

# Estadística 1 inferencial

para ingeniería y ciencias



Eduardo Gutiérrez González  
Olga Vladimirovna Panteleeva



# Estadística Inferencial **1**

para Ingeniería y Ciencias

**EDUARDO GUTIÉRREZ GONZÁLEZ**

PROFESOR DE MATEMÁTICAS DE LA UPIICSA – IPN  
SECCIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

**OLGA VLADIMIROVNA PANTELEEVA**

PROFESORA DE MATEMÁTICAS DE LA UACH  
ÁREA DE MATEMÁTICAS



Para establecer comunicación  
con nosotros puede hacerlo por:



**correo:**  
Renacimiento 180, Col. San Juan  
Tlihuaca, Azcapotzalco,  
02400, Ciudad de México



**fax pedidos:**  
(01 55) 5354 9109 • 5354 9102



**e-mail:**  
info@editorialpatria.com.mx



**home page:**  
www.editorialpatria.com.mx

---

**Dirección editorial:** Javier Enrique Callejas

**Coordinadora editorial:** Estela Delfín Ramírez

**Supervisor de prerensa:** Jorge A. Martínez Jiménez

**Diseño de portada:** Juan Bernardo Rosado Solís/Signx

**Ilustraciones:** Adrian Zamoratequi B.

**Fotografías:** © Thinkstockphoto

**Revisión técnica:**

Ana Elizabeth Gracia Hernández

Instituto Politécnico Nacional

*Estadística Inferencial 1 para ingeniería y ciencias*

Derechos reservados:

© 2016, Eduardo Gutiérrez González, Olga Vladimirovna Panteleeva

© 2016, Grupo Editorial Patria, S.A. de C.V.

Renacimiento 180, Colonia San Juan Tlihuaca

Azcapotzalco, Ciudad de México.

Miembro de la Cámara Nacional de la Industrial Editorial Mexicana

Registro Núm. 43

ISBN ebook: 978-607-744-487-9

Queda prohibida la reproducción o transmisión total o parcial del contenido de la presente obra en cualesquiera formas, sean electrónicas o mecánicas, sin el consentimiento previo y por escrito del editor.

Impreso en México

Printed in Mexico

**Primera edición ebook: 2016**

---

## Agradecimientos

Cuando se termina una obra existen infinidad de compañeros y colegas a quienes se les debe en cierta forma su culminación. Sin la intención de hacer a un lado a nadie, agradecemos infinitamente a todos nuestros compañeros de trabajo, tanto de las academias de Matemáticas como de Investigación de Operaciones y de la Sección de Graduados de la UPIICSA-IPN, así como a los compañeros del Programa en Estadística del colegio de Posgraduados, campus Montecillo, donde adquirimos grandes conocimientos sobre la probabilidad y la estadística que han hecho posible la escritura de este texto. Agradecemos también a los compañeros del área de matemáticas de la UACH, y en particular a los compañeros del grupo GITAM (Grupo de Investigación y Trabajos Académicos de Matemáticas, de las academias de Matemáticas UPIICSA-IPN, fundado en 2013) mediante la línea 2 de investigación sobre probabilidad y Estadística por las aportaciones obtenidas durante el Seminario de Probabilidad y Estadística (2013), así como a los integrantes del diplomado en formación docente en Probabilidad y Estadística con vigencia 2013-2015. Por último, reconocemos a todos los revisores de la editorial, cuyas contribuciones han sido inmejorables para que el texto tenga una mejor presentación y calidad. Por su parte, el doctor Gutiérrez agradece el apoyo brindado a las autoridades de EDD y COFAA para la elaboración de esta obra.

**E.G.G. y O.V.P.**

## Autores

### Eduardo Gutiérrez González

Es doctor en Ciencias (Físico-matemáticas). Realizó estudios de licenciatura, maestría y doctorado en la Universidad Estatal de San Petersburgo, Federación Rusa en Análisis matemático de 1984-1994. Es doctor en Ciencias (Estadística) y realizó estudios de maestría de 2002 a 2004 y el doctorado de 2005 a 2009 en el Colegio de Posgraduados-México en el programa en Estadística. Es maestro en Ingeniería, realizó estudios de maestría en el Posgrado de Ingeniería de la UNAM-México en Ingeniería de Sistemas en el campo disciplinario de Investigación de operaciones de 2004 a 2006. Actualmente es un académico de tiempo completo en la Sección de Estudios de Posgrado e Investigación de la UPIICSA-IPN, además de becario por la DEDICT-COFAA y E.D.D.

### Olga Vladimirovna Panteleeva

Es maestra en Ciencias Físico-Matemáticas (matemáticas aplicadas) y realizó estudios de licenciatura y maestría en la Universidad Estatal de San Petersburgo, Federación Rusa, en matemáticas aplicadas y procesos de control de 1986 a 1992. Es doctora en Ciencias (Estadística) y realizó estudios de maestría de 2005 a 2007 y de doctorado de 2008 a 2012 en el Colegio de Posgraduados-México en el programa en Estadística. Actualmente es una académica de tiempo completo en la Universidad Autónoma de Chapingo en el área de matemáticas.

## Contenido

<b>UNIDAD 1 Estadística descriptiva .....</b>	<b>1</b>
Competencias específicas a desarrollar .....	1
¿Qué sabes? .....	1
Introducción .....	2
<b>1.1 Estadística.....</b>	<b>3</b>
<b>1.2 Población y muestra .....</b>	<b>4</b>
Probabilidad contra estadística .....	4
Caracteres y variables estadísticas .....	5
Escala de medición de una variable .....	6
Escala de medidas cualitativas o no métricas .....	6
Escala de medidas cuantitativas o métricas.....	7
<b>1.3 Técnicas de muestreo .....</b>	<b>9</b>
Muestreo aleatorio simple.....	10
Muestreo estratificado.....	10
Muestreo sistemático con iniciación aleatoria.....	11
Muestreo por conglomerados.....	12
Tamaño de la muestra .....	12
<b>1.4 Parámetros y estadísticos.....</b>	<b>16</b>
<b>1.5 Medidas centrales .....</b>	<b>16</b>
La media .....	16
La mediana .....	18
Cálculo de la mediana .....	18
La moda.....	19
Otros valores medios.....	20
<b>1.6 Medidas de dispersión .....</b>	<b>24</b>
Rango .....	25
Variancia y desviación estándar.....	25
Otra expresión para cálculos de las variancias .....	27
Desviación media .....	27
Covarianza .....	28
<b>1.7 Parámetros de forma en la distribución de la muestra .....</b>	<b>31</b>
<b>1.8 Aplicación de las medidas a inversiones.....</b>	<b>34</b>
<b>1.9 Clases de frecuencia.....</b>	<b>39</b>
Cálculo de las frecuencias acumuladas.....	40
Distribución de frecuencias para variables cuantitativas .....	41
Cantidad de clases para un conjunto de datos cuantitativos .....	41
Amplitud o longitud de clase para datos cuantitativos .....	42
Construcción de clases de frecuencia para datos cuantitativos.....	42
<b>1.10 Gráficos .....</b>	<b>45</b>
Histogramas .....	47
Gráficos lineales, polígonos de frecuencias .....	48
Preguntas de autoevaluación .....	50
Ejercicios complementarios con grado de dificultad uno.....	50
Ejercicios complementarios con grado de dificultad dos.....	52
Ejercicios complementarios con grado de dificultad tres.....	53

<b>UNIDAD 2 Distribuciones muestrales y teorema del límite central .....</b>	<b>55</b>
Competencia específica a desarrollar .....	55
¿Qué sabes? .....	55
Introducción .....	56
<b>2.1 Modelo normal .....</b>	<b>56</b>
Cálculo de probabilidades .....	58
Propiedades de la distribución normal estándar .....	59
Uso de tablas de la función acumulada .....	60
Uso de tablas porcentuales .....	63
<b>2.2 Distribución <i>ji</i> cuadrada .....</b>	<b>67</b>
Uso de tablas de la distribución <i>ji</i> cuadrada .....	67
<b>2.3 Distribución t-Student .....</b>	<b>69</b>
Uso de tablas de la distribución t-Student .....	70
<b>2.4 Distribución <i>F</i> .....</b>	<b>72</b>
Uso de tablas de la distribución <i>F</i> .....	72
<b>2.5 Muestra aleatoria .....</b>	<b>74</b>
<b>2.6 Estadísticas importantes .....</b>	<b>75</b>
Media .....	77
Diferencia de medias .....	77
Varianza insesgada o muestral .....	77
Proporciones .....	77
Media y varianza de la media muestral .....	77
Media y varianza de una diferencia de medias .....	78
<b>2.7 Distribuciones muestrales asociadas a la normal .....</b>	<b>78</b>
Sumas, promedios y combinaciones lineales de variables aleatorias normales con la misma media y varianza .....	79
Cálculo del tamaño de la muestra en distribuciones normales .....	81
Explicación de la desigualdad anterior .....	81
Explicación de la desigualdad anterior .....	82
Fórmulas para el tamaño mínimo de muestra en distribuciones normales .....	82
Diferencia de medias de distribuciones normales .....	85
Cálculo del tamaño de la muestra para diferencia de medias .....	86
<b>2.8 Distribuciones de Bernoulli .....</b>	<b>88</b>
Distribución de la suma de variables de Bernoulli (binomial) .....	88
Media y varianza de una proporción .....	89
Media y varianza de una diferencia de proporciones .....	90
<b>2.9 Teorema central del límite media y suma muestral .....</b>	<b>91</b>
Teorema central del límite para la media de variables .....	91
Teorema central del límite suma de variables .....	92
<b>2.10 Teorema central del límite para diferencia de medias .....</b>	<b>95</b>
<b>2.11 Teorema central del límite para proporciones .....</b>	<b>98</b>
Teorema central del límite para diferencia de proporciones .....	99
Cálculo del tamaño mínimo de muestra para proporciones de muestras grandes .....	100
Teorema central del límite para distribuciones discretas .....	103
Distribuciones a las que no se puede aplicar el teorema central del límite .....	105
Preguntas de autoevaluación .....	105



Ejercicios complementarios con grado de dificultad uno.....	105
Ejercicios complementarios con grado de dificultad dos.....	107
<b>UNIDAD 3 Estimación puntual y por intervalos de confianza.....</b>	<b>109</b>
Competencia específica a desarrollar .....	109
¿Qué sabes? .....	109
Introducción .....	110
<b>3.1 Conceptos básicos sobre estimadores puntuales.....</b>	<b>111</b>
Espacio paramétrico.....	112
Valores de los estimadores puntuales .....	113
Estimadores insesgados .....	115
Estimadores insesgados de distribuciones específicas .....	118
<b>3.2 Conceptos básicos de los intervalos de confianza .....</b>	<b>121</b>
<b>3.3 Intervalos de confianza para los parámetros de una población normal .....</b>	<b>122</b>
Intervalos de confianza para la media de poblaciones normales o aproximadamente normales cuando se conoce $\sigma$ .....	122
Intervalos de confianza para medias de poblaciones normales o aproximadamente normales cuando se desconoce $\sigma$ .....	123
Ejemplos variados para la estimación de la media .....	125
Intervalos de confianza para la varianza de poblaciones normales .....	129
Ejemplos variados para varianzas.....	130
<b>3.4 Intervalos de confianza para comparar dos poblaciones normales.....</b>	<b>134</b>
Resultados posibles de las comparaciones entre dos medias .....	135
Intervalos de confianza para la diferencia de medias, poblaciones aproximadamente normales cuando se conocen $\sigma_1$ y $\sigma_2$ .....	135
Intervalos de confianza para la diferencia de medias de poblaciones normales cuando se desconocen $\sigma_1$ y $\sigma_2$ , pero se sabe que $\sigma_1^2 = \sigma_2^2$ .....	136
Intervalos de confianza para la diferencia de medias de poblaciones normales cuando se desconocen $\sigma_1$ y $\sigma_2$ , pero se sabe $\sigma_1^2 \neq \sigma_2^2$ .....	138
Intervalos de confianza para la diferencia de medias de poblaciones aproximadamente normales, se desconocen $\sigma_1$ y $\sigma_2$ muestras grandes.....	139
Intervalos de confianza para la diferencia de medias de observaciones pareadas con diferencias normales .....	141
Ejemplos variados para la estimación de diferencia de medias.....	144
Intervalos de confianza para la razón entre varianzas de poblaciones normales .....	148
<b>3.5 Intervalos de confianza para proporciones .....</b>	<b>156</b>
Intervalos de confianza para proporciones de muestras grandes .....	156
Ejemplos variados para proporciones .....	157
Con una estimación puntual preliminar .....	157
Con una cota inferior .....	157
Intervalo de confianza de diferencia de proporciones muestras grandes.....	159
Tamaño de muestras en diferencia de proporciones.....	160
Con una estimación puntual preliminar .....	160
Con una cota inferior .....	161
Preguntas de autoevaluación.....	165
Ejercicios complementarios con grado de dificultad uno.....	165
Ejercicios complementarios con grado de dificultad dos.....	165
Ejercicios complementarios con grado de dificultad tres.....	170



<b>UNIDAD 4 Pruebas de hipótesis .....</b>	<b>171</b>
Competencia específica a desarrollar .....	171
¿Qué sabes? .....	171
Introducción .....	172
<b>4.1 Conceptos básicos sobre pruebas de hipótesis.....</b>	<b>172</b>
Regiones de rechazo y no rechazo .....	173
Tipos de errores en una prueba de hipótesis.....	174
Función de potencia y tamaño de la prueba .....	178
Elección de la hipótesis nula y alterna .....	181
Cálculo de las probabilidades para los dos tipos de errores .....	182
Conceptos básicos sobre los tipos de pruebas de hipótesis .....	187
Metodología para probar una hipótesis estadística .....	188
<b>4.2 Pruebas de hipótesis para los parámetros de una distribución normal.....</b>	<b>188</b>
Pruebas de hipótesis para la media de poblaciones aproximadamente normales cuando se conoce $\sigma$ .....	188
Pruebas de hipótesis para la media de poblaciones aproximadamente normales cuando se desconoce $\sigma$ .....	194
Pruebas para la varianza de poblaciones normales.....	199
<b>4.3 Pruebas de hipótesis para comparar dos poblaciones normales .....</b>	<b>206</b>
Pruebas de hipótesis para la diferencia de medias sobre poblaciones aproximadamente normales cuando se conocen .....	206
Pruebas de hipótesis para la diferencia de medias sobre poblaciones aproximadamente normales cuando se desconocen $\sigma_1^2$ y $\sigma_2^2$ pero $\sigma_1^2 = \sigma_2^2$ .....	210
Pruebas de hipótesis para la diferencia de medias sobre poblaciones aproximadamente normales cuando se desconocen $\sigma_1^2$ y $\sigma_2^2$ pero $\sigma_1^2 \neq \sigma_2^2$ .....	214
Pruebas de hipótesis para la diferencia de medias de observaciones pareadas con diferencias normales .....	218
Pruebas de hipótesis para la razón entre varianzas de poblaciones normales .....	222
<b>4.4 Pruebas para poblaciones tipo Bernoulli, proporciones.....</b>	<b>229</b>
Preguntas de autoevaluación .....	240
Ejercicios complementarios con grado de dificultad uno.....	240
Ejercicios complementarios con grado de dificultad dos.....	240
Ejercicios complementarios con grado de dificultad tres.....	244
 <b>UNIDAD 5 Pruebas de bondad de ajuste.....</b>	 <b>247</b>
Competencias específicas a desarrollar .....	247
¿Qué sabes? .....	247
Introducción .....	248
<b>5.1 Pruebas de bondad de ajuste de forma gráfica.....</b>	<b>248</b>
Cuantiles.....	248
Técnica gráfica Q-Q para una prueba de ajuste de distribuciones .....	250
Ejemplo de la técnica gráfica Q-Q para una prueba de normalidad .....	250
Técnica analítica Q-Q para una prueba de normalidad .....	253
<b>5.2 Prueba de bondad de ajuste ji cuadrada.....</b>	<b>254</b>
Metodología de la prueba ji cuadrada.....	254
Valor-p en una prueba de hipótesis.....	256
<b>5.3 Uso de las pruebas de bondad de ajuste K-S y A-D.....</b>	<b>256</b>
Prueba de bondad de ajuste Kolmogorov-Smirnov .....	256

Prueba de bondad de ajuste Kolmogorov-Smirnov con Minitab.....	261
Prueba de bondad de ajuste Anderson-Darling con Minitab .....	263
Preguntas de autoevaluación .....	266
Ejercicios complementarios con grado de dificultad dos.....	266
<b>UNIDAD 6 Regresión lineal simple y múltiple .....</b>	<b>271</b>
Competencias específicas a desarrollar .....	271
¿Qué sabes? .....	271
Introducción .....	272
<b>6.1 Regresión lineal simple .....</b>	<b>273</b>
Diagrama de dispersión .....	274
Supuestos de la variable dependiente en el análisis de regresión .....	276
Supuestos del error en un modelo lineal .....	277
<b>6.2 Método de mínimos cuadrados para optimizar el error .....</b>	<b>278</b>
<b>6.3 Error estándar de estimación y propiedades de los estimadores.....</b>	<b>286</b>
<b>6.4 Prueba de hipótesis para el parámetro de la pendiente .....</b>	<b>289</b>
<b>6.5 Coeficientes de correlación y determinación .....</b>	<b>291</b>
Coeficiente de correlación lineal.....	291
Coeficiente de determinación.....	297
<b>6.6 Intervalos de confianza para la predicción y estimación .....</b>	<b>299</b>
<b>6.7 Regresión lineal múltiple .....</b>	<b>305</b>
Planteamiento general del modelo de regresión lineal múltiple .....	305
Generalización de resultados de la regresión lineal y prueba F.....	307
Coeficiente de determinación ajustado.....	308
Prueba $F$ , análisis de varianza .....	309
Uso de Excel para la regresión lineal múltiple .....	310
Solución de un modelo de regresión lineal múltiple.....	313
Análisis de residuales en la regresión lineal múltiple .....	321
Independencia y valor esperado cero de los errores.....	321
Varianza constante de los errores .....	321
Observaciones atípicas o aberrantes .....	322
Problemas en la regresión lineal múltiple.....	324
Regresión curvilínea .....	328
Modelos de regresión con errores multiplicativos .....	332
Modelos de regresión con variables de respuesta transformadas.....	336
Preguntas de autoevaluación.....	343
Ejercicios complementarios con grado de dificultad 1 .....	343
Ejercicios complementarios con grado de dificultad 2.....	345
Ejercicios complementarios con grado de dificultad 3.....	345
Caso de estudio .....	348



## Prefacio

### Palabras de los autores

En términos generales, el libro está dividido en cuatro partes. En la primera, trabajamos con estadística descriptiva, en la segunda con estadística inferencial, en la tercera parte con las pruebas de bondad de ajuste y en la cuarta con los modelos de regresión lineales. Con estas cuatro partes, el libro se complementa realizando un avance completo de los conceptos básicos que tienen mayor aplicación en problemas prácticos de las diferentes esferas de la estadística descriptiva e inferencial en ingeniería y ciencias.

La primera parte de la obra la dedicamos al estudio de la estadística descriptiva para datos no agrupados, en la que analizamos las diferentes medidas, tanto centrales como de desviación. Dentro de las medidas centrales estudiamos la media, mediana, moda, media geométrica, media ponderada y media armónica. En las medidas de desviación analizamos el rango, la varianza y la desviación estándar. Revisamos los coeficientes de variación y covarianza, los parámetros de forma para un conjunto de datos y continuamos con algunas aplicaciones de los datos no agrupados a inversiones. Estudiamos las clases de frecuencias y sus gráficas por medio de histogramas y polígonos de frecuencia, con los que se analizan las distribuciones de los datos; simetría y sesgo que serán utilizadas en la unidad 5 para llevar a cabo una prueba de bondad de ajuste.

El estudio sobre las distribuciones muestrales lo iniciamos en la unidad 2, donde hablamos detalladamente sobre las distribuciones muestrales de la media y diferencia de medias para variables normales. Ampliamos las distribuciones muestrales para la suma y el promedio de la distribución Bernoulli. Por último, hacemos una revisión detallada del teorema central del límite en sus diferentes presentaciones: media, suma y distribuciones específicas.

En la unidad 3 hablamos brevemente sobre los estimadores puntuales y sus propiedades más importantes. Revisamos con mucho detalle los intervalos de confianza. Iniciamos con los conceptos básicos sobre las propiedades de un buen intervalo de confianza y con estos conceptos revisamos a detalle la parte metodológica de los intervalos de confianza para los parámetros de poblaciones normales o aproximadamente normales, para una población y comparación de poblaciones. Finalizamos con intervalos de confianza para proporciones y diferencia de proporciones en muestras grandes.

En la unidad 4 hacemos una revisión similar a la unidad 3, pero ahora utilizamos las pruebas de hipótesis. Iniciamos con la revisión de los conceptos básicos sobre pruebas de hipótesis y después revisamos la metodología para las pruebas de hipótesis. Primero, revisamos qué es una hipótesis estadística y cuáles son los errores que cometemos al llevar a cabo una prueba. Asimismo, tratamos con detenimiento la potencia de la prueba. Al final, revisamos la parte metodológica de las pruebas de hipótesis para los parámetros de poblaciones normales o aproximadamente normales y poblaciones tipo Bernoulli.

En la tercera parte del texto revisamos brevemente un contraste de hipótesis muy particular; nos referimos a las pruebas de bondad de ajuste, con lo que podemos justificar estadísticamente la distribución de la que provienen los datos muestrales. Se revisan las principales pruebas, desde una prueba gráfica llamada Q-Q, una prueba de frecuencias, conocida como prueba  $\chi^2$ -cuadrada, hasta dos pruebas no paramétricas que trabajan con las funciones de distribución acumulada, las pruebas Kolmogorov-Smirnov (K-S) y Anderson-Darling (A-D).

En la cuarta parte del texto en un solo capítulo hacemos una revisión, a detalle, de los modelos de regresión, tanto simples como múltiples. En el caso de una regresión simple explicamos cómo llevar a cabo un análisis sobre la regresión, desde la construcción de un diagrama de dispersión hasta los intervalos de confianza y pruebas de hipótesis de los parámetros de la regresión. Durante el desarrollo de los resultados de una regresión vemos cómo encontrar e interpretar la ecuación de regresión, cómo obtener predicciones y calculamos intervalos de confianza para las predicciones. Con la regresión múltiple ampliamos los modelos a regresiones curvilíneas, casos con errores multiplicativos y problemas de Cobb-Douglas. Además, explicamos a detalle los diferentes problemas que se pueden presentar con las observaciones de una muestra, como puede ser la multicolinealidad, datos aberrantes, transformaciones Box-Cox para variables de respuesta no normales, etcétera.

Sin importar los avances que tenemos en computación y en la teoría de la estadística en los textos metodológicos sobre aplicaciones de la estadística inferencial, se conserva el viejo esquema del uso exclusivo de la distribución normal para las fórmulas y métodos que se acostumbra usar en los intervalos de confianza y la prueba de hipótesis. Por otro lado, los textos que hablan sobre las bases teóricas para diferentes tipos de distribuciones resultan ser demasiado teóricos, de manera que a un lector sin formación matemática se le dificulta comprender el desarrollo del libro.

En la presente obra damos un enfoque metodológico. Así, el lector que solo tenga interés en la parte metodológica de la estadística descriptiva e inferencial, pruebas de bondad de ajuste y los modelos de regresión lineales podrá avanzar en su estudio sin problemas.

### Unas palabras del estilo y la forma de escritura

El estilo de escritura del libro es muy sencillo, pero a diferencia de otros libros metodológicos muestra conceptos que son la base para los desarrollos teóricos más complejos. Cada tema tratado está reforzado por una gran cantidad de ejemplos y ejercicios prácticos en cada sección que abarcan diferentes formas de ver un problema. Las soluciones y sugerencias a la mayoría de los problemas están en la página electrónica del libro en SALI y fueron hechas en Excel-Microsoft considerando todos los dígitos; por estas razones las soluciones que obtenga el lector pueden variar ligeramente con respecto a las mostradas en la página electrónica del libro en SALI, pero estas variaciones deben ser mínimas.

Cada sección se escribe con el número del capítulo al que pertenece, seguida de un punto y el número correspondiente a la sección dada, comenzando con la sección *uno* en cada capítulo. Ejemplo **4.3**, significa la sección 3 del capítulo 4.

### Bases teóricas requeridas

Para la total comprensión de los temas se requieren solo conocimientos básicos de los cursos de cálculo diferencial e integral. En los temas no es necesario el manejo de las demostraciones, pero no así en los ejemplos y ejercicios correspondientes.

### Objetivos del texto

El objetivo de este libro es presentar a los futuros profesionistas herramientas cuantitativas que puedan aplicar en los problemas que les corresponda resolver dentro de su ámbito de trabajo y llegar a una mejor toma de decisiones. Al final del texto esperamos que el lector sea capaz de:

- *Describir* las diferentes técnicas de la estadística descriptiva, para llevar a cabo un estudio detallado del comportamiento de los datos.
- *Definir* los conceptos de parámetros y estadísticos.
- *Nombrar* las diferentes técnicas que se pueden usar para llevar a cabo inferencias.
- *Identificar* en un problema dado, cuándo un dato se refiere a un parámetro o a un estadístico.
- *Aplicar* las inferencias a su área de trabajo.
- *Experimentar* desde el punto de vista de la estadística inferencial.
- *Proponer e investigar* experimentos en los que se tengan distribuciones muestrales para hacer inferencias con respecto a sus parámetros.
- *Aplicar* la regresión lineal para determinar relaciones entre variables y poder hacer predicciones en situaciones de su área de trabajo.



# Estadística descriptiva

UNIDAD

1



## Competencias específicas a desarrollar

- Comprender que mediante el estudio de la muestra se puede estimar el comportamiento poblacional.
- Recolectar de manera adecuada los datos para poder realizar estudios estadísticos, así como tomar decisiones.

## ¿Qué sabes?

- ¿Qué es estadística descriptiva?
- ¿Qué es una muestra?
- ¿Cómo puedes estimar el comportamiento de una muestra?
- ¿Cómo haces la recolección de datos?
- ¿Cómo se calcula la media y la desviación estándar?

## Introducción

Desde tiempos muy remotos, el ser humano ha tenido que analizar una gran cantidad de datos o información referente a los problemas o actividades de sus comunidades. Por ejemplo, a partir del inicio de la civilización se utilizaban representaciones gráficas y otros símbolos en pieles, rocas, palos de madera y paredes de cuevas para contar el número de personas, animales u objetos. Hacia el año 3 000 a.C., los babilonios usaban pequeñas tablillas de arcilla para recopilar datos sobre la producción agrícola y los productos vendidos o intercambiados mediante el trueque. Se sabe que mucho antes de construir las pirámides los egipcios analizaban los datos de la población y la renta del país.

Carolingia, dinastía de reyes francos (también llamada Carlovingia), que gobernaron un vasto territorio en Europa occidental desde el siglo VII hasta el siglo X d.C., tomó su nombre de su más renombrado miembro, Carlomagno.

Otro ejemplo muy claro de la recopilación y análisis de datos es el Imperio Romano, el primer gobierno que, al verse en la necesidad de mantener un control sobre sus esclavos y riqueza, recopiló una gran cantidad de datos referentes a la población, superficie y renta de todos los territorios bajo su control. Luego, a mediados del primer milenio, debido al gran crecimiento de las poblaciones, surgió la necesidad de tener un control sobre éstas. Así, empezaron a aplicarse los primeros censos poblacionales, como los que se llevaron a cabo durante la Edad Media en Europa. Los reyes carolingios Pipino “el Breve” y Carlomagno fueron algunos de los primeros monarcas en ordenar la realización de estudios minuciosos de las propiedades que tenía la Iglesia entre los años 758 y 762, respectivamente.

Con el paso del tiempo, la recopilación y el análisis de datos comenzaban a tener otro fin, además de los censos y el conocimiento de propiedades. Por ejemplo, en Inglaterra, a principios del siglo XVI, se realizó el registro de nacimientos y defunciones, lo que dio origen al primer estudio de datos poblacionales, *Observations on the London Bills of Mortality (Comentarios sobre las partidas de defunción en Londres)*, publicado 1662. Un estudio similar sobre la tasa de mortalidad en la ciudad de Breslau, en Alemania, realizado en 1691, fue utilizado por el astrónomo inglés Edmund Halley como base para la primera tabla de mortalidad. Más tarde, en el siglo XIX, con la generalización del método científico para estudiar todos los fenómenos de las ciencias naturales y sociales, los investigadores aceptaron la necesidad de reducir la información a valores numéricos, con el fin de evitar la ambigüedad de las descripciones verbales.

Ahora bien, respecto a los orígenes de la palabra estadística, se sabe que el vocablo *statistik* proviene de la palabra italiana *statista* (que significa “estadista”), la cual que fue utilizada por primera vez por Gottfried Achenwall (1719-1772), un profesor de Marborough y de Göttingen. Por su parte, el doctor W. Zimmerman introdujo el término *statistics* (estadística) a Inglaterra, cuyo uso fue popularizado por sir John Sinclair en su obra *Statistical Account of Scotland 1791-1799 (Informe estadístico sobre Escocia 1791-1799)*. Sin embargo, mucho antes del siglo XVIII, la gente ya utilizaba y registraba datos.

Hoy día, la estadística es un área de las matemáticas que tiene gran auge en las aplicaciones y los estudios de investigación de todo tipo. Por ejemplo, es difícil imaginar a un ingeniero industrial sin conocimientos de estadística; más aún, no es posible considerar una aseveración de investigación de cualquier área de las ciencias, medicina, ingeniería mecánica, eléctrica, automotriz, ingeniería civil, ingeniería económica, economía, finanzas, etc., sin que esta aseveración tenga fundamentos estadísticos. Un ingeniero eléctrico tiene que hacer diferentes pruebas para comparar la conductividad de algunos materiales e indicar cuál de los que está comparando es mejor conductor de la electricidad y con qué porcentaje de error. Por su parte, un ingeniero civil realiza cálculos de resistencia de materiales para la construcción, cuyos resultados presenta con base en las pruebas realizadas, acompañados de un rango de error. En otro ejemplo, un médico aplica nuevos medicamentos a pacientes que han aceptado formar parte de un proceso de experimentación del o los laboratorios que lo producen; así, mediante un análisis estadístico se conocen las probabilidades de éxito y riesgos en su aplicación.

En esta unidad damos inicio a la estadística como contraparte de la teoría de las probabilidades revisada en un curso previo al de Estadística. Desde este momento, y a diferencia de lo que se ha revisado en cursos previos sobre los conceptos teóricos en las probabilidades, trabajaremos con datos u observaciones de experimentos, a partir de los cuales buscaremos el comportamiento general del que provienen. Por ejemplo, un ingeniero industrial que trabaje en el área de control de calidad de los procesos de producción de la empresa, desea conocer cuál es el comportamiento



probabilístico de la producción de artículos buenos o defectuosos en las líneas de producción. En general, conocer el resultado del comportamiento de un fenómeno aleatorio mediante un conjunto de datos no resulta una tarea sencilla. Por tanto, a partir de ahora lo resolveremos de manera paulatina.

El análisis estadístico de las observaciones de cualquier estudio que revisaremos en el texto inicia en esta unidad, en la cual veremos cómo hacer un análisis estadístico de tipo descriptivo acerca de un conjunto de observaciones que se desee investigar.

La unidad inició con la definición de estadística, después se habla sobre población y muestra, y se realiza un comparativo entre estadística y probabilidad. Continuamos con las variables estadísticas y una clasificación de éstas, según la escala de medición de la variable. Luego, explicamos de manera breve las principales técnicas de muestreo y el tamaño de muestra. Después de esto, definiremos y ejemplificaremos qué es un parámetro y un estadístico, extendiendo su estudio a las medidas centrales más comunes en la estadística descriptiva. Explicamos que el conocimiento de una sola medida central no es suficiente para conocer el comportamiento de las observaciones, para esto se introducen las medidas de desviación y forma mediante el coeficiente de asimetría. Por último, mostramos una aplicación de los conceptos estudiados a las inversiones.

## 1.1 Estadística

Después de revisar la introducción de la unidad, estamos de acuerdo en que el ser humano se vio en la necesidad de crear una ciencia encargada de reducir la información a valores numéricos para la mejor y fácil interpretación de los fenómenos que lo rodeaban, a la que dio el nombre de **estadística**. Ahora bien, ¿qué entendemos por estadística?

La **estadística** es la rama de las matemáticas que proporciona métodos para reunir, organizar y analizar información y usarla para obtener diversas conclusiones que pueden ayudar a resolver problemas en la toma de decisiones y el diseño de experimentos.

En este sentido, ¿cuál es la función actual de la estadística? La estadística se ha convertido en un método efectivo para describir con un grado de exactitud los valores de datos económicos, políticos, sociales, psicológicos, químicos, biológicos y físicos, entre otros; y sirve como herramienta para relacionarlos y analizarlos. Por estas razones, se clasifica en diferentes campos. Entre los que tienen mayor aplicación y una importancia más relevante en nuestros días destacan:

- Estadística descriptiva
- Estadística inferencial
- Regresión lineal
- Diseños de experimentos
- Análisis multivariado
- Estadística no paramétrica
- Estadística espacial

En el presente texto estudiamos:

- Estadística descriptiva
- Estadística inferencial
- Regresión lineal

La estadística descriptiva se estudia esta unidad, mientras que la inferencial se revisa en las unidades 2, 3, 4 y 5. Por último, en la unidad 6 veremos la regresión lineal.

## 1.2 Población y muestra

La materia prima de la estadística consiste en conjuntos de números obtenidos al contar o medir los resultados de un experimento. Entonces, al recopilar datos estadísticos, debemos tener especial cuidado en garantizar que la información sea completa y correcta. Así, el primer problema para los estadísticos reside en determinar qué información y en qué cantidad deberá reunirse. Por ejemplo, en la práctica, la dificultad al llevar a cabo un censo reside en obtener el número de habitantes de forma completa y exacta; de la misma manera, cuando un ingeniero o un físico quiere contar el número de colisiones por segundo entre las moléculas de un gas, debe empezar por determinar con precisión la naturaleza de los objetos a contar; para ello es necesario saber cómo se obtienen los conjuntos de datos en esta disciplina.

Debido a que la naturaleza de los fenómenos que podemos analizar varía mucho, es necesario definir los conjuntos de datos que vamos a revisar.

Se llama **población** al conjunto de todos los elementos de un tipo particular cuyo conocimiento es de nuestro interés.

Cada uno de los elementos que intervienen en la definición de población es un individuo u objeto. Los elementos de la población se denominan así debido a que originalmente el campo de actuación de la estadística fue la demografía.

Con frecuencia, la información disponible para un estudio consta de una porción o subconjunto de la población. Por este motivo, introducimos un segundo concepto, **muestra** de una población.

Se llama **muestra** a cualquier subconjunto de la población, en realidad en el texto nos interesan los subconjuntos no vacíos y finitos.

### Ejemplos 1.1 Muestra

1. Si el conjunto de datos de interés está constituido por todos los estudiantes de licenciatura en el Tecnológico de Tlalnepantla, cada uno de los estudiantes será un individuo estadístico, mientras que el conjunto de todos los estudiantes será la población. Una muestra podría ser el conjunto de todos los estudiantes del cuarto semestre de la licenciatura en administración.
2. La producción de autos de una armadora ubicada en Morelos. En este ejemplo, la población es cada uno de los autos ensamblados (individuos estadísticos); por su parte, una muestra se puede proponer como los autos subcompactos fabricados en mayo.
3. El estudio de ciertos experimentos químicos. En este caso, cada uno de los experimentos será un individuo estadístico y el conjunto de todos los posibles experimentos en esas condiciones será la población, mientras que una muestra podría ser un conjunto de resultados experimentales en ciertas condiciones.
4. El conjunto de mediciones, en toneladas, de la carga máxima soportada por los cables de acero producidos por la compañía CM. En este caso, los individuos se refieren a los cables producidos por esta empresa durante un periodo determinado.

## Probabilidad contra estadística

En los cursos de probabilidad se revisa a detalle que las variables aleatorias están definidas por algún comportamiento particular y por estas razones se les asignan nombres para diferenciarlas; por ejemplo, una variable aleatoria normal sigue una distribución normal cuya forma se caracteriza por ser semejante a una campana y por este motivo también se le conoce como campana de Gauss. El comportamiento de una variable aleatoria está

determinado por diferentes medidas que llamamos parámetros, que pueden ser de localidad, escala o forma. En el caso de una variable aleatoria normal, se trata de la media o parámetro de localidad,  $\mu$ , y la desviación estándar o parámetro de escala  $\sigma$ . Un repaso de la distribución normal lo veremos en la unidad 2, junto con algunas otras distribuciones de gran importancia para la estadística inferencial.

Cuando hablamos de los parámetros, implícitamente decimos que conocemos cómo se comporta toda la población. De esta forma, en probabilidad siempre se parte del hecho de que conocemos los parámetros de la distribución y calculamos probabilidades de que ocurra algún valor o valores particulares de la población (valor muestral).

Veremos que en la estadística inferencial se tiene el proceso inverso al de la probabilidad o, dicho de otra forma, el proceso aplicado a observaciones o datos, muestra que a partir de estos queremos conocer cómo es la población, cómo son sus parámetros, etcétera. Este proceso inverso es de gran importancia en las aplicaciones, pues en general en un problema estadístico trabajamos con los valores de las observaciones sin conocer la distribución y parámetros que explican cuál es la población. Por ejemplo, un ingeniero industrial que desee hacer una planeación sobre la parte de inventarios de la cadena de suministro de su empresa, requiere hacer pronósticos sobre la demanda que tiene el almacén, para lo cual debe contar con alguna base de datos de las demandas. Por su parte, el ingeniero civil requiere de información previa acerca de la frecuencia e intensidad de los temblores en el lugar en que iniciará una construcción, además de hacer un estudio previo sobre el suelo.

En la presente unidad revisamos con mucho detalle diferentes técnicas para el manejo de datos de forma descriptiva, con las cuales es posible decir que empezamos a tener conocimientos intuitivos sobre el comportamiento de la población de forma descriptiva. Es decir, el conocimiento del presente material es fundamental para poder describir los procesos de producción, la logística de una empresa, los problemas de inventarios y los problemas de ingeniería relacionados con mediciones, como la dureza de material, conductividad de metales, etcétera.

## Caracteres y variables estadísticas

Antes, cuando definimos una población hablamos sobre sus elementos, a los que llamamos individuos; además, en los ejemplos 1.1 se nota que éstos pueden ser descritos por una o varias de sus propiedades o características, situación que da origen a la definición de carácter.

Se llama **carácter** de un individuo u objeto a cualquier característica o propiedad por medio de la cual es posible clasificar y estudiar a dicho individuo.

Veamos algunos ejemplos de carácter para poder utilizar esta definición con mayor libertad.

### Ejemplos 1.2 Carácter

1. Si los individuos son personas, entonces el sexo, el estado civil, el número de hermanos o su estatura son caracteres.
2. Si los individuos son computadoras, entonces un carácter podría ser la rapidez del procesador y la capacidad del disco duro, entre otras.
3. Si el individuo es una reacción química, entonces el tiempo de la reacción, la cantidad de producto obtenido o si éste es ácido o básico son posibles caracteres.

Un carácter puede ser:

- Cualitativo o no métrico, si no admite medición numérica.
- Cuantitativo o métrico, si es contable o medible numéricamente.

Ahora definamos qué es un carácter cualitativo y qué es un carácter cuantitativo. Los datos no métricos o caracteres cualitativos son atributos, características o propiedades categóricas que identifican o describen a un sujeto. Describen diferencias en tipo o clase, e indican la presencia o ausencia de una característica propia. Por ejemplo, si una persona es mujer, se excluye que sea hombre. Es decir, no hay cantidad de género, solo la condición de ser mujer u hombre. Por otro lado, los datos métricos o caracteres cuantitativos están constituidos de manera que los sujetos pueden estar identificados por diferencias entre sus cantidades. Es decir, las variables medidas métricamente reflejan cantidades relativas. Por esta razón, las medidas métricas son las más apropiadas para casos que involucren cantidad o magnitud, como la demanda de trabajo, el nivel de ozono en la atmósfera, etcétera.

Ahora deseamos responder la pregunta, ¿qué es una variable estadística? Los distintos valores que puede tomar un carácter cuantitativo configuran una **variable estadística**. Ésta es de dos tipos: discreta y continua.

Una **variable estadística** es **discreta** cuando solo permite valores aislados que pueden ser numerables y proviene de un conteo.

### Ejemplo 1.3 Variable discreta

En cierta población, la variable que representa al número de hermanos puede tomar los valores: 0, 1, 2, 3, 4 y 5. Este tipo de variables se caracterizan por obtenerse mediante un proceso de conteo (véase el tema semejanza con las variables aleatorias discretas de la teoría de las probabilidades).

Una **variable estadística** es **continua** cuando admite todos los valores de un intervalo y proviene de una medición.

### Ejemplos 1.4 Variable continua

1. En cierta población, la variable que representa la estatura de una persona adulta que se mide, puede tomar cualquier valor en el intervalo 135-215 cm.
2. La variable temperatura de una persona puede tomar cualquier valor en el intervalo 20-41 °C.

Como se puede ver, este tipo de variables se caracterizan porque se obtienen por medio de mediciones.

## Escalas de medición de una variable

Se mencionó que tenemos dos tipos de datos: cualitativos o no métricos y cuantitativos o métricos. En esta sección discutimos un poco más sobre sus escalas de medición.

### Escalas de medidas cualitativas o no métricas

Las medidas no métricas pueden tener escalas nominales y ordinales.

- **Escala nominal o de categorías.** En esta escala podemos usar números para etiquetar o identificar a los sujetos u objetos, pero no hay relación de orden. El número asignado solo sirve para determinar la cantidad de ocurrencias en cada clase o categoría de la variable que estamos estudiando. Por ejemplo, los números que se asignan al sexo o al estado civil de una persona, solo sirven para indicar la presencia o ausencia del atributo o característica bajo estudio. Esta escala es propicia solo para variables discretas y sirve para clasificar a la población.

### Ejemplos 1.5 Escala nominal o de categorías

1. El sector económico se clasifica en: primario, industrial y de servicios.
  2. Profesión: ingeniero, médico, matemático, abogado, etcétera.
  3. Propiedad del suelo: agrícola, forestal, urbano, etcétera.
  4. Sexo o género de la persona: masculino y femenino.
  5. Colores de un objeto: blanco, negro, rojo, entre otros.
- **Escala ordinal.** Este tipo de escala cualitativa presenta un nivel superior de precisión de la medida que la escala nominal. Las variables pueden ser ordenadas o clasificadas en escalas ordinales en relación con la cantidad del atributo poseído. Podemos realizar una relación de orden entre las clases con base en un gradiente ascendente mayor que o descendente menor que. Esta escala es propicia solo para variables discretas y sirve para ordenar los datos.

### Ejemplos 1.6 Escala ordinal

1. Diferentes niveles de satisfacción de una persona sobre un producto determinado, pueden ser: muy satisfecho, medio satisfecho y no muy satisfecho. Estos atributos representan una relación de orden en forma descendente.

Observe que los números posibles a utilizarse en esta escala no son cuantitativos, dado que indican posiciones relativas en series ordenadas. Esto se debe a que no hay medida de cuánta satisfacción recibe el consumidor en términos absolutos; más aún, el investigador ni siquiera conoce con certeza la diferencia exacta entre distintos puntos de la escala de satisfacción.
2. Niveles de estudio de un candidato a ocupar un puesto en la empresa: pasante, licenciado, maestría, doctorado. Estos atributos representan una relación de orden en forma ascendente.
3. Clases sociales respecto a su poder adquisitivo: baja, media y alta. Estos atributos representan una relación de orden en forma ascendente.
4. Clases de autos: lujo, deportivo, automático equipado, automático, estándar equipado, estándar, austero. Estos atributos representan una relación de orden en forma descendente.

Cuando una variable cualitativa solo puede tener dos categorías que podemos nombrar de presencia-ausencia, suelen llamarse **variables cualitativas binarias**. Es decir, la variable indica la presencia o ausencia de un atributo; este tipo de variables suelen encontrarse en los cuestionarios; por ejemplo: la persona tiene casa propia o no; la persona tiene trabajo o no; un adolescente estudia o no, etcétera.

## Escalas de medidas cuantitativas o métricas

Las medidas métricas pueden tener escalas por intervalos y razón, éstas proporcionan el nivel más alto de medida de precisión, permitiendo realizar casi todas las operaciones matemáticas. Las dos escalas tienen unidades constantes de medida, de manera que las diferencias entre dos puntos adyacentes de cualquier parte de la escala son iguales. La única diferencia real entre las escalas de intervalo y las de razón es que la primera tiene un punto cero arbitrario, mientras que la segunda tiene un cero absoluto.

- **Escala de intervalos.** Con esta escala podemos medir distancias; el cero es arbitrario como punto de referencia. Esta escala es propicia para variables tanto discretas como continuas. Las escalas de intervalos más comunes son las de temperatura Celsius y Fahrenheit. Ambas tienen un punto del cero arbitrario, pero éste no indica una cantidad cero o ausencia de temperatura, dado que se pueden registrar por debajo del punto cero de esa escala. Por tanto, no podemos decir que un valor cualquiera situado en un intervalo de la escala es un múltiplo de cualquier otro punto de la misma escala.

## Ejemplo 1.7 Escala de intervalos

Si en un día se registra una temperatura de 80 °F (Fahrenheit) no podemos decir que ésta es dos veces más calurosa que otro día con una temperatura de 40 °C. Esto se debe a que en la escala Celsius las temperaturas equivalen a 26.7 y 4.4°, respectivamente y obviamente  $26.7 \neq 2 \times 4.4$ , de tal manera que no se puede afirmar que el calor de 80 °F sea dos veces el calor de 40 °C, porque al usar diferentes escalas éste no es dos veces mayor.

- **Escala de razón (proporción).** Este tipo de escala es la más relevante, pues con esta es posible tomar prácticamente cualquiera de las medidas que se estudian más adelante. También, en esta escala podemos hacer todas las operaciones aritméticas y es propicia para variables, tanto discretas como continuas. En el caso de escala, el cero toma un valor absoluto, por lo que las medidas pueden expresarse en múltiplos cuando se relaciona un punto con otro de la escala.

## Ejemplo 1.8 Escala de razón (proporción)

El peso de un mueble de 40 kg es el doble de uno de 20 kg. Entre otros ejemplos, se encuentran el nivel de inflación, el producto interno bruto, la tasa de interés, el tipo de cambio, los precios de la mezcla de petróleo, etcétera.

Para finalizar la sección precisamos la definición de estadística descriptiva.

El uso adecuado de las diferentes escalas de medición es muy importante para que el estudiante o el investigador identifique la escala de medición de cada variable empleada, de manera que no utilice datos no métricos como si lo fueran.

La parte de la estadística que analiza, estudia y describe a la totalidad de individuos de una población o muestra se llama **estadística descriptiva**.

Ahora bien, ¿cuál es la finalidad de la estadística descriptiva?

La estadística descriptiva busca obtener información para después analizarla, elaborarla y simplificarla lo necesario para que pueda ser interpretada, cómoda y rápidamente y, en consecuencia, se pueda utilizar de manera eficaz para algún fin deseado.

El proceso que sigue la estadística descriptiva para el estudio de una población o muestra consta de los siguientes pasos:

- Selección de caracteres dignos de ser estudiados.
- Obtención del valor de cada individuo mediante una encuesta o medición, con respecto a cada uno de los caracteres seleccionados.
- Obtención de números que sinteticen los aspectos más relevantes de una distribución estadística (más adelante a dichos números en el caso de la población les llamaremos parámetros, mientras que en el caso de las muestras, estadísticos).
- Elaboración de tablas de frecuencias, mediante la adecuada clasificación de los individuos dentro de cada carácter.
- Representación gráfica de los resultados.

## Ejercicios 1.1

En cada caso indique al sujeto estadístico, observación y población; discuta sobre una posible muestra, indique el carácter de interés y el tipo del carácter.

1. El director de una escuela primaria lleva un control de la edad en años de los alumnos de la escuela.
2. El director de una escuela primaria mide las estaturas de los alumnos de la escuela.
3. El supervisor de una línea de producción de botes de jugo lleva el control sobre la cantidad de líquido envasado, con la finalidad de llevar un control por día.
4. El supervisor de una línea de producción de botellas de refresco lleva el control sobre la cantidad de botellas envasadas en la línea de producción que estén en alguno de los tres rangos (llenado alto, medio y bajo) establecidos por el departamento de control de calidad de la envasadora.
5. El gerente de mercadería de una compañía recibe los informes sobre el volumen de ventas diarias de la compañía durante un año y le interesa conocer su utilidad diaria.
6. El gerente de mercadería de una compañía recibe los informes sobre el volumen de ventas diarias de la compañía durante un año.

En cada uno de los ejercicios indique el tipo de escala que se utilizaría para llevar a cabo un estudio estadístico y explique.

7. En el caso de los promedios de los estudiantes de licenciatura en la universidad, el carácter se refiere a la calificación promedio de cada uno y es de tipo métrico continuo.
8. En el caso de los promedios de los grupos de licenciatura en la universidad, el carácter se refiere a la calificación promedio de los grupos y es de tipo métrico continuo.
9. En el caso del gerente de mercadería sobre el volumen de ventas diarias de la compañía durante un año, el carácter se refiere al volumen de ventas al día y es de tipo métrico discreto.
10. En el caso del gerente de mercadería sobre el volumen de ventas diarias de la compañía durante un año, en donde interesaba la utilidad diaria, el carácter se refiere a la utilidad y es de tipo métrico discreto.
11. Cuando los individuos son personas, entonces el sexo y el estado civil son caracteres de tipo cualitativo.
12. Si el individuo es una reacción química, entonces si éste es ácido o básico se trata de un carácter de tipo no métrico.

## 1.3 Técnicas de muestreo

Los estadísticos enfrentan un problema complejo cuando deben seleccionar una muestra para un sondeo de opinión o una encuesta electoral, puesto que seleccionar una muestra capaz de representar con exactitud las preferencias del total de la población no es tarea fácil. Más aún, un buen muestreo debe proporcionar resultados más oportunos que permitan la obtención rápida de información de toda una población o sobre un proceso variable.

Además de lo anterior, vemos que el buen muestreo es indispensable para los problemas estadísticos donde el estudio de toda la población resulta ser muy caro o, en los casos en que la información se destruye, no sería factible. Por ejemplo, en el control de calidad sobre la vida media de las bombillas se llevan a cabo pruebas de tipo destructivo puesto que la muestra se analiza hasta que las bombillas dejen de funcionar.

En muchas situaciones, el muestreo produce resultados más exactos que en un censo (un censo se lleva a cabo cuando es indispensable analizar todos los casos de una población), dado que la pesada carga de trabajo de procesar la información de un censo produce una gran fatiga que, a su vez, puede ser la responsable de prácticas poco adecuadas por parte de los investigadores. Asimismo, la población puede ser muy dinámica y no mantenerse en un estado el tiempo necesario para medir sus características.

Por otro lado, existe infinidad de casos de laboratorio o experimentos que no tienen todos los datos de la población, ya que ocurren solo con las repeticiones de los experimentos que pueden ser infinitas. En este sentido, es necesario saber, ¿qué entendemos por muestreo?

El **muestreo** es simplemente un conjunto de métodos para obtener muestras.



Pero, ¿qué buscamos con el muestreo? Obtener con el mínimo costo, la máxima información sobre las medidas de la población (parámetros). En otras palabras, encontrar con una muestra pequeña la mayor información posible de los parámetros.

Al usar un muestreo deben tomarse las precauciones necesarias para asegurar la aleatoriedad de las muestras. Por consiguiente, existen diferentes técnicas para llevarlo a cabo. A continuación, revisaremos un breve resumen de las más comunes en los muestreos probabilísticos.

## Muestreo aleatorio simple

Cuando hablamos de un muestreo aleatorio simple debe entenderse un muestreo sin reemplazo. El muestreo aleatorio simple se recomienda cuando las características de interés presentan gran homogeneidad, pues en caso contrario su uso requeriría muestras grandes, para lograr una precisión aceptable. Además, cuando se presenta cierta heterogeneidad en los datos podrían seleccionarse muestras indeseables.

El muestreo aleatorio simple es aquel método que asigna la misma probabilidad de selección a todas y cada una de las muestras posibles y distintas. Siendo esta probabilidad  $1/C_n^N$ , donde  $N$  representa al tamaño de la población y  $n$ , el tamaño de la muestra.

Una forma equivalente de seleccionar la muestra es elegir las unidades de una en una y en forma consecutiva y asignar una probabilidad de selección a las unidades en cada caso.

### Ejemplo 1.9 Muestreo aleatorio simple

De la población estudiantil del Tecnológico de Huatabampo seleccionamos de manera aleatoria una muestra de 10 estudiantes para encuestar y obtener cierta información. En estos casos, para respetar la aleatoriedad podemos llevar a cabo la obtención de la muestra de diferentes formas, la más común consiste en asignar un número diferente a cada estudiante y luego, con la ayuda de una tabla de números aleatorios o un programa generador de éstos, elegir 10 números aleatorios y proceder a realizar las entrevistas a los alumnos seleccionados.

Suponga que contamos a todos los estudiantes de la población del Tecnológico de Huatabampo, y el resultado es 366 estudiantes. Luego, los etiquetamos con los números 0, 1, 2, hasta 365. Ahora, mediante tablas de números aleatorios o un programa generador de éstos. Se generan los 10 números entre 0 y 365, suponga que resultan los números 45, 78, 92, 184, 197, 236, 248, 269, 275 y 291. Es decir, hemos seleccionado a los 10 estudiantes con la técnica de muestreo aleatorio simple.

## Muestreo estratificado

Cuando se tiene una población que puede ser dividida en varias subpoblaciones a las que llamamos estratos, de acuerdo con ciertas propiedades que deben cumplir sus integrantes, pensamos en un muestreo de tipo estratificado, cuando éste cumple estas condiciones:

- La población se divide en subpoblaciones denominadas **estratos**, en las cuales los integrantes de cada uno cumplen ciertas propiedades comunes.
- Seleccionar una muestra en forma independiente de cada estrato. Si las muestras por estrato se eligen con el muestreo aleatorio simple, entonces éste se denomina **muestreo aleatorio estratificado** (este tipo de muestreo es el que comúnmente se utiliza). No hay reglas determinantes para elegir el tamaño de cada estrato, pero se sugiere que sea de forma proporcional a los tamaños de los estratos con respecto al tamaño poblacional.
- Los estimadores para los parámetros de la población completa se proponen como una combinación de los correspondientes a los parámetros de los estratos.

Este método de muestreo es flexible en cuanto a la selección de la muestra en cada estrato. Es válido señalar aquí que los estratos se construyen sin importar que sean geográficamente contiguos o no. Además, tanto el tamaño de la población completa, como el del estrato deben ser conocidos.

El muestreo estratificado es ampliamente usado por varias razones, entre las que destacan:

- Proporciona estimadores (véase la unidad 3) para la población más precisos (esto se logra mediante la construcción de estratos que sean lo más homogéneos posible).
- Proporciona información sobre los estratos.
- Permite una mejor organización del muestreo.
- Permite una mejor administración de la encuesta.
- Permite una mejor administración de la población.
- Este tipo de muestreo se recomienda cuando se desea tener en la muestra representantes de cada subpoblación.

### Ejemplo 1.10 Muestreo estratificado

Suponga que se pide seleccionar una muestra de tamaño 2% de toda la población estudiantil de la UPIICSA, que tiene 12 500 alumnos. La muestra debe cumplir la condición de que exista al menos un representante de cada una de las carreras que se imparten en este centro educativo que tiene las siguientes carreras: administración industrial (4 200), ingeniería industrial (3 250), ingeniería en transporte (850), ingeniería en informática (1 700) y licenciatura en informática (2 500).

Por el entorno del ejemplo, podemos decir que están todas las condiciones para llevar a cabo un muestreo estratificado, donde el tamaño de la muestra es de 250 alumnos (2% de 12 500). El tamaño de la muestra por estrato, se obtiene de esta forma:

$$\text{Administración industrial } \frac{4\,200}{12\,500} \approx 0.336 \Rightarrow n_1 = 0.336 \times 250 = 84$$

$$\text{Ingeniería industrial } \frac{3\,250}{12\,500} \approx 0.26 \Rightarrow n_2 = 0.260 \times 250 = 65$$

$$\text{Ingeniería en transporte } \frac{850}{12\,500} \approx 0.068 \Rightarrow n_3 = 0.068 \times 250 = 17$$

$$\text{Ingeniería en informática } \frac{1\,700}{12\,500} \approx 0.136 \Rightarrow n_4 = 0.136 \times 250 = 34$$

$$\text{Licenciatura en informática } \frac{2\,500}{12\,500} \approx 0.200 \Rightarrow n_5 = 0.200 \times 250 = 50$$

Se cumple que  $n = n_1 + n_2 + n_3 + n_4 + n_5 = 84 + 65 + 17 + 34 + 50 = 250$

## Muestreo sistemático con iniciación aleatoria

El método de muestreo con iniciación aleatoria es un método de muestreo probabilístico que simplifica la selección de una muestra. En este caso, la primera unidad se selecciona en forma aleatoria y los restantes elementos, para formar la muestra del tamaño requerido, se toman siguiendo un patrón establecido. Las ventajas del muestreo sistemático son:

- Es más fácil de realizar en el campo y aun en la oficina.
- Se eliminan errores de los enumeradores, en especial cuando se tiene un marco de muestreo defectuoso.
- Extiende la muestra a toda la población, se distribuye mejor y de manera uniforme sobre la población.
- No precisa la distinción entre muestreo sin reemplazo y con reemplazo.

- Recoge el posible efecto de la estratificación debido al orden en que figuran las unidades en la población.
- Si la disposición de las unidades en la población es aleatoria, la selección sistemática equivale a un muestreo aleatorio simple.

Este tipo de muestreo es propicio para realizar estudios como:

1. En una línea de producción que esté trabajando en forma continua, se puede hacer un muestreo de tamaño determinado cada 200 unidades.
2. En la línea de producción anterior el muestreo puede llevarse a cabo cada determinado tiempo. Por ejemplo, cada hora se selecciona una muestra para su análisis.
3. En el estudio de árboles de un bosque, en el que podemos establecer un patrón de revisión, elegir el primero y después seleccionar un árbol de cada 100 para su estudio.
4. Cuando se requiere llevar a cabo encuestas a los usuarios del metro, el mejor muestreo es el sistemático.

## Muestreo por conglomerados

Este tipo de muestreo, en cierta forma, es similar al estratificado, puesto que la población se divide en subpoblaciones (estratos), pero a diferencia del estratificado en éste no se requiere un representante de cada estrato en la muestra, ya que en primer lugar elegimos una muestra de estratos y, en segundo, seleccionamos una muestra de cada uno para conformar la muestra deseada.

El ejemplo del muestreo por conglomerados es de una etapa, pero en general existen conglomerados de varias etapas.

El muestreo por conglomerados se usa en poblaciones en extremo grandes, y a diferencia de las técnicas mencionadas antes, no requiere de un marco de muestreo que liste las unidades con anterioridad. Proporciona un mayor ahorro de recursos que con cualquiera de las anteriores, pero se pierde precisión. Además, se usa cuando las unidades se encuentran muy dispersas geográficamente.

### Ejemplo 1.11 Muestreo por conglomerados

Suponga que se quiere llevar a cabo una encuesta de los usuarios del metro de la Ciudad de México (alrededor de cinco millones de usuarios diarios). Como la población en estudio es demasiado grande, podemos dividir en estratos; por ejemplo, estaciones del metro. Después, elegimos una muestra de las estaciones y procedemos a realizar la encuesta a los usuarios en las estaciones seleccionadas (puede ser con el muestreo sistemático). Esta forma de muestreo disminuye considerablemente el costo de la muestra ya que no se requiere numerar con anterioridad las unidades poblacionales.

## Tamaño de la muestra

En general, los investigadores y estudiantes de áreas aplicadas requieren conocer el tamaño ideal de una muestra para realizar el estudio de campo de la investigación que llevan a cabo. Pero, de manera errónea se piensa que existe una fórmula mágica para calcular el valor deseado de muestra que sea posible aplicar a cualquier situación o investigación. Al igual que el muestreo, en el que existen diferentes técnicas para determinar el tamaño de la muestra, también hay diferentes situaciones. Pero, siempre debe tenerse en cuenta que las muestras deben cumplir estas características:

- **Representativa.** Todos y cada uno de los elementos de la población deben tener la misma oportunidad de ser tomados en cuenta para conformar la muestra.
- **Adecuada y válida.** El error de la muestra debe ser el mínimo posible respecto de la población.
- **Confiable.** El tamaño de la muestra debe obtenerse mediante algún proceso matemático que elimine la incidencia del error.

Podemos establecer que el cálculo del tamaño de la muestra resulta ser uno de los aspectos clave en la fase previa de cualquier investigación científica o de mercado, ya que con ésta es posible determinar el grado de credibilidad

que podemos asignar a los resultados de la investigación. Además, al elegir un buen tamaño de muestra y una técnica adecuada de muestreo, implícitamente hemos reunido información que cumple las características enunciadas antes para una muestra: representativa, válida y confiable a un costo mínimo.

Resolver el problema del tamaño no resulta tan simple como algunos lectores lo imaginen. En general, se requiere de conocimientos previos de temas estadísticos que estudiaremos más adelante, como los intervalos de confianza (unidad 3) y la prueba de hipótesis (unidad 4), además de los objetivos de estudio y características de la población de interés. Debido a esto, la gran mayoría de textos de estadística no presentan fórmulas o métodos para calcular el tamaño de la muestra.

En esta subsección mostramos las fórmulas básicas para determinar el tamaño de la muestra, pero haciendo énfasis en que el problema no quedará resuelto con las fórmulas que serán revisadas en esta sección. Cualquier fórmula para calcular el tamaño de la muestra debe tener tres factores:

1. Un porcentaje de confianza con el que se desea generalizar los datos de la muestra a la población. Éste lo representaremos por  $1 - \alpha$  y está en estrecha relación con los intervalos de confianza (unidad 3). En el tamaño de la muestra se utiliza para calcular  $Z_{1-\alpha}$ . En general, se considera  $1 - \alpha = 0.95$ , en este caso  $Z_{1-\alpha} = 1.96$  (véase uso de tablas de la distribución normal en la unidad 2).
2. Un porcentaje de error permitido para aceptar la generalización. Lo representaremos por  $\varepsilon$  y está estrechamente relacionado con las pruebas de hipótesis (unidad 4). Suele tomar valores entre 0 y 0.10, mientras más pequeño el tamaño de muestra, aumenta el porcentaje de error.
3. Nivel de variabilidad, la probabilidad con la que se presentan los fenómenos de estudio, los valores serán denotados por  $p$  y  $q = 1 - p$ , en donde  $p$  es el porcentaje de confiabilidad. Cuando no se tienen antecedentes sobre la investigación (no hay otras o no se pudo aplicar una prueba previa), entonces el valor de variable se considera máximo, este valor resulta cuando  $p = q = 0.5 = \sigma$ .

Para tener una comprensión total del término variabilidad se requieren conocimientos de la unidad 4 sobre las pruebas de hipótesis, ya que una definición formal de la variabilidad sería: La **variabilidad** es la probabilidad (o porcentaje) con el que se aceptó y se rechazó la hipótesis que se quiere investigar con base en alguna investigación anterior o en un ensayo previo. El porcentaje con que se aceptó esa hipótesis se denomina variabilidad positiva y se denota por  $p$ , y el porcentaje con el que se rechazó es la variabilidad negativa, denotada por  $q = 1 - p$ .

### Caso 1. Tamaño de la muestra cuando no se conoce $N$ o la población es infinita.

Cuando no conocemos el tamaño de la población y las observaciones presentan normalidad, el tamaño de la muestra para estimar la media se puede calcular con la fórmula 1.1

$$n \geq \frac{p(1-p)Z_{1-\alpha}^2}{\varepsilon^2} \quad (1.1)$$

Donde,  $n$  es el tamaño de la muestra;  $Z_{1-\alpha}$  es el valor de tablas de la distribución normal estándar para una probabilidad central de  $1 - \alpha$ ;  $\varepsilon$  es el error muestral permitido y  $p$  es la variabilidad positiva (confiabilidad).

#### Ejemplo 1.12 Tamaño de la muestra

El ingeniero de control de calidad de una línea de producción de envases de refresco debe tomar una muestra diaria de la línea de producción para inspeccionarla. Decide que su estudio tenga una confianza de 96% y permitir un error de 5%. Calcule el tamaño de la muestra para inspeccionar la producción.

- a) Cuando inicia y no tiene información previa.
- b) Cuando tiene la información de varias revisiones diarias, con las que se ha obtenido una variabilidad positiva de 0.75.

**Solución**

Ambos casos tienen en común 96% de confianza  $Z_{1-\alpha} = 2.054$  y  $\varepsilon = 0.05$

- a) Como no se tiene información previa sobre la investigación, entonces consideramos la variabilidad  $p = 0.5$ . Si se sustituye en la fórmula 1.1:

$$n \geq \frac{(0.5)(0.5)(2.054)^2}{(0.05)^2} = 421.89 \Rightarrow n = 422$$

- b) En este inciso tenemos información previa sobre la investigación con una confiabilidad  $p = 0.75$ . Si se sustituye en la fórmula 1.1:

$$n \geq \frac{(0.75)(0.25)(2.054)^2}{(0.05)^2} = 316.41 \Rightarrow n = 317$$

Como era de esperarse, se requiere un tamaño de muestra mayor cuando no se tiene información previa.

**Caso 2. Tamaño de la muestra cuando se conoce el tamaño de la población  $N$ .**

Cuando conocemos el tamaño de la población y las observaciones presentan normalidad, el tamaño de la muestra para estimar la media se calcula con la fórmula 1.2:

$$n \geq \frac{Np(1-p)Z_{1-\alpha}^2}{(N-1)\varepsilon^2 + p(1-p)Z_{1-\alpha}^2} \quad (1.2)$$

Donde  $n$  es el tamaño de la muestra;  $N$  es el tamaño de la población;  $Z_{1-\alpha}$  es el valor de tablas de la distribución normal estándar para una probabilidad central de  $1 - \alpha$ ;  $\varepsilon$  es el error muestral permitido,  $p$  es la variabilidad positiva. La fórmula 1.2 se obtiene del intervalo:

$$\bar{x} - Z \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \leq \mu \leq \bar{x} + Z \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

**Ejemplo 1.13 Tamaño de la población**

El ingeniero de control de calidad de cinescopios debe revisar lotes de 2000 artículos cada uno. Él decide que su estudio tenga una confianza de 95% y permitir un error de 8%. Calcule el tamaño de la muestra para inspeccionar un lote.

- a) Cuando inicia y nunca ha revisado un lote; es decir, no tiene información previa.  
 b) Cuando se han realizado varias revisiones de lotes del mismo tamaño y se ha obtenido una variabilidad positiva de 0.8.

**Solución**

Ambos casos tienen en común 95% de  $Z_{1-\alpha} = 1.96$ ,  $N = 2000$  y  $\varepsilon = 0.08$ .

- a) Como en este caso no se tiene información previa sobre la investigación, entonces consideramos la variabilidad máxima  $p = 0.5$ . Si se sustituye en la fórmula 1.2:

$$n \geq \frac{2000(0.5)(0.5)(1.96)^2}{(2000-1)(0.08)^2 + (0.5)(0.5)(1.96)^2} = 139.65 \Rightarrow n = 140$$

- b) Como aquí tenemos información previa sobre la investigación con una variabilidad  $p = 0.8$ . Si se sustituye en la fórmula 1.2:

$$n \geq \frac{2000(0.8)(0.2)(1.96)^2}{(2000 - 1)(0.08)^2 + (0.8)(0.2)(1.96)^2} = 91.68 \Rightarrow n = 92$$

Como era de esperarse, se requiere un tamaño de muestra mayor cuando no se tiene información previa.

### Caso 3. Tamaño de muestra cuando los datos son cualitativos.

Cuando se conoce el tamaño de la población, éste es finito y los datos son cualitativos. Por ejemplo, en fenómenos sociales en los que se utiliza la escala nominal para comprobar la ausencia o presencia del o de los atributos a estudiar, el tamaño de muestra para estimar la proporción se calcula con la fórmula 1.3:

$$n \geq \frac{P^*}{1 + P^*/N} \quad (1.3)$$

Donde,  $n$  es el tamaño de la muestra,  $N$  es el tamaño de la población;  $P^* = \frac{p(1-p)}{(se)^2}$  con  $se$ , error estándar y  $p$ , la variabilidad positiva o porcentaje de confiabilidad.

#### Ejemplo 1.14 Datos cualitativos

Suponga que se tiene una población de 2 150 personas que se van a entrevistar para conocer si están a favor de un candidato para presidente municipal. Se pretende conocer la aceptación hacia el candidato y el estudio se realizará a través de una muestra. ¿Cuál deberá ser el tamaño mínimo de muestra que cumpla con un error estándar de 0.02 y una confiabilidad de 90%?

#### Solución

Podemos notar que se trata de un problema de variables cualitativas, el atributo es que están o no a favor del candidato. Donde,  $N = 2150$ ,  $se = 0.02$ ,  $p = 0.90$ , entonces  $P^* = \frac{0.90(0.10)}{(0.02)^2} = 225$

$$n \geq \frac{225}{1 + 225/2150} = 203.68 \Rightarrow n = 204$$

### Caso 4. Disminución del tamaño de la muestra cuando $N$ es pequeña.

En los casos 2 y 3 se consideró que es conocido el tamaño de la población, fórmulas 1.2 y 1.3. Al emplear las fórmulas mencionadas, si el tamaño de la población no es muy grande, tendremos que éste representará un gran porcentaje de  $N$ . Para disminuir el tamaño muestral se puede hacer una corrección. Denote por  $n^*$  al tamaño de muestra calculado con alguna de las fórmulas, el cual se disminuye utilizando la fórmula 1.4:

$$n \geq \frac{n^*N}{n^* + N} \quad (1.4)$$

Donde,  $n$  es el tamaño nuevo de la muestra,  $N$  es el tamaño de la población;  $n^*$  es el tamaño previo de la muestra.

#### Ejemplo 1.15 Disminución del tamaño de muestra

El ingeniero de control de calidad de cinescopios debe revisar lotes de 200 artículos cada uno. Decide que su estudio tenga una confianza de 95% y permitir un error de 5%. Calcule el tamaño de la muestra para inspeccionar un lote cuando va iniciando y nunca ha revisado un lote, es decir, no tiene información previa. Después, disminuya el tamaño de la muestra con la corrección.

**Solución**

Tenemos 95% de confianza de  $Z_{1-\alpha} = 1.96$ ,  $N = 200$   $\varepsilon = 0.05$ , si se sustituye en la fórmula 1.2:

$$n^* \geq \frac{200(0.5)(0.5)(1.96)^2}{(200-1)(0.05)^2 + (0.5)(0.5)(1.96)^2} = 131.75 \Rightarrow n = 132$$

Como el tamaño de la población es pequeño, el tamaño de la muestra representa un porcentaje elevado de  $N$  (66%). Entonces, es posible disminuir el tamaño de la muestra con la fórmula 1.4:

$$n \geq \frac{132(200)}{132 + 200} = 79.5 \Rightarrow n = 80$$

Este tamaño de muestra es menor al obtenido con la fórmula 1.2 y representa 40% de  $N$ .

Si las poblaciones son homogéneas (la característica de interés es poco variable) no existe problema para realizar o diseñar un muestreo, ya que basta con un muestreo pequeño. Sin embargo, si la población es heterogénea, entonces se tendrán ciertas dificultades para decidir sobre qué tipo de muestreo debe emplearse. De acuerdo con esto, podemos decir que es necesario preparar gente que sea capaz de muestrear en poblaciones heterogéneas.

## 1.4 Parámetros y estadísticos

Los números que sintetizan los aspectos más relevantes de una distribución estadística pueden obtenerse, tanto de una población como de una muestra, por consiguiente, el investigador tiene la obligación de clasificarlos.

Estos números reciben el nombre de **parámetros** y cuando son obtenidos de una muestra se llaman **estadísticos**. En las variables aleatorias estudiamos tres tipos de parámetros que fueron utilizados para describir una variable aleatoria, localidad, escala y forma. En la parte estadística vamos a ampliar este concepto a operaciones entre variables, como la suma y el promedio, entre otras.

Para un estudio más detallado de los parámetros y estadísticos, véase la unidad 2 sobre distribuciones muestrales.

Ahora bien, ¿qué es un parámetro?, ¿qué es un estadístico?

Los parámetros y estadísticos más comunes de la estadística descriptiva que estudiaremos en esta unidad se clasifican en dos tipos:

1. Medidas centrales: media, mediana, moda, media geométrica, media armónica, media ponderada.
2. Medidas de dispersión: rango, varianza y desviación estándar.

## 1.5 Medidas centrales

Si el conjunto de datos numéricos de una muestra de tamaño  $n$  (o población de tamaño  $N$ ) es de la forma  $x_1, x_2, \dots, x_n$  (o para la población  $x_1, x_2, \dots, x_N$ ). Podemos preguntar, ¿qué características del conjunto de números son de interés?

En esta sección discutiremos los métodos para describir su localización y en particular el centro de los datos.

### La media

Cuando una persona tiene en sus manos un conjunto de datos para analizarlos, en general una de sus primeras inquietudes consiste en calcular su promedio.

#### Ejemplo 1.16

El señor Luis Martínez tiene las cantidades mensuales que ha ganado en el último medio año (\$10 800, \$9 700, \$11 100, \$8 950, \$9 750 y \$10 500) y desea conocer un valor que represente al salario promedio durante este tiempo.



En este caso, el señor Luis calculará su ingreso promedio al sumar los sueldos y dividir entre la cantidad de meses:

$$\frac{10\,800 + 9\,700 + 11\,100 + 8\,950 + 9\,750 + 10\,500}{6} = \$10\,133.33$$

De esta forma, el sueldo promedio de los últimos seis meses del señor Luis es de \$10 133.33.

Así como el problema anterior, existe una infinidad de casos prácticos sencillos donde, dado un conjunto de datos es de interés conocer un valor central que refleje la influencia que tiene cada uno de los valores de las observaciones en el valor central. La medida central más propicia para estos fines se define a continuación.

Dado el conjunto finito de datos muestrales  $x_1, x_2, \dots, x_n$ , se llama **media muestral** (promedio aritmético) o estadístico media del conjunto, al valor que representa el promedio de los datos, y se simboliza por  $\bar{x}$  (x barra o x testada) y se calculará por:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i \quad (1.5)$$

En la unidad 2 se verá una definición más precisa de estadístico media.

En la unidad se cita una definición más general para  $\mu$ , la cual se puede aplicar tanto a poblaciones finitas como infinitas. La definición que aquí se trata solo se refiere a las poblaciones finitas.

De forma similar, se simboliza con la letra griega miu ( $\mu$ ) al parámetro media para las poblaciones finitas,  $x_1, x_2, \dots, x_N$  y llamaremos media poblacional o parámetro media del conjunto a:

$$\mu = \frac{x_1 + x_2 + \dots + x_N}{N} = \frac{1}{N} \sum_{i=1}^N x_i$$

La medida que se defina de aquí en adelante para una población será la misma que para una muestra, cambiando  $n$  (tamaño muestral) por  $N$  (tamaño poblacional). Por esas razones omitiremos la definición de la medida para la población.

A continuación, se ilustra la definición de media muestral por medio de dos ejemplos.

### Ejemplo 1.17 Media muestral

Un fabricante de pistones toma una muestra aleatoria de 20 para medir su diámetro interno promedio. Los diámetros, en centímetros, que el fabricante obtuvo están dados a continuación. Calcule el diámetro medio de dichos pistones (véase tabla 1.1).

Tabla 1.1

10.1	10.1	9.8	9.7	10.3	9.9	10.0	9.9	10.2	10.1
9.9	9.9	10.1	10.3	9.8	9.7	9.9	10.0	10.0	9.8

#### Solución

Como se trata de una muestra, utilizamos la fórmula de la definición de media muestral.

$$\begin{aligned} \bar{x} &= \frac{1}{20} [10.1 + 10.1 + 9.8 + 9.7 + 10.3 + 9.9 + 10.0 + 9.9 + 10.2 + 10.1 + \\ &\quad + 9.9 + 9.9 + 10.1 \\ &\quad + 10.3 + 9.8 + 9.7 + 9.9 + 10.0 + 10.0 + 9.8] \\ &= 9.975 \end{aligned}$$

La **media** representa un valor promedio de todas las observaciones, por consiguiente, cada uno de los datos influye de igual forma en su resultado. Por tanto, cuando se tienen datos aberrantes, que se alejan de manera considerable del resto de los demás valores, el valor promedio encontrado no refleja la realidad del caso (véase ejemplo 1.18).

### Ejemplo 1.18 La media

Suponga que se quiere estimar el sueldo promedio de los trabajadores de una fábrica, al elegir de manera aleatoria a 10 de todos los trabajadores y obtener las observaciones de la tabla 1.2.

**Tabla 1.2**

Dato	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$	$x_8$	$x_9$	$x_{10}$
Sueldo	2000	2200	2500	2200	1800	25000	2400	2300	2800	2400

Si se calcula el sueldo promedio obtenemos:

$$\begin{aligned}\bar{x} &= \frac{1}{10} [2000 + 2200 + 2500 + 2200 + 1800 + 25000 + 2400 + 2300 + 2800 + 2400] \\ &= 4560\end{aligned}$$

Donde, obviamente el estadístico no refleja una realidad de los datos, puesto que el sueldo de \$25 000 es mucho mayor (sueldo aberrante) que los sueldos restantes, influyendo considerablemente en el valor promedio.

En situaciones como la anterior, el uso del valor promedio no es tan acertado de manera que se suele recurrir a otra medida de tipo central.

## La mediana

De lo expuesto al final de la subsección anterior, comprendemos la necesidad de introducir otro tipo de medida central con la que los valores extremos, con respecto al resto, no tenga una influencia tan marcada como en la media. Debido a su naturaleza, a esta medida se le conoce con el nombre de mediana y la definiremos a continuación.

La **mediana** de un conjunto de datos es el valor central de los datos cuando estos se han ordenado en forma no decreciente en cuanto a su magnitud.

### Cálculo de la mediana

Sea el conjunto de datos muestrales  $x_1, x_2, \dots, x_n$ , la mediana muestral o estadístico mediana del conjunto la denotamos por  $\tilde{x}$  ( $x$  tilde) y se obtiene al ordenar primero en forma no decreciente los  $n$  datos, renombrando según su posición por medio de tildes de la siguiente forma:

$$\tilde{x}_1 \leq \tilde{x}_2 \leq \dots \leq \tilde{x}_n$$

Enseguida, localizamos el punto medio de los datos ordenados pudiendo ocurrir alguno de los siguientes dos casos:

- Cuando la cantidad de observaciones es impar: el punto medio del ordenamiento es el dato que se encuentra en la posición  $\frac{n+1}{2}$ .
- Cuando la cantidad de datos es par: en este caso resultan dos datos medios localizados en las posiciones  $\frac{n}{2}$  y  $\frac{n}{2} + 1$ , por lo que la mediana se considera el promedio de estos datos medios.

Por último, el cálculo de la mediana se resume con la siguiente fórmula:

$$\tilde{x} = \begin{cases} \frac{\tilde{x}_{\frac{n+1}{2}}}{2}, & \text{cuando la cantidad de datos es impar.} \\ \frac{\tilde{x}_{\frac{n}{2}} + \tilde{x}_{\frac{n}{2}+1}}{2}, & \text{cuando la cantidad de datos es par.} \end{cases} \quad (1.6)$$

En el siguiente ejemplo mostramos el cálculo de la mediana.

### Ejemplo 1.19 Cálculo de la mediana

Sea el conjunto muestral de datos del ejemplo anterior referente a los sueldos promedios de los salarios. Encuentre la mediana de los salarios.

#### Solución

La tabla 1.3 muestra el conjunto de los 10 datos:

Tabla 1.3

Dato	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$	$x_8$	$x_9$	$x_{10}$
Sueldo	2 000	2 200	2 500	2 200	1 800	25 000	2 400	2 300	2 800	2 400

Al ordenar los salarios en forma no decreciente y renombrarlos obtenemos:

$$1\,800 \leq 2\,000 \leq 2\,200 \leq 2\,200 \leq 2\,300 \leq 2\,400 \leq 2\,400 \leq 2\,500 \leq 2\,800 \leq 25\,000$$

Tabla 1.4

Datos originales	$x_5$	$x_1$	$x_2$	$x_4$	$x_8$	$x_7$	$x_{10}$	$x_3$	$x_9$	$x_6$
Sueldos	1 800	2 000	2 200	2 200	<b>2 300</b>	<b>2 400</b>	2 400	2 500	2 800	25 000
Datos ordenados	$\tilde{x}_1$	$\tilde{x}_2$	$\tilde{x}_3$	$\tilde{x}_4$	$\tilde{x}_5$	$\tilde{x}_6$	$\tilde{x}_7$	$\tilde{x}_8$	$\tilde{x}_9$	$\tilde{x}_{10}$

La cantidad de datos es 10, este es un número par. Entonces,

la *mediana muestral* se calcula con el promedio de los datos ordenados en las *posiciones*  $\frac{n}{2} = \frac{10}{2} = 5$  y  $\frac{n}{2} + 1 = \frac{10}{2} + 1 = 6$ .

Es decir,

$$\tilde{x} = \frac{\tilde{x}_5 + \tilde{x}_6}{2} = \frac{2\,300 + 2\,400}{2} = 2\,350$$

Podemos observar que el valor \$25 000 que sobresalía con respecto a todos los demás sueldos no influye en la mediana. Pues, si en lugar de \$25 000 elegimos \$5 000 o \$100 000, el sueldo medio de los 10 trabajadores seguirá siendo \$2 350. Por esta razón, decimos que la mediana es una medida central insensible a los cambios.

## La moda

En algunos estudios es necesario encontrar el valor central de un conjunto de datos, en el cual la medida de interés está basada en su repetición. Por esta razón, no es conveniente usar ninguna de las dos medidas vistas. Debido a su naturaleza, a la medida a la que hacemos referencia se le denomina **moda** y la definimos a continuación.

La **moda** de un conjunto de datos es el valor de éstos que se presenta en su distribución con mayor frecuencia.

Con respecto a la notación de la moda, a diferencia de las dos medidas centrales anteriores, no existe notación estándar, por lo que empleamos la letra  $M$  para la muestra.

### Ejemplo 1.20 La moda

En la lista se muestran las calificaciones de 20 exámenes de lingüística. Encuentre la calificación que más se repite, es decir, la moda de la distribución de las calificaciones.

Tabla 1.5

5	8	9	9	8	10	9	5	10	5
6	5	10	10	8	9	7	9	5	9

#### Solución

Si se realiza un conteo de los datos podemos verificar que resultan:

- Cinco datos con valor 5.
- Un dato con valor 6 y otro con valor 7.
- Tres datos con valor 8.
- Seis datos con valor 9.
- Cuatro datos con valor 10.

Por último, la moda es igual a 9; la calificación que se repite más veces.

Al calcular la moda podemos observar que se trata de una medida completamente opuesta a la mediana en cuanto a su sensibilidad. Si en el ejemplo anterior un alumno con calificación 9 hubiera obtenido 5, ¡la moda cambiaría a 5! (serían seis 5 y cinco 9). Como se puede notar con la alteración de un solo dato se modificó por completo la moda. Entonces, se dice que ésta es una medida sensible a los cambios.

La moda presenta los siguientes problemas:

- Puede no existir.

Por ejemplo, al calcular la moda de las muestras:

Muestra 1: 6, 7, 9, 4, 8.

Muestra 2: 6, 3, 8, 9, 3, 8, 6 y 9.

Resulta que en ambas muestras los datos se repiten la misma cantidad de veces, es decir, no tienen moda. En tales situaciones, a la muestra se le llama amodal o sin moda. En este sentido, cabe preguntarnos: ¿cuándo un conjunto de datos es amodal?

- Puede no ser única.

Por ejemplo, la moda de la muestra:

6, 7, 9, 4, 8, 6, 6, 8, 9, 6, 8, 6, 9, 3, 9 y 9

tiene al 6 y al 9 con mayor frecuencia, puesto que ambos se repiten cinco veces.

Cuando el conjunto de datos tiene más de una moda se llama multimodal: bimodal si son dos modas, trimodal si son tres, etcétera.

## Otros valores medios

Hasta ahora se han estudiado los tres valores centrales más conocidos y utilizados en la estadística descriptiva. El primero fue el valor medio definido como una media aritmética, pero se comentó que existen distribuciones de datos para las que dicha medida no es muy propicia. Por consiguiente, se recurre a otras medidas de tipo central.

Dos de estas fueron la mediana y la moda, ahora se verán otros tipos de medidas que en muchas ocasiones son de gran utilidad en la estadística descriptiva.

- 1. Valor geométrico o media geométrica.** La media geométrica de los datos  $x_1, x_2, \dots, x_n$  se simbolizará por  $MG$ , está definida como la raíz  $n$ -ésima del producto de las  $n$  mediciones.

$$MG = \sqrt[n]{x_1 \cdot x_2 \cdots x_n} \quad (1.7)$$

### Ejemplo 1.21 Valor geométrico o media geométrica

Calcule la media geométrica de las 20 calificaciones de los exámenes psicológicos que se muestran en la tabla 1.6.

**Tabla 1.6**

5	8	9	9	8	10	9	5	10	5
6	5	10	10	8	9	7	9	5	9

De la definición de media geométrica se deduce con facilidad que ésta no se puede aplicar cuando algún dato vale cero o la cantidad de datos es par y existen algunos de estos negativos, sin embargo, tiene cierta aplicación en la psicofísica.

#### Solución

Si se emplea la fórmula 1.7 tenemos:

$$\begin{aligned} MG &= \sqrt[20]{5 \times 8 \times 9 \times 9 \times 8 \times 10 \times 9 \times 5 \times 10 \times 5 \times 6 \times 5 \times 10 \times 10 \times 8 \times 9 \times 7 \times 9 \times 5 \times 9} \\ &= 7.5446868 \end{aligned}$$

Otra aplicación importante de la media geométrica se presenta en las tasas de interés al considerar su factor de crecimiento medio, entendido como:

$$\text{Factor de crecimiento} = 1 + \frac{\text{tasa de interés}}{100}$$

Entonces, el factor de crecimiento medio es un valor medio de estos factores, y la mejor medida central que debe utilizarse es la media geométrica.

#### Explicación

En las economías emergentes, durante los periodos de crisis se registra un alto índice de inflación y los bancos deben pagar altas tasas de interés para atraer a los ahorradores. Vamos a suponer que tenemos en un periodo de cuatro años, en tiempo de crisis, las tasas de interés anual de 100, 200, 250 y 350%. Es decir, tenemos los factores de crecimiento, 2, 3, 3.5 y 4.5, respectivamente. Queremos conocer cuánto crecerá un depósito inicial de \$1 000 a cuatro años.

En estas condiciones el banco pagará a cuatro años  $\left(\left(\left(1000 \times 2\right) \times 3\right) \times 3.5\right) \times 4.5 = 94500$ :

$$1000 \times 2 = 2000 \text{ primer año}$$

$$2000 \times 3 = 6000 \text{ segundo año}$$

$$6000 \times 3.5 = 21000 \text{ tercer año}$$

$$21000 \times 4.5 = 94500 \text{ cuarto año}$$

Es decir, al término del cuarto año el banco tendría que pagar \$94 500 por la inversión de \$1 000. Pero, qué pasa si dicho monto se quiere calcular con un factor de crecimiento medio y utilizamos la media:

$$\text{Factor de crecimiento} = \frac{2 + 3 + 3.5 + 4.5}{4} = 3.25$$

Luego, el monto a pagar por el banco con este promedio sería ( $1000 \times 3.25^4 = 111566.41$ ):

$$1000 \times 3.25 = 3250 \text{ primer año}$$

$$3250 \times 3.25 = 10562.5 \text{ segundo año}$$

$$10562.5 \times 3.25 = 34328.125 \text{ tercer año}$$

$$34328.125 \times 3.25 = 111566.41 \text{ cuarto año}$$

Valor que difiere del real \$94500.00.

Por otro lado, si el valor medio del factor de crecimiento lo calculamos con la media geométrica y realizamos los cálculos:

$$\text{Factor de crecimiento medio geométrico} = \sqrt[4]{2 \times 3 \times 3.5 \times 4.5} = 3.11787$$

Entonces, el monto a pagar por el banco con este promedio sería  $1000 \times 3.11787^4 = 94500$ , que corresponde al valor real del pago del banco.

**2. Valor medio armónico o media armónica.** La media armónica de los datos  $x_1, x_2, \dots, x_n$  se denota por  $MA$  y se define como el recíproco de la media aritmética de los recíprocos.

$$MA = \frac{1}{\frac{1}{n} \sum_{i=1}^n \frac{1}{x_i}} = \frac{1}{\frac{1}{n} \left[ \frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n} \right]} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}} \quad (1.8)$$

Las principales aplicaciones de esta media se basan en promediar variaciones respecto del tiempo, es decir, cuando la misma distancia se recorre a diferentes tiempos.

Por la naturaleza de la definición, esta medida central tiene su mayor uso en física e ingeniería cuando se trabaja con engranes o poleas para determinar velocidades promedio de recorridos.

### Ejemplo 1.22 Media armónica

Suponga que viaja de una ciudad a otra y recorre los primeros 100 km a 80 km/h, los siguientes 100 km a una velocidad de 100 km/h y los otros 100 km a una velocidad de 120 km/h. Calcule la velocidad media realizada con la media armónica y compárela con las medias aritmética y geométrica.

#### Solución

Si se utilizan las fórmulas 1.5, 1.7 y 1.8 tenemos:

$$\bar{x} = \frac{1}{3} [80 + 100 + 120] = \frac{300}{3} = 100$$

$$MG = \sqrt[3]{80 \times 100 \times 120} = 98.6485$$

$$MA = \frac{1}{\frac{1}{3} \left[ \frac{1}{80} + \frac{1}{100} + \frac{1}{120} \right]} = 97.2973$$

#### Observación

Para tomar una decisión de qué media parece ser la más correcta, calculamos la velocidad promedio:

$$\text{Velocidad promedio} = \frac{\text{Distancia total recorrida}}{\text{Tiempo total}}$$

Distancia total recorrida es igual a  $100 + 100 + 100 = 300$  km.

$$\text{Tiempo total de recorrido } \frac{100}{80} + \frac{100}{100} + \frac{100}{120} = 3.0833 \text{ horas.}$$

Ahora, comparamos con la distancia total real recorrida. Es decir, se comparan las distancias que presumiblemente recorrería el automóvil con cada una de las velocidades promedio calculadas:

$$\text{Media aritmética: } 3.0833 \times 100 = 308.33 \text{ km}$$

$$\text{Media geométrica: } 3.0833 \times 98.6485 = 304.166 \text{ km}$$

$$\text{Media armónica: } 3.0833 \times 97.2973 = 300 \text{ km}$$

Note que el mejor resultado se obtiene con la media armónica.

**3. Valor medio ponderado o media ponderada.** En los casos en que cada dato tiene una importancia relativa llamada peso (véase definición), la media más apropiada se obtiene sumando los productos de cada dato por su peso; esta medida se llama media ponderada.

Dado un conjunto de datos  $x_1, x_2, \dots, x_n$  se llama pesos o ponderaciones, a las cantidades  $w_1 + w_2 + \dots + w_n = 1$  que cumplen:

- a)  $w_i \in [0, 1]$  para todo valor de  $i$ .
- b)  $w_1 + w_2 + \dots + w_n = 1$

La media ponderada del conjunto de datos  $x_1, x_2, \dots, x_n$ , con pesos respectivos  $w_1, w_2, \dots, w_n$ , la denotaremos por  $MP$  y se calcula por medio de:

$$MP = \sum_{i=1}^n w_i x_i \quad (1.9)$$

A continuación, se muestra un ejemplo para calcular la media ponderada.

### Ejemplo 1.23 Media ponderada

Calcule la calificación promedio de un estudiante del Instituto Tecnológico de Celaya en la materia de fundamentos de física, si la calificación está ponderada de la siguiente forma: 10% tareas, 40% laboratorio y 50% de teoría. Suponga que las calificaciones del estudiante fueron 8, 9 y 4, respectivamente.

#### Solución

La calificación ponderada se calcula con la fórmula 1.9

$$MP = 0.1 \times 8 + 0.4 \times 9 + 0.5 \times 4 = 6.4$$

Al realizar un estudio del conjunto de datos cabe preguntarse si el conocimiento de sus medidas centrales es suficiente para reconocer la distribución de dichos datos. Respuesta que puede darse solo después de estudiar la sección 1.6.

## Ejercicios 1.2

1. Obtenga la media, mediana, moda del siguiente conjunto de datos.

150 150 165 155 155 145 150 140 145 150 160 175 145 160



2. Calcule la media y mediana de los tiempos entre llegadas de seis aviones al aeropuerto Benito Juárez, de la Ciudad de México, cuyos tiempos, en minutos, son:

3.5 4.2 2.9 3.8 4.0 2.8.

3. Calcule la media armónica del viaje redondo que realiza el chofer de una línea de autobuses al dirigirse de México a Acapulco (460 km). Si de ida viajó por la Autopista del Sol a una velocidad de 90 km/h y de regreso por otra carretera a una velocidad promedio de 60 km/hora.
4. Calcule la media geométrica del conjunto de datos del ejercicio 2.
5. En una muestra de 100 pistones se encontró que 55 tenían un diámetro interno de 10.5 cm, 25 de 10.0 cm y el restante de 10.75 cm. Utilice las frecuencias de los diámetros internos de los pistones para determinar la media ponderada de su diámetro interno.
6. En los envases de leche de chocolate de una prestigiosa marca, la cantidad de líquido no es siempre un litro, se toma una muestra de 10 paquetes, obteniéndose las mediciones siguientes, en litros:

0.97 1.01 0.97 0.95 1.0 0.95 0.95 1.01 0.95 0.98.

Calcule la cantidad promedio de leche en los envases de la muestra.

7. Sean las calificaciones de 30 estudiantes en la materia de probabilidad y estadística descriptiva que se muestran en la tabla 1.7.

**Tabla 1.7**

27	72	83	15	96	30	8	98	86	5	39	86	87	100	56
88	31	3	30	57	22	7	20	62	95	35	73	66	56	57

Calcule la media, mediana y moda de las calificaciones.

8. La Bolsa Mexicana de Valores registró diferentes alzas y bajas en puntos porcentuales durante la primera quincena de junio de 2015 (véase tabla 1.8).

**Tabla 1.8**

3.4%	1.7%	-0.5%	0.7%	-2.4%	-1.8%	-0.9%	2.5%	0.3%	0.8%
------	------	-------	------	-------	-------	-------	------	------	------

Considerando solo los porcentajes, calcule el porcentaje medio obtenido en dicha quincena para la bolsa de valores.

9. Sean  $x_1, x_2, \dots, x_n$  los valores muestrales y  $\bar{x}$  su media, pruebe que el siguiente promedio siempre vale cero

$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}).$$

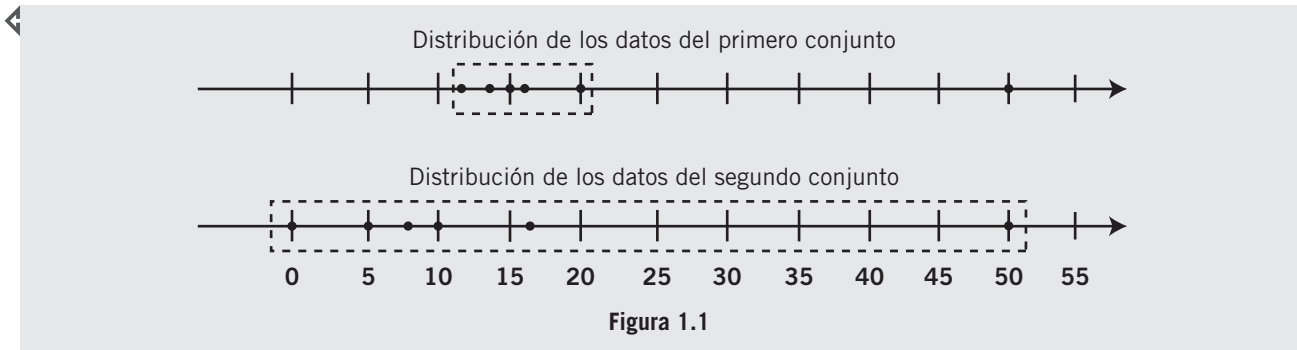
10. Pruebe que cualquier medida central siempre es un valor que se encuentra entre el menor y mayor de los valores de los datos.
11. Pruebe, en general, que la media geométrica aplicada al factor de crecimiento promedio en las inversiones siempre dará el resultado más preciso.

## 1.6 Medidas de dispersión

Cuando se lleva a cabo un análisis de la distribución de los datos de una muestra, el estudio de sus medidas centrales no es suficiente, debido a que en diferentes conjuntos de datos pueden dar medidas centrales iguales. Por tanto, no se tendría el conocimiento de la forma de su distribución.

### Ejemplo 1.24 Medidas de dispersión

Si un conjunto de datos contiene los valores 20, 12, 15, 16, 13 y 14; un segundo conjunto los valores 5, 0, 50, 17, 8 y 10, podemos comprobar que en ambos casos se obtiene un promedio de 15 (¡verifíquelo!). Pero, si representamos en una recta los datos no es difícil comprobar que las observaciones del segundo conjunto tienen una dispersión mucho mayor (véase figura 1.1).



De los conjuntos de datos anteriores podemos observar que es necesario introducir una medida con la que sea posible comparar qué datos son más dispersos. Dichas medidas se conocen como valores de dispersión o variabilidad del conjunto de datos.

En síntesis, una medida de dispersión indica qué tan cercanos o separados están los valores con respecto a la media u otra medida de tendencia central. En las siguientes subsecciones se muestran las medidas de dispersión más comunes de la estadística descriptiva.

## Rango

El primer valor que muestra cómo están dispersos los datos es muy sencillo y lo llamamos **rango** de las observaciones, lo denotaremos por  $r$ .

El **rango** es una medida variacional de los datos que lo único que indica es el tamaño o longitud del intervalo en el que estos se encuentran distribuidos y se calcula por:

$$\text{Rango} = \text{El valor mayor menos el valor menor de los datos.} \quad (1.10)$$

### Ejemplo 1.25 Rango

Para los datos muestrales de los dos conjuntos anteriores se tiene:

- En los datos anteriores en el primer conjunto su rango vale  $r_1 = 20 - 12 = 8$ . Es decir, los datos de este conjunto están distribuidos a lo largo de un intervalo de longitud 8.
- En el segundo conjunto su rango vale  $r_2 = 50 - 0 = 50$ . Es decir, los datos de este conjunto están distribuidos a lo largo de un intervalo de longitud 50.

Obviamente, de los dos resultados anteriores es fácil concluir que los elementos del segundo conjunto tienen una separación mayor. Aunque el resultado no muestra el comportamiento de los datos con respecto a su media o algún valor central.

## Variancia y desviación estándar

Sean una muestra  $x_1, x_2, \dots, x_n$  con  $n$  datos y valor medio  $\bar{x}$ , los cuadrados de las desviaciones de cada uno de los datos con respecto a su valor medio son:

$$(x_1 - \bar{x})^2, (x_2 - \bar{x})^2, \text{ etcétera.}$$

De manera que otra medida de dispersión de los datos que está relacionada directamente con su media aritmética es la siguiente.

Sea  $x_1, x_2, \dots, x_n$  los valores de una muestra aleatoria, llamaremos:

$$\text{Variancia (varianza) sesgada o poblacional a: } s_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (1.11)$$

$$\text{Variancia (varianza) insesgada o muestral a: } s_{n-1}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (1.12)$$

Tal vez surga la pregunta: ¿por qué dos definiciones diferentes? La respuesta es sencilla, la variancia sesgada refleja a la perfección el significado de una medida de dispersión como un promedio de los cuadrados de las desviaciones y tiene una gran aplicación en el estudio de las probabilidades. Mientras que la variancia insesgada es más propicia en los cálculos estadísticos y se usa en las muestras (véase unidad 3). Precisamente de aquí surge el nombre de variancia muestral.

En la definición de variancia se puede notar que se calcula con los cuadrados de las desviaciones; por tanto, no está en las mismas unidades que los datos, una razón para introducir una nueva medida de dispersión.

Se llama **desviación estándar** de un conjunto de datos a la raíz cuadrada positiva de la variancia, la cual dependerá del tipo de variancia que se esté empleando.

### Ejemplo 1.26 Varianza insesgada

Calcule la variancia insesgada y su desviación estándar correspondiente en cada uno de los dos conjuntos dados que se dieron al inicio de la sección. Conjunto uno: 20, 12, 15, 16, 13 y 14; el segundo conjunto: 5, 0, 50, 17, 8 y 10.

#### Solución

Sea el conjunto de 20, 12, 15, 16, 13 y 14,  $\bar{x} = 15$ . Entonces de la fórmula 1.12:

$$\begin{aligned} s_{n-1}^2 &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{6-1} [(20-15)^2 + (12-15)^2 + (15-15)^2 + (16-15)^2 + (13-15)^2 + (14-15)^2] \\ &= \frac{1}{5} [25 + 9 + 0 + 1 + 4 + 1] = 8 \end{aligned}$$

La desviación estándar  $s_{n-1} = \sqrt{8} \approx 2.8284$ .

Para otro conjunto de datos 5, 0, 50, 17, 8 y 10,  $\bar{x} = 15$ , pero su variancia insesgada:

$$\begin{aligned} s_{n-1}^2 &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= \frac{1}{6-1} [(5-15)^2 + (0-15)^2 + (50-15)^2 + (17-15)^2 + (8-15)^2 + (10-15)^2] \\ &= \frac{1}{5} [100 + 225 + 1225 + 4 + 49 + 25] \\ &= 325.6 \end{aligned}$$

De igual manera, en el primer conjunto de datos la desviación estándar es  $s_{n-1} = \sqrt{325.6} \approx 18.0444$ .

Como se puede observar en ambos resultados, aunque la media es la misma para los datos de ambos conjuntos, el segundo conjunto de datos tiene una variabilidad considerablemente mayor que los datos del primer conjunto. Entonces, se dice que los resultados del primer conjunto son más homogéneos que los del segundo.

## Otra expresión para cálculos de las variancias

En los cálculos de la variancia se acostumbra emplear otra representación equivalente a 1.11 o 1.12, la cual está dada por las fórmulas 1.13 y 1.14, respectivamente:

$$\text{Variancia sesgada o poblacional a: } s_n^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2 = \frac{n-1}{n} s_{n-1}^2 \quad (1.13)$$

$$\text{Variancia insesgada o muestral a: } s_{n-1}^2 = \frac{1}{n-1} \sum_{i=1}^n x_i^2 - \frac{n}{n-1} \bar{x}^2 = \frac{n}{n-1} s_n^2 \quad (1.14)$$

### Demostración

$$\begin{aligned} s_n^2 &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n (x_i^2 - 2x_i \bar{x} + \bar{x}^2) = \frac{1}{n} \left\{ \sum_{i=1}^n (x_i^2) - 2\bar{x} \sum_{i=1}^n x_i + \sum_{i=1}^n \bar{x}^2 \right\} \\ &= \frac{1}{n} \left\{ \sum_{i=1}^n (x_i^2) - 2\bar{x}n\bar{x} + n\bar{x}^2 \right\} = \frac{1}{n} \left\{ \sum_{i=1}^n (x_i^2) - n\bar{x}^2 \right\} = \frac{1}{n} \sum_{i=1}^n (x_i^2) - \bar{x}^2 \end{aligned}$$

### Ejemplo 1.27 Varianza insesgada

Calcule la variancia insesgada para los conjuntos de datos del ejemplo 1.26 empleando las últimas fórmulas para la variancia y compruebe que los resultados coinciden.

#### Solución

Sea el conjunto de datos 20, 12, 15, 16, 13 y 14, empleando la fórmula 1.14:

$$\begin{aligned} s_{n-1}^2 &= \frac{1}{n-1} \sum_{i=1}^n x_i^2 - \frac{n}{n-1} \bar{x}^2 = \frac{1}{6-1} [20^2 + 12^2 + 15^2 + 16^2 + 13^2 + 14^2] - \frac{6}{6-1} (15)^2 \\ &= \frac{1}{5} [400 + 144 + 225 + 256 + 169 + 196] - \frac{6}{5} \times 225 = 278 - 270 = 8 \end{aligned}$$

Valor que coincide calculado con la fórmula 1.12.

Para el conjunto de datos 5, 0, 50, 17, 8 y 10, aplicamos la fórmula 1.14:

$$\begin{aligned} s_{n-1}^2 &= \frac{1}{n-1} \sum_{i=1}^n x_i^2 - \frac{n}{n-1} \bar{x}^2 = \frac{1}{6-1} [5^2 + 0^2 + 50^2 + 17^2 + 8^2 + 10^2] - \frac{6}{6-1} (15)^2 \\ &= \frac{1}{5} [25 + 0 + 2500 + 289 + 64 + 100] - \frac{6}{5} \times 225 = 595.6 - 270 = 325.6 \end{aligned}$$

Valor que coincide al emplear la fórmula 1.12.

## Desviación media

Otra medida de dispersión para los datos que está relacionada directamente con su promedio se define a continuación.

Sean  $x_1, x_2, \dots, x_n$  los datos en estudio, llamaremos **desviación media (DM)** o **desviación media absoluta (DMA)** del conjunto de datos al promedio de los valores absolutos de las desviaciones de cada uno de los datos con respecto a la media. Es decir:

$$DM = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}| \quad (1.15)$$

Es posible que para algunos lectores esta medida de variación resulte la más adecuada para medir la dispersión de los datos; en efecto, la desviación media tiene buenas propiedades para medir la dispersión, pero en muchos cálculos

no es recomendable trabajar con el valor absoluto ya que este no es una función derivable, mientras que la función cuadrática utilizada en la varianza sí lo es.

### Ejemplo 1.28 Desviación media

Calcule la desviación media para los conjuntos de datos del ejemplo 1.26 y compare los resultados con la desviación estándar de la varianza insesgada.

#### Solución

Conjunto de datos 20, 12, 15, 16, 13 y 14 cuya media fue 15. Luego:

$$\begin{aligned} DM &= \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}| = \frac{1}{6} \{ |20 - 15| + |12 - 15| + |15 - 15| + |16 - 15| + |13 - 15| + |14 - 15| \} \\ &= \frac{1}{6} \{ |5| + |-3| + |0| + |1| + |-2| + |-1| \} = \frac{1}{6} \{ 5 + 3 + 0 + 1 + 2 + 1 \} = 2 \end{aligned}$$

En este caso, la desviación estándar fue  $\sqrt{8} = 2.828 > DM$ .

Para el conjunto de 5, 0, 50, 17, 8 y 10. Su media también resultó ser igual a 15:

$$\begin{aligned} DM &= \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}| = \frac{1}{6} \{ |5 - 15| + |0 - 15| + |50 - 15| + |17 - 15| + |8 - 15| + |10 - 15| \} \\ &= \frac{1}{6} \{ |10| + |-15| + |35| + |2| + |-7| + |-5| \} = \frac{1}{6} \{ 10 + 15 + 35 + 2 + 7 + 5 \} = 12.3333 \end{aligned}$$

La desviación estándar es  $\sqrt{325.6} = 18.044 > DM$ .

## Covarianza

Por último, una medida que que representa la dependencia entre dos muestras o poblacionales está definida de la siguiente forma.

Sean los datos de dos muestras del mismo tamaño,  $x_1, x_2, \dots, x_n$  y  $y_1, y_2, \dots, y_n$ , referentes a diferentes características, llamaremos **covarianza** a la medida que refleja el grado de dependencia entre los datos de las dos muestras, la denotaremos por  $s_{xy}$  y calcularemos por:

$$s_{xy} = \text{cov}(x, y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y} \quad (1.16)$$

### Ejemplo 1.29 Covarianza

Suponga que se estudia un grupo de 10 personas en sus características de ingresos y años de estudio (véase tabla 1.9).

Tabla 1.9

Persona	1	2	3	4	5	6	7	8	9	10
Ingreso en miles	10.5	6.8	20.7	18.2	8.6	25.8	22.2	5.9	7.6	11.8
Años de estudio	17	18	21	16	16	21	16	14	18	18

Calcule la covarianza entre las dos características.

#### Solución

En el citado ejemplo calculamos sus promedios  $\bar{x} = 13.81$  y  $\bar{y} = 17.5$ . Ahora, debemos calcular los productos (véase tabla 1.10).

Tabla 1.10

Persona $i$	1	2	3	4	5	6	7	8	9	10
Ingreso en miles ( $x$ )	10.5	6.8	20.7	18.2	8.6	25.8	22.2	5.9	7.6	11.8
Años de estudio ( $y$ )	17	18	21	16	16	21	16	14	18	18
$x_i y_i$	178.5	122.4	434.7	291.2	137.6	541.8	355.2	82.6	136.8	212.4

De tal forma que la covarianza entre ingresos y años de estudio se calcula con la fórmula 1.16:

$$s_{xy} = \text{cov}(x, y) = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y} = \frac{2493.2}{10} - 13.81(17.5) = 7.645$$

Al finalizar el ejemplo anterior es probable que surja la pregunta sobre la interpretación del valor obtenido, 7.645, para la covarianza de las dos características. Ahora bien, ¿qué significa 7.645, la dependencia entre las características es grande o pequeña?

Con los datos que tenemos, la respuesta no resulta tan obvia porque, en realidad, depende del tamaño de las unidades en la que están los datos. Para evitar este problema de interpretación, se introduce una nueva medida basada en un coeficiente que representa la medida relativa de dependencia entre los caracteres en estudio.

Sean los datos de dos muestras del mismo tamaño,  $x_1, x_2, \dots, x_n$  y  $y_1, y_2, \dots, y_n$ , referentes a características diferentes, llamaremos **coeficiente de correlación muestral** a la medida que refleja el grado de dependencia entre las dos muestras y lo denotaremos por  $r_{xy}$ , se calcula:

$$r_{xy} = \frac{\text{cov}(x, y)}{\sqrt{s_n^2(x)} \sqrt{s_n^2(y)}} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x} \bar{y}}{\sqrt{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \sqrt{\sum_{i=1}^n y_i^2 - n\bar{y}^2}} \quad (1.17)$$

El coeficiente de correlación se encuentra  $[-1, 1]$ .

### Ejemplo 1.30 Coeficiente de correlación muestral

En el ejemplo anterior referente al grupo de 10 personas con características de ingresos y años de estudio, calcule su coeficiente de correlación para las características.

#### Solución

En el ejemplo anterior calculamos la covarianza de las dos muestras  $\text{cov}(x, y) = 7.645$ , mientras que en el ejemplo 1.26, sus varianzas insesgadas, luego las varianzas sesgadas estarán dadas por:

$$s_n^2(x) = \frac{n-1}{n} s_{n-1}^2(x) = \frac{9}{10} (52.5899) = 47.3309 \quad \text{y} \quad s_n^2(y) = \frac{n-1}{n} s_{n-1}^2(y) = \frac{9}{10} (4.9444) = 4.4496$$

De tal forma que el coeficiente de correlación entre ingresos y años de estudio se calcula con la fórmula 1.17 y resulta:

$$r_{xy} = \frac{\text{cov}(x, y)}{\sqrt{s_n^2(x)} \sqrt{s_n^2(y)}} = \frac{7.645}{\sqrt{47.3309} \sqrt{4.4496}} = 0.5268$$

Después de calcular el coeficiente de correlación muestral es natural que surjan preguntas sobre la interpretación del valor obtenido, por ejemplo: ¿cómo interpretar el coeficiente de correlación muestral?

A diferencia de la covarianza, el coeficiente de correlación muestral tiene una interpretación que no depende del tamaño de los valores de los datos, debido a que siempre será una cantidad adimensional entre  $[-1, 1]$ , sin importar qué tan grandes o pequeños sean los datos muestrales.

- Si  $r_{xy} > 0$ , esto es,  $r_{xy} \in (0, 1]$ , entonces se dice que los caracteres son directamente proporcionales. Es decir, cuando el valor de uno de los dos aumenta o disminuye el otro también lo hace.
- Si  $r_{xy} < 0$ , esto es,  $r_{xy} \in [-1, 0)$ , entonces se dice que los caracteres son inversamente proporcionales. Es decir, cuando el valor de uno de los dos caracteres aumenta o disminuye el otro también lo hace.
- Si  $r_{xy} = 0$ , entonces se dice que los caracteres no son dependientes. Es decir, el aumento o disminución de uno de éstos no influye en el del otro.

En forma numérica, si  $|r_{xy}| \approx 1$ , se dice que los caracteres en estudio tienen un alto grado de dependencia, ya sea directa o inversa, según sea el signo de  $r_{xy}$ . Por otro lado, si  $r_{xy} \approx 0$  se dice que los caracteres en estudio tienen un grado muy pequeño de dependencia, ya sea directa o inversa, según sea el signo de  $r_{xy}$ . Para valores intermedios la interpretación, en general, depende del investigador, a partir de qué valores de  $r_{xy}$  se considera que las muestras sean dependientes. Así, para algunos investigadores el valor 0.7268 puede ser considerado como una alta dependencia entre las muestras y para otros puede considerarse una dependencia moderada.

### Ejercicios 1.3

1. Para el siguiente conjunto de datos, calcule su rango, varianza insesgada y desviación media.

145 150 165 155 155 145 150 140 145 150 160 175 150 160

2. Calcule la desviación estándar muestral de los tiempos entre llegadas indicadas (en minutos) de aviones en el aeropuerto Benito Juárez, de la Ciudad de México:

3.5, 4.2, 2.9, 3.8, 4.0, 5.3, 2.4, 3.8, 4.6, 3.9, 5.2, 4.3, 3.9 y 2.8

3. En los envases de leche, la cantidad de líquido no siempre es un litro, se toma una muestra de 10 paquetes, obteniéndose las mediciones de abajo, en litros. Calcule el rango, la varianza insesgada y la desviación media de los contenidos de leche.

0.95 1.01 0.97 0.95 1.0 0.97 0.95 1.01 0.95 0.98

4. Sean las calificaciones de 30 estudiantes en la materia de probabilidad (véase tabla 1.11).

Tabla 1.11

27	72	83	15	96	30	8	98	86	5	39	86	87	100	56
88	31	3	30	57	22	7	20	62	95	35	73	66	56	57

Calcule el rango, la varianza insesgada y la desviación media de las calificaciones.

5. La Bolsa Mexicana de Valores registró alzas y bajas en puntos porcentuales indicadas en la tabla 1.12 durante la primera quincena de junio de 2015. Calcule su varianza muestral.

Tabla 1.12

3.4%	1.7%	-0.5%	0.7%	-2.4%	-1.8%	-0.9%	2.5%	0.3%	0.8%
------	------	-------	------	-------	-------	-------	------	------	------

6. En la tabla 1.13 se muestran las calificaciones de 30 alumnos correspondientes a las materias de cálculo diferencial y algebra lineal.

Tabla 1.13

Álgebra lineal	80	70	43	55	23	98	42	73	20	35
	75	95	70	75	57	32	32	82	50	96
	46	83	45	75	60	65	100	86	30	10

Continúa



Tabla 1.13 (Continuación)

Cálculo diferencial	90	100	38	30	10	70	30	65	10	45
	80	50	52	80	40	40	10	65	40	90
	30	43	30	90	35	40	90	60	25	10

Calcule las medidas de variabilidad por grupo:

a) Rango y varianza insesgada.

b) Covarianza y el coeficiente de correlación entre las dos materias.

7. Se llevó a cabo un experimento y se anotaron sus valores  $\bar{x} = 53.48$ , con  $\sum_{i=1}^{25} x_i^2 = 86\,463$ . Calcule la varianza insesgada de los datos.
8. Para determinar la dependencia entre dos caracteres se hizo un estudio de 20 y se anotaron sus valores  $\sum_{i=1}^{20} x_i = 208$ ,  $\sum_{i=1}^{20} x_i^2 = 2\,540.5$ ,  $\sum_{i=1}^{20} y_i = 1\,067$ ,  $\sum_{i=1}^{20} y_i^2 = 65\,713$  y  $\sum_{i=1}^{20} x_i y_i = 12\,884.5$ . Calcule el coeficiente de correlación de los datos muestrales.
9. Se llevó a cabo un experimento y se anotaron sus valores  $\sum_{i=1}^{50} x_i = 1\,634$ , con  $\sum_{i=1}^{50} x_i^2 = 94\,492$ . Calcule la varianza insesgada de los datos.
10. Se conoce que  $\sum_{i=1}^{30} x_i = 331.3$ ,  $\sum_{i=1}^{30} y_i = 1\,673.5$  y  $\sum_{i=1}^{30} x_i y_i = 22\,414$ . Calcule la covarianza de los valores muestrales para  $x$  y  $y$ .

## 1.7 Parámetros de forma en la distribución de la muestra

En esta sección revisaremos cómo podemos definir y calcular un estadístico de **forma** para la distribución de los datos muestrales.

Sean  $x_1, x_2, \dots, x_n$ ,  $n$  datos con media  $\bar{x}$  y desviación estándar muestral  $s_{n-1}$ , entonces se llama **coeficiente de sesgo** o **coeficiente de asimetría** a la medida que representa el grado de asimetría de la gráfica y lo denotaremos por  $CA$ , en la literatura se usan por lo regular dos fórmulas para el cálculo:

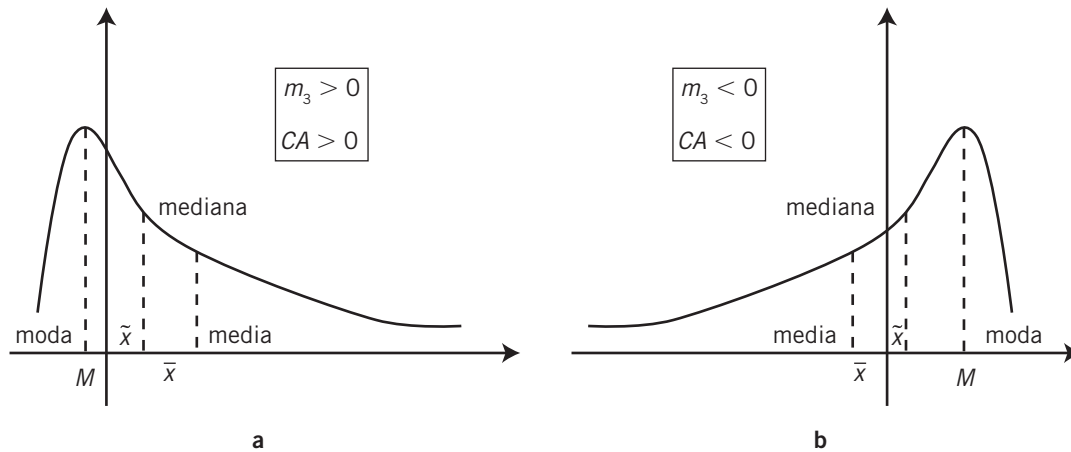
$$CA_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s_{n-1}} \right)^3 \quad (1.18a)$$

$$CA_2 = \frac{m_3}{(s_{n-1})^3} = \frac{1}{n} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s_{n-1}} \right)^3 \quad (1.18b)$$

Donde,  $m_3 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3$  y  $s_{n-1}$  es la desviación estándar correspondiente a la varianza insesgada. Como podemos apreciar, las dos fórmulas casi coinciden cuando  $n$  es grande y están relacionadas por  $CA_2 = \frac{(n-1)(n-2)}{n^2} CA_1$ , que en el valor límite  $n \rightarrow \infty$  son iguales.

El coeficiente de asimetría caracteriza el grado de alejamiento de los datos con respecto a su media y en general se encuentra entre  $-4$  y  $4$ . Cuando el coeficiente de asimetría vale cero, significa que su distribución es simétrica. Para calcular el coeficiente de asimetría se requieren al menos tres datos. El paquete Excel utiliza la fórmula  $CA_1$  para calcular la asimetría o sesgo (véase figura 1.2).

En la figura 1.2 podemos apreciar que en el caso de la asimetría positiva (a), el valor correspondiente a la moda es más pequeña que la mediana y ésta más pequeña que la media. De forma contraria, cuando el sesgo es negativo (b), la media es la más pequeña, le sigue en tamaño la mediana y, por último, la moda es más grande. En el caso de que la distribución sea simétrica, las tres medidas centrales coinciden. Si resumimos, tenemos:



**Figura 1.2** Tipos de asimetría. a) Asimetría derecha, datos sesgados a la derecha. b) Asimetría izquierda, datos sesgados a la izquierda.

$$CA = \begin{cases} 0, & \text{la distribución de los datos es simétrica.} \\ < 0, & \text{los datos están sesgados a la izquierda.} \\ > 0, & \text{los datos están sesgados a la derecha.} \end{cases}$$

### Ejemplo 1.31 Coeficiente de asimetría

Calcule el coeficiente de asimetría para los datos referentes a la estatura de 50 estudiantes del Instituto Tecnológico de Mexicali (véase tabla 1.14).

**Tabla 1.14**

173.5	171.4	178.2	165.7	180.0	174.6	176.0	168.5	180.1	165.9
169.0	175.4	176.5	164.0	167.5	158.4	168.0	172.8	172.5	173.2
170.5	180.5	171.8	184.3	178.5	172.0	174.5	173.0	176.3	186.4
167.5	165.7	165.0	178.0	177.5	181.0	179.5	174.6	173.2	172.9
170.5	181.3	160.6	168.5	170.0	171.0	176.5	178.9	180.0	169.0

### Solución

Primero, se calcula la media y desviación estándar muestral de los 50 datos con las fórmulas 1.5 y 1.12, respectivamente:

$$\bar{x} = 173.204 \text{ y } s_{n-1} = 5.9554$$

El coeficiente de asimetría se calcula con la fórmula 1.18a:

$$CA_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s_{n-1}} \right)^3 = \frac{50}{49 \times 48} \sum_{i=1}^{50} \left( \frac{x_i - 173.204}{5.9554} \right)^3$$

Enseguida, se calcula cada uno de los 50 sumandos  $\left( \frac{x_i - 173.204}{5.9554} \right)^3$ , con lo que se obtiene:

Tabla 1.15

0.000	-0.028	0.590	-2.001	1.486	0.013	0.103	-0.493	1.553	-1.845
-0.352	0.050	0.170	-3.691	-0.879	-15.360	-0.667	0.000	-0.002	0.000
-0.094	1.839	-0.013	6.468	0.703	-0.008	0.010	0.000	0.140	10.879
-0.879	-2.001	-2.614	0.522	0.375	2.243	1.182	0.013	0.000	0.000
-0.094	2.512	-9.480	-0.493	-0.156	-0.051	0.170	0.875	1.486	-0.352

Luego, se suman y se obtiene  $-8.167$ , de manera que:

$$CA_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s_{n-1}} \right)^3 = \frac{50}{49 \times 48} (-8.167) = -0.1736$$

De igual modo, podemos obtener el coeficiente de asimetría con la fórmula 1.18b:

$$CA_2 = \frac{(n-1)(n-2)}{n^2} CA_1 = \frac{49 \times 48}{50 \times 50} (-0.1736) = -0.1633$$

En este caso, resulta que los datos tienen un pequeño sesgo a la izquierda.

## Ejercicios 1.4

1. Las calificaciones de 30 estudiantes en la materia de probabilidad se muestran en la tabla 1.16.

Tabla 1.16

27	72	83	15	96	30	8	98	86	5	39	86	87	100	56
88	31	3	30	57	22	7	20	62	95	35	73	66	56	57

Calcule su coeficiente de asimetría e indique cómo es la distribución de los datos con respecto a su simetría.

2. En la tabla 1.17 se muestran las calificaciones de 30 alumnos correspondientes a las materias de cálculo diferencial y álgebra lineal.

Tabla 1.17

Álgebra lineal	80	70	43	55	23	98	42	73	20	35
	75	95	70	75	57	32	32	82	50	96
	46	83	45	75	60	65	100	86	30	10
Cálculo diferencial	90	100	38	30	10	70	30	65	10	45
	80	50	52	80	40	40	10	65	40	90
	30	43	30	90	35	40	90	60	25	10

Calcule su coeficiente de asimetría por grupo e indique cómo es la distribución de los datos con respecto a su simetría.

3. Considere los valores del IPC de una cadena de supermercados y calcule el coeficiente de asimetría (véase tabla 1.18).

Tabla 1.18

37.1	36.99	37.83	36.36	36.17	35.98	35.87	35.68	35.92	35.91
35.29	34.86	34.83	35.12	34.71	34.48	34.74	34.85	34.79	34.95
34.29	34.52	35.02	34.87	34.91	34.74	34.71	34.63	34.85	34.84
34.59	34.65	34.68	34.50	34.41	34.13	33.93	33.73	33.32	33.78

4. La tabla 1.19 representa el año y la cantidad de divorcios de mutuo consentimiento en México desde 1985 hasta 2011. Calcule el coeficiente de asimetría.

Tabla 1.19

1985	1986	1987	1988	1989	1990	1991	1992	1993
24 207	26 110	29 934	30 341	29 676	30 060	33 403	34 524	25 343
1994	1995	1996	1997	1998	1999	2000	2001	2002
26 940	28 254	28 364	29 847	32 889	35 003	37 937	40 796	43 351
2003	2004	2005	2006	2007	2008	2009	2010	2011
46 285	49 046	51 091	52 712	55 995	59 543	58 502	58 466	62 744

## 1.8 Aplicación de las medidas a inversiones

Los conceptos vistos sobre estadística descriptiva tienen una gama muy amplia de aplicaciones, en esta sección hablaremos brevemente sobre las inversiones, para esto iniciamos explicando qué se entiende por título en el contexto de inversiones.

En finanzas, el término **título** o activo financiero se aplica al conjunto de instrumentos legales que incluyen bonos, acciones y préstamos otorgados por instituciones financieras, cuyos propietarios tienen ciertos derechos para percibir en el futuro una determinada cantidad monetaria.

Casi todos los títulos de valores que se negocian en los mercados secundarios suelen pertenecer a uno de los siguientes dos grandes grupos: bonos o acciones. Los bonos son instrumentos crediticios (deuda emitida en general por el gobierno o las empresas), a cambio de cierta cantidad de dinero, proporcionan un rendimiento fijo. Las acciones preferentes son parecidas a los bonos, puesto que tienen un valor facial y proporcionan un dividendo predeterminado (parecido al cupón de los bonos). La diferencia estriba en que las acciones preferentes, a diferencia de los bonos, no tienen un plazo de vencimiento; además, se puede no pagar los dividendos anualmente durante varios años, sin que ello implique la quiebra del emisor. Estos títulos de valores tienen un periodo de vida ilimitado y solo se pagarán dividendos si el emisor obtiene algunos beneficios satisfactorios. Dado que los rendimientos de los bonos son los más seguros, constituyen la inversión menos arriesgada, pero a su vez tienen un menor rendimiento. Las acciones preferentes toleran mayores riesgos que los bonos, pero menores que las acciones ordinarias. Las acciones preferentes son las más arriesgadas, por lo que su tasa de rendimiento esperada es también la más elevada.

Así, sabemos que el rendimiento de un título es proporcional al riesgo, pero ¿cómo se miden el riesgo y el rendimiento de un título?

Para hacerlo, necesitamos conocer los diferentes precios en diferentes intervalos de los títulos. Para saber el historial de cuánto se gana y cuánto se pierde se define al rendimiento como:

$$\text{Rendimiento (hoy)} = R = \frac{\text{Precio de hoy} - \text{Precio de ayer}}{\text{Precio de ayer}} \quad (1.19)$$

Por su parte, el riesgo puede ser definido por medio de una medida de variabilidad, puesto que estas representan un riesgo a la perfección. Mayor variabilidad, mayor inestabilidad y esto implica un riesgo elevado. Menor variabilidad, mayor estabilidad y esto implica poco riesgo. Entonces se define al riesgo de la inversión:

Riesgo = desviación estándar, de la varianza sesgada, de los rendimientos =  $s_n(R)$

A continuación, se detalla un ejemplo para explicar rendimientos y riesgos.

### Ejemplo 1.32 Rendimientos y riesgos

Dados los títulos de las empresas WM y TX en 15 días sucesivos, calcule:

- Rendimientos
- Rendimientos promedio en dicho periodo
- Riesgos

**Tabla 1.20**

Día	WM	TX
1	34.79	17.80
2	34.85	17.57
3	34.74	17.59
4	34.48	17.85
5	34.71	17.87
6	35.12	18.17
7	34.83	18.17
8	34.86	18.30
9	35.29	18.36
10	35.91	18.40
11	35.92	18.40
12	35.68	18.48
13	35.87	18.51
14	35.98	18.45
15	36.17	18.49

### Solución

Primero, calculamos sus rendimientos (véase tabla 1.21).

**Tabla 1.21**

Día	WM	Rendimientos WMt	TX	Rendimientos TX
1	34.79		17.8	
2	34.85	0.00172	17.57	-0.01292
3	34.74	-0.00316	17.59	0.00114
4	34.48	-0.00748	17.85	0.01478
5	34.71	0.00667	17.87	0.00112
6	35.12	0.01181	18.17	0.01679
7	34.83	-0.00826	18.17	0.00000
8	34.86	0.00086	18.3	0.00715

Continúa 

Tabla 1.21 (Continuación)

Día	WM	Rendimientos WMt	TX	Rendimientos TX
9	35.29	0.01234	18.36	0.00328
10	35.91	0.01757	18.4	0.00218
11	35.92	0.00028	18.4	0.00000
12	35.68	-0.00668	18.48	0.00435
13	35.87	0.00533	18.51	0.00162
14	35.98	0.00307	18.45	-0.00324
15	36.17	0.00528	18.49	0.00217
<b>Rendimiento promedio</b>		<b>0.00281</b>		<b>0.00274</b>
<b>Riesgo</b>		<b>0.00776</b>		<b>0.00717</b>

En resumen, los valores del título de WM resultan ser un poco más dispersos que los de TX.

Ahora bien, puede ocurrir que el decisor quiera arriesgar su capital de inversión en varios títulos, en este caso estamos hablando de portafolios. En un portafolio, el inversionista destina proporciones de su capital a cada título, de manera que requiere conocer el rendimiento promedio del portafolio.

Suponga que el inversionista tiene  $n$  títulos en los que invierte su capital de acuerdo con las siguientes proporciones  $p_1, p_2, \dots, p_n$ , con:

$$p_1 + p_2 + \dots + p_n = 1$$

Entonces, el rendimiento promedio del portafolio no es otra cosa que la media ponderada de los rendimientos promedios de cada uno de los  $n$  títulos del inversionista. Luego, el rendimiento promedio del portafolio

$$\bar{R}_p = \sum_{i=1}^n p_i \bar{R}_i \quad (1.20)$$

Donde,  $\bar{R}_i$  representa el rendimiento promedio del título  $i$ . Mientras que el riesgo del portafolio está representado con la variabilidad de los títulos y se calcula:

$$s_{R_p} = \sqrt{\sum_{i=1}^n p_i^2 s_{R_i}^2 + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n p_i p_j \text{cov}(R_i, R_j)} \quad (1.21)$$

Donde  $s_{R_i}^2$  representa la varianza sesgada de los rendimientos del título  $i$ ;  $\text{cov}(R_i, R_j)$  es la covarianza que existe entre los rendimientos de los títulos  $i$  con  $j$ .

Finalizamos la unidad con un ejemplo en el que se propone un portafolio y se calcula su rendimiento promedio y riesgo del portafolio.

### Ejemplo 1.33 Portafolio

Considere un portafolio con los dos títulos del ejemplo 1.32, WM y TX en 15 días sucesivos y calcule:

- Rendimiento promedio del portafolio para una inversión de 30% en WM y 70% en TX.
- En la inversión del inciso a), el riesgo del portafolio.

#### Solución

El rendimiento promedio del portafolio.

$$\bar{R}_p = \sum_{i=1}^2 p_i \bar{R}_i = 0.30(0.00281) + 0.70(0.00274) = 0.00276$$

Para el riesgo del portafolio, primero necesitamos calcular la covarianza entre los rendimientos. Así, encontramos que  $\text{cov}(R_1, R_2) = 0.0000019498$ .

Ahora, calculemos el riesgo del portafolio:

$$\begin{aligned} s_{R_p} &= \sqrt{\sum_{i=1}^2 p_i^2 s_{R_i}^2 + 2 \sum_{i=1}^{2-1} \sum_{j=i+1}^2 p_i p_j \text{cov}(R_i, R_j)} = \sqrt{p_1^2 s_{R_1}^2 + p_2^2 s_{R_2}^2 + 2p_1 p_2 \text{cov}(R_1, R_2)} \\ &= \sqrt{0.3^2(0.00776)^2 + 0.7^2(0.00717)^2 + 2(0.3)(0.7)(0.00000195)} \\ &= 0.00561 \end{aligned}$$

## Ejercicio 1.5

Sean los siete títulos de la tabla 1.22, con ayuda de algún paquete calcule:

- Rendimientos de los títulos.
- Rendimientos promedio de cada título.
- Riesgos de cada título.
- Determine un portafolio de tres títulos (WM, BB y EK) y calcule el rendimiento promedio del portafolio para una inversión de 30, 20 y 50%, respectivamente.
- En la inversión del inciso anterior calcule el riesgo del portafolio de la tabla 1.22.

Tabla 1.22

Fecha	WM	SOR	BB	CX	CM	EK	GMO
16/08/2013	35.57	42.31	42.71	15.58	51.33	470.39	116.42
15/08/2013	35.54	42.34	43.45	15.53	52.31	476.17	116.42
14/08/2013	35.06	42.25	44.29	15.87	53.15	482.47	117.00
13/08/2013	35.20	41.06	44.31	15.79	54.55	489.95	117.00
12/08/2013	34.64	42.78	44.61	15.40	56.13	508.86	117.35
09/08/2013	35.23	41.80	45.80	15.37	56.44	510.95	118.00
08/08/2013	34.81	42.00	45.49	15.47	56.19	500.88	115.87
07/08/2013	35.05	41.90	45.13	15.11	55.50	491.26	117.31
06/08/2013	34.42	42.00	44.74	15.28	54.48	495.05	114.02
05/08/2013	34.83	42.99	45.43	15.27	54.01	486.28	115.12
02/08/2013	35.62	43.59	45.62	15.13	54.38	482.90	116.38
01/08/2013	35.63	42.50	45.43	15.24	54.28	471.63	115.87
31/07/2013	34.95	42.81	43.20	14.72	54.05	462.88	117.77
30/07/2013	34.89	43.08	42.99	14.50	53.90	467.24	116.23
29/07/2013	35.01	43.02	42.86	14.54	52.22	462.88	115.80
26/07/2013	35.88	43.06	43.74	14.78	51.62	464.49	115.20
25/07/2013	35.53	44.60	42.75	14.58	51.66	464.99	115.30
24/07/2013	35.97	44.46	42.55	14.37	49.71	459.28	115.30
23/07/2013	35.53	42.82	41.31	14.34	49.06	467.03	114.70
22/07/2013	35.02	40.90	39.98	14.40	47.98	459.74	114.80
19/07/2013	35.18	42.21	40.78	14.26	48.50	462.00	115.77
18/07/2013	34.69	43.53	41.43	14.22	49.09	468.35	113.92
17/07/2013	35.22	43.08	41.63	14.10	48.13	464.78	115.90



Tabla 1.22 (Continuación)

Fecha	WM	SOR	BB	CX	CM	EK	GMO
16/07/2013	35.91	43.38	39.83	13.97	48.20	461.93	115.90
15/07/2013	36.37	43.89	41.50	14.29	48.32	473.21	116.85
12/07/2013	35.99	44.90	40.49	14.50	47.47	480.64	116.85
11/07/2013	36.72	45.32	39.50	14.22	46.81	482.86	118.01
10/07/2013	36.07	44.35	38.63	13.60	46.27	479.66	117.90
09/07/2013	36.06	44.37	39.90	13.57	46.10	473.72	117.68
08/07/2013	36.44	45.97	39.89	13.44	47.45	465.67	119.02
05/07/2013	36.93	46.00	39.44	13.61	47.85	458.80	119.45
04/07/2013	37.11	46.44	41.05	13.69	47.92	453.54	118.30
03/07/2013	36.74	45.30	40.35	13.52	47.88	451.97	118.61
02/07/2013	36.88	46.92	39.27	13.51	48.01	444.76	119.26
01/07/2013	37.11	48.52	40.34	13.91	48.87	438.78	118.40
28/06/2013	36.48	48.60	39.25	13.76	49.25	428.95	117.97
27/06/2013	35.58	47.06	37.46	13.80	47.96	432.89	118.76
26/06/2013	34.53	44.93	35.82	13.70	46.01	436.36	120.42
25/06/2013	34.18	42.39	34.68	13.24	44.97	430.23	120.63
24/06/2013	34.20	43.70	34.64	12.80	44.71	420.08	119.16
21/06/2013	34.96	43.61	35.57	13.14	45.32	426.28	120.69
20/06/2013	34.65	44.50	34.12	13.32	46.15	423.99	121.03
19/06/2013	35.99	45.61	35.02	13.83	45.91	461.69	121.52
18/06/2013	36.45	46.12	35.90	13.57	45.94	479.33	117.27
17/06/2013	36.33	46.17	36.51	13.22	45.30	488.37	115.26
14/06/2013	36.20	46.14	36.48	13.32	45.36	480.99	115.64
13/06/2013	36.29	45.24	37.32	13.44	45.29	493.12	116.49
12/06/2013	36.00	44.80	37.22	13.24	45.66	489.36	117.92
11/06/2013	35.76	44.59	37.51	13.49	46.28	481.73	119.31
10/06/2013	36.26	46.70	38.22	14.14	45.87	495.22	117.49
07/06/2013	36.40	46.53	36.88	13.98	45.72	474.13	116.20
06/06/2013	36.91	46.74	37.13	13.89	46.56	461.05	116.58
05/06/2013	36.33	46.68	37.03	13.70	45.93	465.48	116.76
04/06/2013	36.40	47.26	37.69	13.94	47.82	481.39	115.22
03/06/2013	37.08	46.81	37.60	14.30	47.98	484.96	115.72
31/05/2013	37.72	47.00	37.83	14.93	47.20	471.47	115.36
30/05/2013	36.57	46.26	37.48	14.79	47.50	474.21	115.83
29/05/2013	36.04	46.30	36.91	14.37	47.46	466.97	115.45
28/05/2013	36.67	48.79	37.62	14.30	47.50	452.78	114.81
27/05/2013	36.86	48.95	36.01	14.20	47.50	447.20	113.04
24/05/2013	37.15	48.82	36.24	14.35	47.07	447.00	114.09
23/05/2013	37.17	48.45	36.99	14.23	46.42	450.57	113.25
22/05/2013	36.31	48.48	36.69	14.27	46.53	446.55	113.29
21/05/2013	36.03	48.36	37.21	14.59	45.84	460.66	112.21
20/05/2013	36.10	49.92	36.28	14.88	47.53	476.79	112.12

Tabla 1.22 (Continuación)

Fecha	WM	SOR	BB	CX	CM	EK	GMO
17/05/2013	36.58	50.25	37.55	15.22	48.32	495.71	112.62
16/05/2013	36.32	50.30	37.58	15.14	48.98	502.63	111.95
15/05/2013	35.99	50.60	38.04	15.14	48.44	513.34	111.47
14/05/2013	36.75	50.70	38.40	14.70	49.79	523.10	111.16
13/05/2013	37.03	50.72	37.49	14.43	47.94	544.61	110.86
10/05/2013	36.37	50.28	38.16	14.51	48.09	546.25	110.23
09/05/2013	35.85	49.99	38.86	13.96	47.37	543.11	109.36
08/05/2013	36.04	50.20	38.85	14.20	47.59	533.25	109.79
07/05/2013	36.75	50.19	39.44	14.47	47.90	536.52	109.75
06/05/2013	37.94	50.31	38.30	14.49	46.98	539.13	110.41
03/05/2013	38.82	49.89	39.19	14.40	47.03	534.94	110.32
02/05/2013	38.33	48.89	39.35	13.85	46.90	524.75	111.03
01/05/2013	38.64	49.99	39.51	13.69	46.88	523.28	110.48
30/04/2013	38.64	49.99	39.51	13.69	46.88	523.28	110.48
29/04/2013	37.85	49.83	41.00	13.69	46.69	517.17	111.10
26/04/2013	38.29	49.45	40.16	14.06	46.96	514.50	110.64
16/08/2013	35.57	42.31	42.71	15.58	51.33	470.39	116.42
15/08/2013	35.54	42.34	43.45	15.53	52.31	476.17	116.42
14/08/2013	35.06	42.25	44.29	15.87	53.15	482.47	117.00
13/08/2013	35.20	41.06	44.31	15.79	54.55	489.95	117.00
12/08/2013	34.64	42.78	44.61	15.40	56.13	508.86	117.35
09/08/2013	35.23	41.80	45.80	15.37	56.44	510.95	118.00
08/08/2013	34.81	42.00	45.49	15.47	56.19	500.88	115.87
07/08/2013	35.05	41.90	45.13	15.11	55.50	491.26	117.31
06/08/2013	34.42	42.00	44.74	15.28	54.48	495.05	114.02

## 1.9 Clases de frecuencia

Hasta ahora hemos trabajado con muestras que tienen pocos elementos, pero, ¿qué pasará cuando la cantidad de datos sea considerable y solo se requiera un resumen más compacto del conjunto de datos, o incluso tener una representación gráfica de su comportamiento? Además, si se trata de un conjunto que tiene una gran cantidad de datos (por ejemplo, 10000 o más), visualizarlos todos para poder estudiar su distribución no es muy factible, por consiguiente, será necesario emplear alguna otra estrategia de análisis para los datos.

Podemos resolver el problema anterior con facilidad si distribuimos los datos mediante intervalos o clases. Pero, ¿qué es un intervalo o una clase?

Dado un conjunto de datos se llama **intervalos de clase, clases de frecuencia** o simplemente **clases** a los intervalos que por parejas son ajenos o disjuntos y contienen a todos los datos del conjunto.

Por ejemplo, sea un gerente que tiene la información sobre los lugares a los que concurren 4000 personas a comprar y desea analizarlos. Para poder interpretarlos, descompone los datos en clases de frecuencia bajo ciertos criterios, uno de éstos podría ser las distancias de localización, otro los nombres de los centros comerciales a los que concurre la gente.

En estos momentos hemos utilizado la palabra frecuencia, pero ¿qué entendemos por ésta?

Dado un conjunto de datos llamamos **frecuencia absoluta de clase**, **frecuencia absoluta** o simplemente **frecuencia** a la cantidad de observaciones estadísticas que pertenecen a la clase, y la denotamos con  $n_i$ , para la clase  $i$ . De la misma manera, llamamos **frecuencia relativa de clase** o **frecuencia relativa**, al cociente de dividir la frecuencia absoluta entre la cantidad total de elementos, y la denotamos con  $f_i$ , para la clase  $i$ .

### Ejemplo 1.34 Frecuencias

En un club se entrevista a 150 miembros para conocer su profesión, se obtuvieron estos resultados: 40 administradores, 25 contadores, 20 médicos, 35 ingenieros y 30 informáticos. Las frecuencias absolutas en estas condiciones se muestran en la tabla 1.23.

Tabla 1.23

$n_1 = 40$ para los administradores	$n_4 = 35$ para los ingenieros
$n_2 = 25$ para los contadores	$n_5 = 30$ para los informáticos
$n_3 = 20$ para los médicos	

Las frecuencias relativas están dadas por:

$$f_1 = \frac{n_1}{n} = \frac{40}{150} = \frac{4}{15} \text{ para los administradores}$$

$$f_4 = \frac{n_4}{n} = \frac{35}{150} = \frac{7}{30} \text{ para los ingenieros}$$

$$f_2 = \frac{n_2}{n} = \frac{25}{150} = \frac{1}{6} \text{ para los contadores}$$

$$f_5 = \frac{n_5}{n} = \frac{30}{150} = \frac{1}{5} \text{ para los informáticos}$$

$$f_3 = \frac{n_3}{n} = \frac{20}{150} = \frac{2}{15} \text{ para los médicos}$$

Tanto en estadística como en probabilidad, la acumulación de frecuencias tiene un interés particular.

Llamamos **frecuencia acumulada** a la función que representa la suma de las frecuencias por clase, y se denota por  $F$ . De igual manera, llamamos **frecuencia relativa acumulada** a la función que representa la suma de las frecuencias relativas por clases, la denotamos por  $F_r$ .

## Cálculo de las frecuencias acumuladas

Suponga que tenemos un conjunto con  $n$  datos y lo dividimos en  $m$  intervalos de clase con frecuencias  $n_1, n_2, \dots, n_m$ , tales que  $n_1 + n_2 + \dots + n_m = n$  cantidad total de datos. En estas condiciones, la frecuencia acumulada está dada por:

$$F(x) = \sum_{i=1}^{x_i \leq x} n_i$$

En este caso, las frecuencias relativas por clase para la *frecuencia relativa acumulada* son:

$$f_1 = \frac{n_1}{n}, f_2 = \frac{n_2}{n}, \dots, f_m = \frac{n_m}{n}$$

Por tanto, la frecuencia relativa acumulada está dada por:

$$F_r(x) = \sum_{i=1}^{x_i \leq x} f_i = \frac{1}{n} \sum_{i=1}^{x_i \leq x} n_i = \frac{1}{n} F(x)$$

La frecuencia relativa acumulada es el cociente de la frecuencia acumulada de clase entre la cantidad total de datos, donde la suma de todas las frecuencias relativas debe ser igual a 1.

Veamos el siguiente ejemplo para calcular frecuencias acumuladas.

### Ejemplo 1.35 Frecuencias acumulables y relativas

Con las frecuencias del ejemplo 1.34 vamos a obtener las frecuencias acumuladas y las frecuencias relativas acumuladas (véase tabla 1.24).

Tabla 1.24

Profesión	Frecuencia, $n_i$	$F(x) = \sum_{i=1}^{x_i \leq x} n_i$	$F_r(x) = \sum_{i=1}^{x_i \leq x} f_i$
Administrador	40	40	8/30
Contador	25	65	13/30
Médico	20	85	17/30
Ingeniero	35	120	24/30
Informático	30	150	1

## Distribución de frecuencias para variables cuantitativas

En el caso de variables cuantitativas, la distribución de frecuencias es un poco más elaborada, puesto que requerimos de algunos conceptos para obtener una buena distribución de frecuencias que den la mayor información posible de los datos con el mínimo esfuerzo. En esta parte iniciamos desde la determinación de clases, seguida de la amplitud y la construcción de éstas.

### Cantidad de clases para un conjunto de datos cuantitativos

Una pregunta que surge con frecuencia al hablar de clases se refiere a la cantidad de intervalos de clase en que se recomienda dividir los datos. La respuesta es sencilla, cuando se consideran  $n$  datos ;no existe regla determinante!, para obtener un número que se considere apropiado para la cantidad de clases de frecuencia. En la práctica, los investigadores emplean diferentes reglas para encontrar esta cantidad. En la búsqueda, el único requisito que debemos conservar es que la cantidad de clases no sea una cantidad excesiva ni pequeña.

Entre las reglas más comunes tenemos:

- Una regla empírica que consiste en determinar el entero más cercano a  $\sqrt{n}$ , en donde  $n$  es el número total de observaciones.
- La regla que considera el entero más cercano a  $\log_2(n)$ ,  $n$  número total de observaciones.
- La llamada regla de Sturges, donde la cantidad de clases se toma como el entero más cercano a  $1 + \frac{10}{3} \log(n)$ , con  $n$  cantidad total de observaciones.
- En este texto emplearemos una cantidad de intervalos entre cinco y 25 según sea el valor de  $n$ .

### Ejemplo 1.36 Cantidades de clases

Suponga que se tiene una muestra de tamaño  $n = 200$  observaciones, la cantidad de clases que se puede considerar en este caso son:

- Con la primera regla:  $\sqrt{200} \approx 14.14$ , entonces se recomienda 14 clases.

- Con la segunda regla:  $\log_2(200) \approx 7.64$ , entonces se recomienda ocho clases.
- Con la tercera regla:  $1 + \frac{10}{3} \log(200) \approx 8.67$ , entonces se recomienda nueve clases.
- La cuarta regla es subjetiva.

En este ejemplo se puede apreciar que el problema de calcular cantidad de clases depende en gran medida de la regla utilizada y no es determinante saber cuál utilizar. Más adelante veremos que una de las aplicaciones importantes de las clases de frecuencia se refiere a poder realizar una exposición breve en una gráfica que ilustre el comportamiento de la muestra, entonces proporcionar una gran cantidad de clases no es recomendable, como tampoco lo es mostrar tres o cuatro, ya que no podríamos obtener información confiable en cuanto a la forma de la muestra se refiere.

## Amplitud o longitud de clase para datos cuantitativos

En las variables cuantitativas para las clases se usan los intervalos (no los valores de los atributos), los cuales pueden ser: cerrados, abiertos o semiabiertos.

Así, una clase  $k$  de frecuencias tiene dos límites, que llamaremos inferior y superior y denotaremos por  $l_i^{(k)}$  y  $l_s^{(k)}$ , respectivamente. Entonces la amplitud de clase se define de la siguiente forma.

Llamaremos **amplitud** o **longitud de clase** a la cantidad que denotaremos por  $\ell$  y se calcula

$$\ell = \text{Límite superior menos límite inferior de clase}$$

### Ejemplo 1.37 Amplitud o longitud de clase

Suponga que deseamos revisar la edad de los alumnos de la universidad y planteamos las clases, donde una de éstas está dada por  $[17, 19]$ , es decir la clase que tiene alumnos entre 17 y 19 años. Luego,  $l_i^{(k)} = 17$  y  $l_s^{(k)} = 19$ , la amplitud o longitud de esta clase es  $l_s^{(k)} - l_i^{(k)} = 19 - 17 = 2$ .

La amplitud de clase puede variar de clase en clase, aunque es preferible tratar con clases de igual amplitud. Esto se puede hacer al determinar el rango de los datos y dividir entre la cantidad de clases deseadas. El ejemplo 1.38 ilustra cómo obtener clases de frecuencias con la misma longitud.

### Ejemplo 1.38 Amplitud de clase

Suponga que deseamos revisar la edad de los alumnos de la universidad y queremos conocer la longitud de clase, restringiéndonos a que serán de igual longitud. Ahora, el alumno más joven de la universidad tiene 17 años, mientras que el mayor, 37 años. Deseamos tener  $m = 10$  clases, luego la amplitud de clase,  $\ell$ , estará dada por:

$$\ell = \frac{37 - 17}{10} = \frac{20}{10} = 2$$

## Construcción de clases de frecuencia para datos cuantitativos

En la construcción de clases de frecuencia existen diferentes técnicas y, al igual que en la elección de la cantidad de clases, no existe un método determinante o fórmula general que se emplee para la construcción.

En el texto se construyen las clases de frecuencia de un conjunto de datos  $\{x_1, x_2, \dots, x_n\}$  de acuerdo con los siguientes puntos.

1. Calculamos el rango o amplitud del conjunto de datos,  $r$ .
2. Dividimos el rango entre la cantidad de clases,  $m$ , que se quiere tener y el valor calculado será la longitud de clase ( $\ell$ ) en las que se distribuirán los datos.
3. Formación de clases; las *clases* son abiertas en los extremos izquierdos de los intervalos y cerradas en los extremos derechos, considerando la primera clase en ambos extremos cerrada.

A continuación, se muestra un ejemplo de cómo emplear estos tres pasos para la construcción de clases de frecuencia.

En la construcción de clases de frecuencia, siempre debemos respetar que cumplan con los aspectos siguientes:

- a) Un mismo dato no debe pertenecer a dos clases diferentes.
- b) Todos los datos deben estar distribuidos en las clases formadas.

### Ejemplo 1.39 Construcción de clases de frecuencia

Sea un conjunto de datos, donde el valor más pequeño es 5 y el más grande 75, construya 10 clases de frecuencia.

#### Solución

Para este caso, seguimos los tres pasos numerados arriba.

1. Cálculo del rango. De las condiciones del problema, el rango de los datos es:  $r = 75 - 5 = 70$ .
2. Longitud de clase. Como se quiere tener  $m = 10$  intervalos de clase, el rango 70 se divide entre 10, y se obtiene  $\ell = 7$ . Este valor es la longitud de las clases de frecuencia.
3. Formación de clases. Las 10 clases son:

$[5, 12], (12, 19], (19, 26], (26, 33], (33, 40], (40, 47], (47, 54], (54, 61], (61, 68], (68, 75]$ .

Cabe recordar que un intervalo de la  $[26, 33]$ , indica que se consideran todos los valores que estén entre 26 y 33, incluyendo el 33 y excluyendo el 26.

Después de construir los intervalos de clase, procedemos a calcular sus frecuencias absolutas, relativas o acumuladas.

En el siguiente ejemplo se detalla todo el proceso para distribuir los datos muestrales en clases de frecuencia.

### Ejemplo 1.40 Clases de frecuencias

Considere las calificaciones (con escala de 0 a 100) de 80 estudiantes en la materia de vibraciones mecánicas, distribuya en siete clases de frecuencias las calificaciones y calcule las relativas.

Tabla 1.25

30	88	96	100	45	38	78	89	68	88
68	100	100	68	69	79	98	94	30	46
30	86	85	89	94	99	100	45	30	35
36	76	78	81	80	40	67	58	89	58
98	90	100	100	68	70	83	85	68	56
30	67	78	98	100	86	69	79	52	45
89	78	65	60	69	76	78	77	89	98
99	91	100	48	68	84	67	69	46	79

**Solución**

En este ejemplo notamos que si se emplea la regla de Sturges para los 80 datos, resulta  $m = 1 + \frac{10}{3} \log(80) = 7.34 \approx 7$  clases de frecuencia.

Ahora construimos las clases de frecuencia.

- 1. El rango.** El valor más pequeño es 30 y el más grande 100. Luego,  $r = 100 - 30 = 70$ .
- 2. La longitud de clase.** Se desean  $m = 7$  clases de frecuencias. Así,  $\ell = 70/7 = 10$ .
- 3. Los intervalos de clase.** El primer intervalo es  $[30, 40]$ , así que agregamos al extremo derecho la longitud de clase  $\ell = 10$ , se obtiene  $(40, 50]$ , y así sucesivamente.

Para la frecuencia de cada intervalo se cuentan los datos que estarán en éste. Para la primera clase  $[30, 40]$ , los datos deben ser mayores o iguales a 30, pero menores o iguales a 40, así resultan: 30, 38, 30, 30, 30, 35, 36, 40 y 30; en total nueve datos. Este proceso de conteo se mantiene hasta la última **clase**.

Por último, calculamos las frecuencias relativas por clases, dividiendo las frecuencias entre la cantidad total de datos; en este caso, 80. Con lo que obtenemos los resultados de la tabla 1.26.

**Tabla 1.26** Clases de frecuencia del ejemplo 1.40

Clase $i$	Intervalo $i$	Conteo	Frecuencia $n_i$	Frecuencia relativa $f_i = \frac{n_i}{n}$	$F_i(x) = \sum_{j=1}^{x_i \leq x} f_j$
1	[30, 40]	30, 38, 30, 30, 30, 35, 36, 40, 30	9	$\frac{9}{80} = 0.1125$	0.1125
2	(40, 50]	45, 46, 45, 45, 46, 48	6	$\frac{6}{80} = 0.0750$	$0.1125 + 0.0750 = 0.1875$
3	(50, 60]	58, 58, 56, 52, 60	5	$\frac{5}{80} = 0.0625$	$0.1875 + 0.0625 = 0.2500$
4	(60, 70]	68, 68, 68, 69, 67, 68, 68, 67, 69, 65, 69, 68, 67, 69, 70	15	$\frac{15}{80} = 0.1875$	$0.2500 + 0.1875 = 0.4375$
5	(70, 80]	78, 79, 76, 78, 78, 79, 78, 76, 80, 78, 77, 79	12	$\frac{12}{80} = 0.1500$	$0.4375 + 0.1500 = 0.5875$
6	(80, 90]	88, 89, 88, 86, 85, 89, 81, 89, 83, 85, 86, 89, 89, 90, 84	15	$\frac{15}{80} = 0.1875$	$0.5875 + 0.1875 = 0.7750$
7	(90, 100]	96, 100, 100, 100, 98, 94, 94, 99, 100, 98, 100, 100, 98, 100, 98, 99, 91, 100	18	$\frac{18}{80} = 0.2250$	$0.7750 + 0.2250 = 1$
Total			80	1	

**Ejercicios 1.6**

- En la tabla 1.27 se observan los tiempos de llegada (en minutos) de 60 aviones que arriban al aeropuerto Benito Juárez de la Ciudad de México.

**Tabla 1.27**

2.6	3.9	4.5	4.0	3.7	3.2	5.7	4.3	3.8	3.6	8.0	5.6
4.7	6.1	6.0	5.0	4.5	6.2	3.4	2.9	3.6	4.1	3.9	4.6
2.5	2.8	3.2	3.1	4.6	5.2	6.1	4.5	4.1	3.8	4.8	5.9



Tabla 1.27 (Continuación)

7.2	3.4	7.9	3.6	3.6	4.8	5.2	6.3	8.2	5.3	6.2	3.2
3.9	4.6	4.5	5.7	4.8	6.9	6.3	2.6	2.5	6.8	4.5	5.0

Distribúyalos en cinco clases.

- Una máquina despachadora de refrescos de un centro comercial parece tener una falla, dado que el encargado recibió varias quejas en la última semana. Entonces, decide registrar la cantidad de contenido en 40 vasos despachados por la máquina y dividirlos en siete clases de igual longitud; si 85% o más de los refrescos despachados se encuentra en las clases 2 a 6, el encargado deberá conservar la máquina; en caso contrario, deberá enviarla a reparar. Los valores medidos en mililitros se encuentran en la tabla 1.28.

Tabla 1.28

245.6	236.9	240.7	235.9	247.8	246.5	230.8	250.6	248.0	247.4
238.6	240.0	246.9	258.9	245.6	248.5	246.8	245.6	247.8	256.0
243.0	243.3	240.6	250.2	249.6	243.8	246.9	247.8	243.0	246.4
230.5	228.9	235.7	248.9	248.9	245.7	240.8	246.8	246.2	250.0

Divida los valores en siete clases de frecuencia de igual longitud, calcule sus frecuencias relativas e indique si el encargado deberá enviar a reparar la máquina o la dejará seguir trabajando.

- Para estudiar el tiempo de vida de personas enfermas con SIDA se analizó una muestra de 90 pacientes, se anotó su duración de vida en meses y se obtuvieron los resultados de la tabla 1.29.

Tabla 1.29

34.0	28.5	18.0	34.9	25.8	16.9	15.8	19.0	11.5	25.9	38.9	34.0	16.8	27.8	26.5
24.6	22.8	16.8	39.0	42.0	48.0	34.8	33.0	23.9	27.5	35.8	36.9	26.7	26.8	34.7
35.9	25.8	24.8	45.8	18.9	35.8	35.8	46.9	36.8	35.9	52.0	33.6	24.8	25.9	26.8
26.8	29.4	37.8	35.9	10.8	25.8	35.8	26.8	25.7	26.9	27.9	38.5	35.8	30.2	28.6
33.1	34.7	45.9	56.8	45.8	25.8	50.2	42.9	46.8	48.9	47.5	48.2	42.5	40.8	27.9
24.8	46.8	40.7	18.9	22.0	29.5	31.9	48.2	34.8	47.2	27.0	39.8	45.8	40.4	38.2

Ordene en 10 clases de frecuencia y calcule la media de los datos.

## 1.10 Gráficos

La variabilidad es una de las características del conjunto de datos que se analizan y es muy difícil de interpretar cuando se analizan todos los datos en conjunto. Sin embargo, en una gráfica la interpretación de la muestra es mucho más sencilla.

Las gráficas a las que se hacen referencia en la estadística descriptiva deben mostrar la distribución de las frecuencias o frecuencias acumuladas del conjunto de datos, ya que con éstas podremos entender e interpretar con facilidad su comportamiento.

La presentación gráfica de las observaciones es muy usada en informes, artículos periodísticos y presentaciones públicas, ya que tiene un efecto visual difícil de conseguir con otras presentaciones de información. En efecto, en ocasiones, basta con observar un gráfico para detectar con facilidad una posible tendencia en el tiempo, sesgo, simetría, etcétera, para saber cuál de los sectores involucrados es el más importante en una situación dada. El efecto visual de los gráficos se obtiene gracias a que, en general, usan las nociones conocidas de área y volumen, que de manera usual maneja el investigador. Sin embargo, éste debe ser cuidadoso en la construcción de los gráficos ya que esa misma facilidad para verlos puede ser engañosa y obtener conclusiones erróneas en caso de que el gráfico esté mal construido.

Un gráfico es un instrumento cuyo objetivo es presentar datos numéricos mediante magnitudes geométricas, es decir, longitudes, áreas, volúmenes, etcétera. La presentación gráfica de la información numérica se basa en un sistema de coordenadas en el que se ubican los datos.

El gráfico tiene estas ventajas:

- Presenta una idea general de manera atractiva.
- Permite comparar una gran cantidad de valores de modo muy eficiente.

Sin embargo, también presenta estas desventajas:

- Muestra valores aproximados.
- No se puede incluir tanta información como en una tabla de frecuencias.

Por lo anterior, es necesario introducir un método gráfico para la interpretación de datos. Los gráficos más comunes se clasifican en los siguientes tipos:

- Barras o histogramas
- Lineales-poligonales
- Tallo-hoja
- Pastel
- Caja-box

En este texto revisaremos los histogramas y gráficos lineales-poligonales. Antes de iniciar la construcción de los tipos de gráficos mencionados, note que en la construcción de clases de frecuencia para la elaboración de gráficos no existen reglas estrictas ni criterios uniformes a los que debe apegarse una presentación gráfica de las observaciones, ya que debe hacerse de acuerdo con los datos que se usan y su finalidad; no obstante, se recomienda que en un reporte de trabajo o investigación un gráfico cumpla con ciertos criterios.

El gráfico debe explicarse por sí mismo, para ello se recomienda que esté compuesto de:

1. **Número:** para identificar si hay más de uno.
2. **Título:** define el qué, dónde, cómo y cuándo de la información.
3. **Fuente:** incluye el origen de la información utilizada, puede servir para informar al usuario sobre el lugar donde puede obtener más datos al respecto.
4. **Leyenda:** cuando en un gráfico se incluyen varias series de datos es necesario identificar cada una de estas mediante símbolos o leyendas.
5. **Escala:** identificar la unidad de medida correspondiente con los valores en ambos ejes, por ejemplo, 1 cm = 1 000 nacimientos.
6. **Nota introductoria y nota al pie:** se utilizan si son necesarias.
7. **Título de los ejes:** se utiliza para identificar cada uno de los ejes.

En la figura 1.3 se ilustra la presentación general de un gráfico.

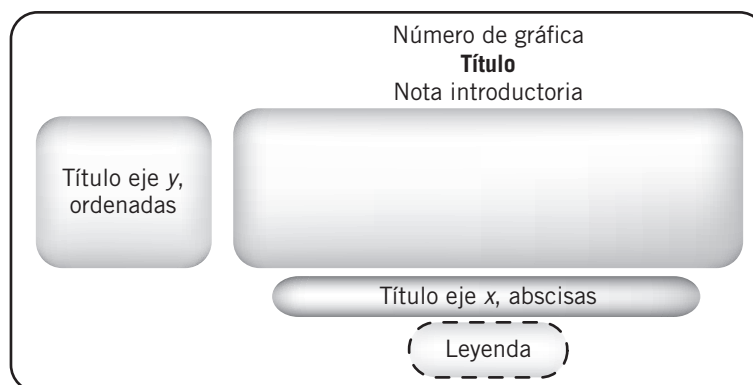


Figura 1.3

## Histogramas

Uno de los gráficos que más se emplea en la práctica es el que se elabora mediante una representación de las frecuencias absolutas o relativas o la acumulada por medio de barras.

Un **histograma** es un gráfico de barras que se utiliza para representar la forma en que están distribuidas las frecuencias, con éste podemos identificar el centro y la variabilidad de los datos.

Para facilitar la construcción de un histograma se recomienda usar solo intervalos de clase de igual longitud, ya que las frecuencias de las clases se grafican de manera proporcional a las alturas de los rectángulos; además, con el histograma es mucho más fácil comparar las diferencias entre frecuencias cuando los rectángulos tienen la misma base.

### Construcción de gráficos de barras para variables cuantitativas

1. Se construyen los intervalos de clase.
2. Se encuentra el punto medio de cada intervalo de clase.
3. En el eje de las abscisas del plano cartesiano se distribuyen los puntos medios de las clases de frecuencia, mientras que en el eje de las ordenadas se distribuyen las frecuencias de los datos. Por último, el histograma se construye al graficar una barra por cada clase, cuyo centro es el punto medio de ésta, de tal manera que la altura de la barra es la frecuencia o frecuencia relativa y la base de los rectángulos estará definida por los límites de cada clase.

### Ejemplo 1.41 Histograma

Construya un histograma para las clases de frecuencia y un histograma para la frecuencia acumulada de los datos de la tabla 1.30.

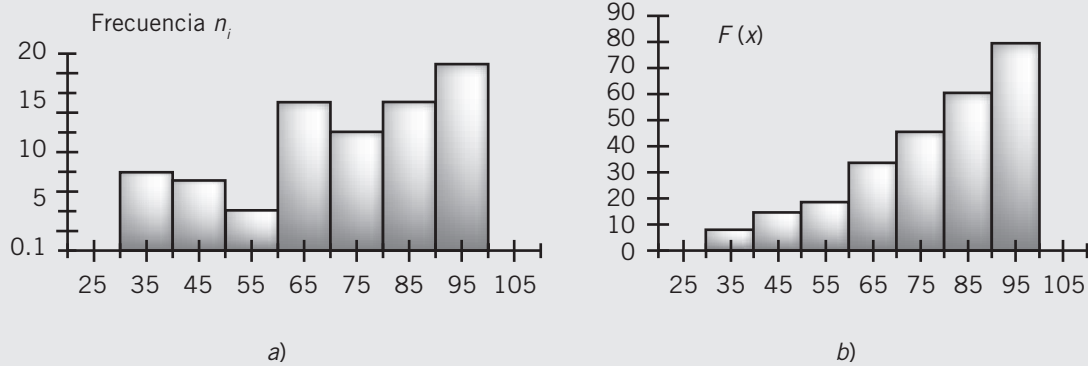
#### Solución

En este caso ya se conocen las clases de frecuencia y las marcas de clase (véase tabla 1.30).

**Tabla 1.30** Frecuencias para el ejemplo 1.41.

Clase $i$	Intervalo $i$	Punto medio $x_i^m$	$n_i$	$F(x) = \sum_{j=1}^{x_i} n_j$	$f_i = \frac{n_i}{n}$	$F_i(x) = \sum_{j=1}^{x_i} f_j$
1	[30, 40]	35	8	8	0.1000	0.1000
2	(40, 50]	45	7	15	0.0875	0.1875
3	(50, 60]	55	4	19	0.0500	0.2375
4	(60, 70]	65	15	34	0.1875	0.4250
5	(70, 80]	75	12	46	0.1500	0.5750
6	(80, 90]	85	15	61	0.1875	0.7625
7	(90, 100]	95	19	80	0.2375	1.0000

Luego, se grafican los puntos medios de los intervalos (columna 3) y se trazan los rectángulos cuya base es igual a la longitud de la clase y su frecuencia a la altura (véase figura 1.4).



**Figura 1.4** a) Histogramas para las clases de frecuencia. b) Histogramas para las frecuencias acumuladas.

Para las frecuencias relativas el histograma es el mismo, solo se divide cada frecuencia entre el total de datos.

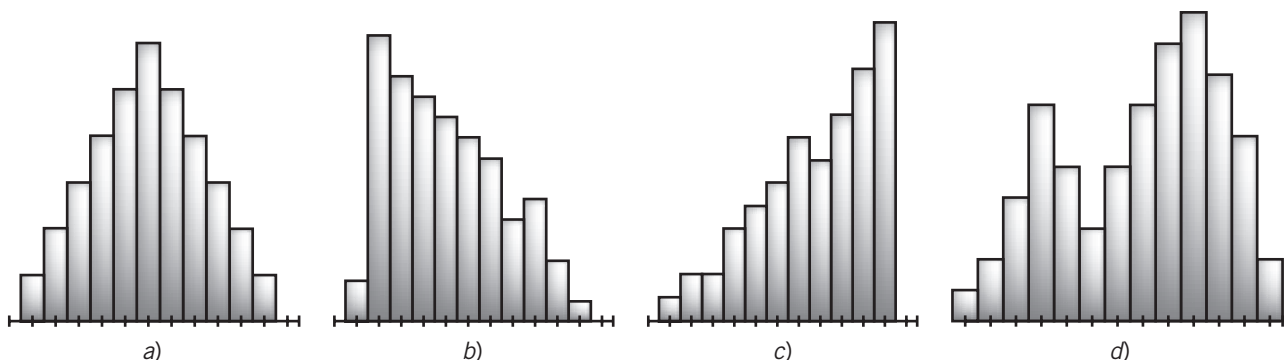
Como se mencionó antes, los histogramas no solo ayudan a ubicar el centro y a visualizar la variabilidad de los datos, sino también la forma en que se distribuyen. Por tanto, se clasifican en:

- a) **Simétricos.** Histogramas en los que su distribución es parecida a una campana. Es decir, la mitad izquierda es una imagen reflejada de la mitad derecha (véase figura 1.5a).
- b) **Sesgados, izquierda y derecha.** Histogramas en los que la distribución de alguna de las colas está más alargada en comparación con la otra. Se llaman sesgados a la derecha o positivamente sesgado si la cola derecha es la que está más alargada (véase figura 1.5b). En caso de que la cola izquierda sea la más alargada, se llaman sesgados a la izquierda o negativamente sesgados (véase figura 1.5c).

Ahora bien, ¿qué entendemos por sesgo en una distribución?

- c) **Multimodales.** Histogramas que tienen en su distribución más de un pico (véase figura 1.5d). En caso de tener dos picos se llaman **bimodal**; si tiene tres, **trimodal**, y así sucesivamente.

Pero, ¿cómo es un histograma sesgado y cuál es su diferencia con un multimodal?



**Figura 1.5** Histogramas para las clases de frecuencia: a) simétrico, b) sesgado a la derecha, c) sesgado a la izquierda y d) multimodal (bimodal).

## Gráficos lineales, polígonos de frecuencias

En ciertas áreas de estudio se requiere que las representaciones gráficas de la distribución de las frecuencias de datos se hagan con líneas en lugar de barras. Por ejemplo, en un estudio sobre los pronósticos de algún evento, la distribución de sus frecuencias y sus tendencias se visualiza mejor si se unen sus puntos medios con segmentos

rectilíneos. En este tipo de gráfico, tanto en su escala horizontal como en la vertical son aritméticas (distancias iguales representan magnitudes iguales).

Un **polígono de frecuencias** es un gráfico de línea que representa las frecuencias de los datos con la unión por líneas de los puntos medios de cada intervalo  $(x_i^m, f_i)$ , o  $(x_i^m, n_i)$ . Donde  $x_i^m$  es el punto medio de la clase  $i$ ;  $f_i$  su frecuencia relativa y  $n_i$  su frecuencia absoluta. Debido a su forma también se le suele llamar **gráfico poligonal**.

#### Construcción de un gráfico poligonal

1. Se crean los intervalos de clase.
2. Se encuentra el punto medio de cada intervalo de clase.
3. En el plano cartesiano, los puntos medios de las clases de frecuencia se distribuyen en el eje de las abscisas, mientras que las frecuencias de los datos, en el eje de las ordenadas. Por último, se construye el gráfico poligonal uniendo los puntos medios de cada intervalo de clase

En el siguiente ejemplo se ilustra cómo construir un gráfico poligonal.

#### Ejemplo 1.42 Gráfico poligonal

Construya un polígono de frecuencias para las clases de frecuencia del ejemplo 1.41.

#### Solución

En este caso también tenemos las clases de frecuencia y sus puntos medios; así, con ayuda de la tabla 1.30 si graficamos los puntos  $(x_i^m, n_i)$  en el plano cartesiano y luego, mediante el uso de las columnas correspondientes a  $x_i^m$  y  $n_i$ , trazamos los segmentos que unen estos puntos (véase figura 1.6).

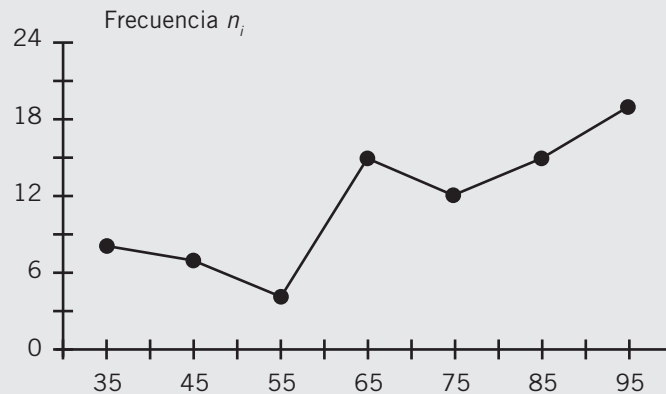


Figura 1.6 Polígono de frecuencias.

Los polígonos de frecuencia tienen un amplio uso en el análisis de las series de tiempo, puesto que es común en esta área de estudio que se desee conocer la tendencia de la distribución de los datos. También se utilizan en ciertas situaciones cuando se quieren comparar las distribuciones de dos o más conjuntos de datos, ya que resulta más factible hacerlo por medio de los polígonos de frecuencias que mediante histogramas, esto se debe a que los primeros pueden superponerse, y los segundos no resulta una buena visión de la comparación. Entonces, los histogramas no se suelen usar para hacer comparaciones.

#### Ejercicios 1.7

Trace el histograma y gráfica poligonal para cada uno de los tres ejercicios de la lista de ejercicios 1.6.

## Ejercicios de repaso

## Preguntas de autoevaluación

- 1.1 ¿Cuál es la diferencia entre un parámetro y un estadístico?
- 1.2 Conteste las siguientes preguntas y explique su respuesta.
- ¿Es cierto que la media se aplica a cualquier escala de medición?
  - ¿Las medidas centrales representan qué tan disperso están el centro de los datos?
  - ¿Existen algunas medidas centrales que se salen del rango de valores de la muestra?
  - ¿Las medidas de desviación representan un valor que indica qué tan juntos o dispersos están los valores de la muestra?
  - ¿El coeficiente de asimetría es un valor que indica la forma de la distribución de la muestra?
  - ¿Para describir de manera estadística el comportamiento de las observaciones de una muestra es suficiente con calcular todas sus medidas centrales?
- 1.3 ¿Si un investigador tiene una población de tamaño 1 000, entonces considerando cualquier muestra de tamaño 400 tendrá excelentes resultados?
- 1.4 ¿La media geométrica solo se puede usar en los casos en que los valores son positivos?
- 1.5 ¿A qué se debe que el riesgo de un título en una inversión queda bien definido por una medida de dispersión?
- 1.6 Con los conceptos revisados en esta unidad, ¿cuál sería la metodología para presentar un informe estadístico sobre el comportamiento de las observaciones de la muestra?
- 1.7 En las siguientes situaciones indique cómo realizaría un muestreo de tamaño  $n$ , considerando al menos una de las técnicas revisadas durante el desarrollo de la unidad.
- Para los usuarios de la sucursal del banco BM en el estado de Michoacán.
  - Para todos los cuentahabientes del banco BAMEX de toda la República Mexicana.
  - Para una línea de producción de refrescos de la marca CC.
  - Para la producción de autos VK en la ciudad de Puebla.
  - Para los huéspedes de un hotel en la ciudad de Cancún.
  - Para el curso del dólar en el último año.
  - Para todos los aficionados al fútbol en el estado de Veracruz.
  - Cualquier medida central siempre existe.
  - El valor promedio nunca será negativo.
  - Las medidas de desviación pueden llegar a ser negativas, cuando una observación de menor valor se coloca a la derecha de otra de mayor valor.

- k) Nunca puede darse el caso de que una medida de desviación sea cero.

## Ejercicios complementarios con grado de dificultad uno

- 1.8 Dado el conjunto de datos, de la tabla 1.31, analícelos, calcule las medidas que se piden y explique el tipo de datos de que se trata.

Tabla 1.31

34	23	45	43	11	10	23	27	31	21	17	25	25	24	31
31	26	33	37	18	11	16	20	18	19	18	16	28	19	16

- Media, mediana, moda.
- Rango y varianza insesgada.
- Coefficiente de asimetría.

- 1.9 Los datos de la tabla 1.32 muestran los diámetros internos en centímetros de 30 pistones.

Tabla 1.32

12.01	12.04	12.05	12.09	11.99	11.92
11.92	12.01	12.01	12.08	12.02	12.03
12.02	11.84	12.0	11.87	11.93	11.93
11.75	12.01	11.76	12.04	11.95	12.01
11.90	11.95	11.92	12.03	11.98	12.02

Calcule:

- Diámetro interno promedio.
  - Rango y varianza insesgada de los diámetros.
  - Coefficiente de asimetría.
- 1.10 En la tabla 1.33 se muestran las calificaciones de la materia de taller de ética para dos muestras de 30 alumnos elegidos aleatoriamente en dos escuelas.

Tabla 1.33

Muestra 1	8	8	3	5	10	9	4	7	1	3
	8	9	7	7	7	2	3	8	8	9
	7	8	4	5	6	6	10	6	3	8

Muestra 2	10	10	8	0	0	2	8	4	1	4
	8	5	2	10	10	10	9	8	9	2
	3	3	1	1	2	4	8	6	3	8

- 1.11 Calcule las medidas muestrales.
- Media, mediana, moda.
  - Rango y varianza insesgada.
  - Coefficiente de asimetría.

¿Qué muestra resultó ser más homogénea en sus calificaciones? ¿Qué se puede decir de posibles comportamientos de las poblaciones de procedencia de los datos?

- 1.12** El precio en dólares por barril de petróleo crudo exportado por México en 15 días en marzo de 2010, de acuerdo con la tabla 1.34 fue:

**Tabla 1.34**

71.5	71.0	74.0	72.5	72.5	75.2	76.5	74.5							
72.0	71.5	74.0	73.9	76.8	80.6	82.0								

Calcule las medidas de los datos de la muestra.

- Media, mediana, moda.
  - Rango y varianza insesgada.
  - Coefficientes de asimetría.
- 1.13** En un experimento de psicología se pide a varios individuos que memoricen cierta secuencia de palabras. En la tabla 1.35 se registran los tiempos (en segundos) que necesitan los participantes para lograrlo.

**Tabla 1.35**

116	45	57	112	73	129	89	128	100	46	107	109
32	122	41	70	96	98	117					

Calcule el tiempo medio en segundos de los individuos de la prueba para memorizar la secuencia de palabras, la variancia de la prueba para la muestra.

- 1.14** En la tabla 1.36 se muestran los salarios mensuales de 15 empleados de una fábrica de artículos para el hogar:

**Tabla 1.36**

3 174	8 277	6 250	6 300	10 215	8 263	5 260	7 228
8 185	5 208	9 260	4 284	7 243	6 195	7 245	

- Calcule el salario muestral medio y su promedio.
  - Calcule la variancia y desviación estándar muestrales de los salarios.
  - Obtenga el coeficiente de asimetría.
- ¿Qué se puede decir con respecto a la distribución de los salarios semanales de los empleados?

- 1.15** En la tabla 1.37 se muestran las calificaciones de tres muestras de 10 alumnos.

**Tabla 1.37**

<b>Muestra 1</b>	8	5	2	10	10	9	4	7	1	3
<b>Muestra 2</b>	1	2	4	8	6	10	10	8	8	9
<b>Muestra 3</b>	7	8	4	5	6	10	9	8	9	2

Determine qué muestra resultó más homogénea en sus calificaciones e indique en qué medida basa su respuesta.

- 1.16** En estadística actualmente tiene gran auge trabajar con un tipo particular de datos, llamados censurados, que corresponden a un experimento en el cual se prueban los componentes hasta obtener los primeros  $n$  que fallen.

Suponga que se lleva a cabo este experimento con los focos, se prueban hasta que se descompone el cuadragesimo y se obtienen los tiempos de falla en horas de 40 focos en secuencia como se averiaron (véase tabla 1.38).

**Tabla 1.38**

690	696	699	702	710	715	716	719	720	722
722	722	724	726	730	731	734	736	738	741
742	745	745	747	748	750	752	753	753	754
759	760	763	765	767	770	772	775	780	781

Con los datos censurados obtenidos, el investigador pretende hacer un reporte que indique la duración de los focos.

- ¿Qué medidas, de las vistas en esta unidad, le recomendaría al investigador calcular e incluir en su reporte?
  - Calcule las medidas sugeridas para la vida de los focos y redacte un reporte dirigido al gerente de mercadotecnia de la empresa para la duración de los focos.
- 1.17** Se realiza un experimento para medir el porcentaje de encogimiento al secar los especímenes de prueba de arcilla plástica con el que se obtuvieron los resultados de la tabla 1.39.

**Tabla 1.39**

17.2	17.7	16.1	19.9	15.6	19.7	16.4	15.5	17.2	16.4
17.3	15.2	18.5	19.2	17.7	16.5	18.8	17.8	18.3	17.4

Con los datos obtenidos el investigador pretende hacer un reporte que indique al comprador el porcentaje de encogimiento al secar los especímenes de arcilla plástica.

- ¿Qué medidas, de las estudiadas en la unidad, le recomendaría al investigador calcular e incluir en su reporte?
  - Calcule las medidas sugeridas para el porcentaje de encogimiento al secar los especímenes de arcilla plástica y redacte un reporte dirigido a los compradores.
- 1.18** Considere los datos de la tabla 1.40 que corresponden al porcentaje de algodón en el material usado para fabricar playeras de caballero.

**Tabla 1.40**

34.2	33.6	33.8	34.7	37.8	32.6	35.8	34.6
33.1	34.7	34.2	33.6	36.6	33.1	37.6	33.6
34.5	35.0	33.4	32.5	35.4	34.6	37.3	34.1
35.6	35.4	34.7	34.1	34.6	35.9	34.6	34.7
34.3	36.2	34.6	35.1	33.8	34.7	35.5	35.7
35.1	36.8	35.2	36.8	37.1	33.6	32.8	36.8
34.7	36.1	35.0	37.9	34.0	32.9	32.1	34.3
33.6	35.3	34.9	36.4	34.1	33.5	34.5	32.7

Con los datos obtenidos el investigador pretende hacer un reporte que indique al distribuidor el porcentaje de algodón usado para fabricar playeras de caballero.



- a) ¿Qué medidas, de las estudiadas en la unidad, le recomendaría al investigador calcular e incluir en su reporte?
- b) Calcule las medidas sugeridas para el porcentaje de algodón usado para fabricar playeras de caballero y redacte un reporte dirigido al distribuidor.

- 1.19** Calcule la media geométrica de las calificaciones de un examen psicológico aplicado a ocho personas cuyos resultados fueron: 7, 8, 7, 9, 6, 8, 9 y 7.
- 1.20** Los datos de la tabla 1.41 representan los salarios diarios de todos los empleados de una fábrica de artículos para el hogar:

Tabla 1.41

119.75	125.70	125.25	124.80	126.34	116.70	123.18	122.40	128.00	121.42
124.17	122.65	126.03	117.45	127.67	125.00	130.60	121.55	126.23	126.00
122.48	118.50	120.48	126.50	128.24	124.73	119.35	124.35	124.32	128.83

Calcule:

- a) La media y su variancia insesgada.
- b) Debajo de qué valor se encuentra 75% de los sueldos de las 30 personas y su salario medio.
- c) El coeficiente de asimetría.
- 1.21** Calcule la media armónica del viaje redondo que realizó el licenciado Jiménez de México a Querétaro (210 km) si de ida lo recorrió a una velocidad de 130 km/h y de regreso a 110 km/h.
- 1.22** Si el señor Alfredo López viajó 400 km en cuatro tramos de 100 cada uno, con velocidades de 100 km/h, 130 km/h, 90 km/h y 110 km/h, respectivamente. Calcule con base en la media armónica la velocidad media con la que realizó el viaje.
- 1.23** El chofer de nombre David Hernández de la línea de autobuses AUTOU viajó 1000 km en cuatro tramos de 250 cada uno, con velocidades de 92 km/h, 85 km/h, 95 km/h y 80 km/h, respectivamente. Calcule, con base en la media armónica, la velocidad media con la que realizó el viaje.
- 1.24** Suponga que tiene los datos de años de estudio de 10 personas y sueldos actuales (véase tabla 1.42).

Tabla 1.42

Años de estudios	17	21	14	12	23	15	28	16	14	22
Sueldo	6 200	14 500	8 300	4 500	14 800	6 300	19 400	8 500	8 200	16 500

Por medio de la variancia indique qué carácter tiene mayor variabilidad, si los años de estudio o el sueldo.

- 1.25** En el ejercicio anterior muestre si existe dependencia entre los dos caracteres.
- 1.26** Para probar si un tratamiento reductor de peso es eficiente se toma una muestra de 10 personas, se anotan sus pesos en kilogramos antes y después del tratamiento, y se obtienen los datos de la tabla 1.43.

Tabla 1.43

Persona	1	2	3	4	5	6	7	8	9	10
Antes	81	75	74	69	71	72	70	84	75	79
Después	75	72	71	67	68	67	68	75	72	73

Calcule la media y desviación estándar muestral para ambos pesos. Mediante la variancia indique qué carácter tiene mayor variabilidad, si los pesos antes o después del tratamiento.

- 1.27** En el ejercicio anterior muestre si existe dependencia entre los dos caracteres.
- 1.28** Un ingeniero de control de calidad de la producción de una empresa que maquila tapas de plástico tiene que tomar una muestra diaria de la línea de producción para inspeccionarla. El ingeniero decide que su estudio tenga una confianza de 90% y un error de 4%. Calcule el tamaño de la muestra para inspeccionar la producción.
- a) Cuando va iniciando y no tiene información previa.
- b) Cuando se tiene la información de varias revisiones diarias, con las que se ha obtenido una variabilidad positiva de 0.80.
- 1.29** En la planta de producción de autos se lleva a cabo la revisión de 3850 unidades. El gerente decide realizar una inspección del lote por medio de una muestra. Desea que su estudio tenga una confianza de 98% y decide permitir un error de 5%. Calcule el tamaño de la muestra para inspeccionar un lote.
- a) Cuando va iniciando y nunca ha revisado un lote, es decir, que no tiene información previa.
- b) Cuando se han realizado varias revisiones de lotes del mismo tamaño y se ha obtenido una variabilidad positiva de 0.75.
- 1.30** Suponga que se tiene una población de 450 amas de casa sobre su preferencia de una pasta de dientes. Se piensa entrevistar solo una muestra para saber si la utilizan. ¿Cuál deberá ser el tamaño mínimo de muestra que cumpla con un error estándar de 0.04 y una confiabilidad de 85%?

## Ejercicios complementarios con grado de dificultad dos

- 1.31** En un importante laboratorio farmacéutico se llevó a cabo un experimento y se anotaron sus valores  $\sum_{i=1}^{70} x_i = 1406$ , con  $\sum_{i=1}^{70} x_i^2 = 29042$ . Calcule la variancia insesgada de los datos.
- 1.32** Para determinar la dependencia entre dos caracteres se hizo un estudio de 50 y se anotaron sus resultados  $\sum_{i=1}^{50} x_i = 1015$ ,  $\sum_{i=1}^{50} x_i^2 = 21101$ ,  $\sum_{i=1}^{50} y_i = 831$ ,  $\sum_{i=1}^{50} y_i^2 = 14419$  y  $\sum_{i=1}^{50} x_i y_i = 16919$ . Calcule el coeficiente de correlación de los datos muestrales.



1.33 Resuelva el ejercicio anterior si  $\sum_{i=1}^{50} x_i = 1531$ ,  
 $\sum_{i=1}^{50} x_i^2 = 48017$ ,  $\sum_{i=1}^{50} y_i = 6547$ ,  $\sum_{i=1}^{50} y_i^2 = 885881$  y  
 $\sum_{i=1}^{50} x_i y_i = 206171$ .

1.34 Se conoce que  $\sum_{i=1}^{50} x_i = 993$ ,  $\sum_{i=1}^{50} y_i = 3592$  y  $\sum_{i=1}^{50} x_i y_i = 76417$ . Calcule la covarianza de los valores muestrales para  $x$  y  $y$ .

1.35 Calcule con los datos de la tabla 1.44:

- Rendimientos de los títulos, rendimientos promedio de cada título y riesgos de cada título.
- Las varianzas de cada título e indique cuál tiene precios más dispersos.
- Con el portafolio de los dos títulos, calcule el rendimiento promedio del portafolio, para una inversión de 40 y 60%, respectivamente. En la inversión calcule el riesgo del portafolio.

Tabla 1.44

WM	EK	WM	EK
34.91	68.90	34.41	63.91
34.74	67.83	34.13	63.32
34.71	67.67	33.93	63.31
34.63	67.55	33.73	63.89
34.85	67.50	33.32	64.34
34.84	65.16	33.59	60.11
34.59	65.00	33.60	60.01
34.65	65.98	33.12	58.65
34.68	64.30	33.58	67.01
34.50	64.34	34.44	67.04

- 1.36 Los fabricantes de cierta marca de llantas quieren saber la duración promedio de su producto, según el uso de diferentes automovilistas, para lo cual consideran una muestra aleatoria de 100 de sus compradores. Cuando sus llantas quedaron en desuso reportaron al fabricante la duración de las llantas en miles de kilómetros y obtuvieron los datos de la tabla 1.45. Divida el conjunto de datos en 10 clases de frecuencias y:
- Trace el histograma de frecuencias.
  - Trace el polígono de frecuencias.

Tabla 1.45

55.3	59.5	60.0	48.6	59.1	63.5	56.3	55.0	53.7	52.8
50.5	56.7	60.8	67.6	68.0	64.4	58.0	49.9	65.4	47.9
45.2	68.1	56.5	50.5	51.2	55.9	61.8	73.0	65.3	60.0
56.6	57.3	49.9	69.5	50.2	52.1	56.7	56.2	52.9	55.0

49.8	51.4	56.8	60.1	56.7	55.9	55.2	65.0	54.8	50.2
56.7	67.0	58.8	57.9	49.9	50.6	58.6	54.8	53.8	52.0
52.8	51.9	61.0	62.5	64.2	67.1	59.9	58.1	56.7	54.0
56.3	53.9	52.0	52.9	51.9	56.0	58.1	52.0	57.0	56.1
49.9	61.0	62.5	51.8	50.1	50.8	60.2	57.8	53.2	51.8
60.1	60.9	56.8	48.0	58.9	57.6	59.7	60.7	63.6	65.3

1.37 En la tabla 1.46 se muestran los salarios diarios de todos los empleados de una fábrica de artículos para el hogar.

Tabla 1.46

120.15	125.68	125.49	122.92	125.34	125.64	129.00	126.16	126.18
126.00	125.24	120.45	119.30	124.00	128.75	124.53	127.10	128.36
119.75	125.70	125.25	124.80	126.34	116.70	123.18	122.40	128.00
124.17	122.65	126.03	117.45	127.67	125.00	130.60	121.55	126.23
122.48	118.50	120.48	126.50	128.24	124.73	119.35	124.35	124.32
124.65	124.50	126.67	124.55	124.75	119.70	128.92	126.00	125.16
123.50	111.54	130.42	142.20	108.24	112.75	132.40	125.25	110.20
140.00	127.78	138.25	131.26	128.25	129.22	124.50	132.00	128.50
131.05	124.50	128.75	136.50	140.00	136.18	120.40	115.50	123.40
125.65	123.08	121.42	126.00	126.00	128.83	130.40	132.50	122.80

Construya una distribución de frecuencias que contenga 10 intervalos de clase de igual longitud.

- Trace el histograma de frecuencias.
- Trace el polígono de frecuencias.

## Ejercicios complementarios con grado de dificultad tres

- 1.38 Pruebe que la media armónica aplicada a los problemas de desplazamiento con la misma distancia siempre dará el resultado más preciso.
- 1.39 Pruebe que el coeficiente de correlación siempre está entre  $[-1, 1]$ .
- 1.40 Pruebe que la media aritmética siempre es mayor a la media geométrica para la misma muestra de datos.
- 1.41 Pruebe que en una muestra unimodal sesgada debe cumplirse  $\bar{x} - M = 3(\bar{x} - \tilde{x})$  o  $Media - Moda = 3(Media - Mediana)$ .
- 1.42 Sea una muestra donde todos sus valores son positivos. Demuestre que en este caso se cumple  $MA \leq MG \leq \bar{x}$  y la igualdad se cumple cuando todos los valores de la muestra son iguales.
- 1.43 Pruebe que en el caso de una variable con distribución normal se cumple  $\mu_4 = 3\sigma^4$ , donde  $\mu_4$  es el cuarto momento central.

## Proyectos de la unidad 1

En equipo resuelvan los siguientes proyectos. Pueden utilizar una hoja de cálculo para su solución.

- I.** En la hoja “IPC-42 Emp” del archivo “Datos IPC Divisas.xlsx” que se encuentra en el portal del libro de SALI, se encuentra una base de datos tomada de una página de internet (<http://economia.terra.com.mx/mercados/acciones/cambios.aspx?idtel=IB032MEXBOL>), de todos los IPC de 41 empresas que cotizan en México, la base de datos está a partir de febrero de 2013. Con esta información realice los siguientes proyectos.
- Con los valores del IPC de la empresa Kimberly® lleve a cabo un muestreo aleatorio simple. Con la muestra seleccionada calcule la media, mediana y desviación estándar muestral.
  - Con los valores del IPC de la empresa refresquera CC® lleve a cabo un muestreo aleatorio simple. Con la muestra seleccionada calcule la media, mediana y desviación estándar muestral.
  - Calcule los rendimientos y el riesgo de inversión de las empresas Pinfra®, Bimbo®, Genomma®, Televisa®, América Móvil® y TV Azteca®. Con los resultados obtenidos lleve a cabo un estudio comparativo de las empresas. Considere un criterio de orden y establezca las prioridades en que invertiría su capital. Explique estadísticamente el criterio que consideró en el ordenamiento.
  - Calcule los rendimientos y el riesgo de inversión de las empresas Comer®, Grupo Modelo®, Chedraui®, Soriana®, Wal-Mart® y Elektra®. Con los resultados obtenidos lleve a cabo un estudio comparativo de las empresas. Considere un criterio de orden y establezca las prioridades en que invertiría su capital. Explique estadísticamente el criterio que consideró en el ordenamiento.
  - Con las tres primeras empresas que eligió para invertir, Pinfra®, Bimbo®, Genomma®, Televisa®, América Móvil® y TV Azteca®, proponga cuatro portafolios y decida cuál es el mejor para invertir.
  - Con las tres primeras empresas que eligió para invertir, Comer®, Grupo Modelo®, Chedraui®, Soriana®, Wal-Mart® y Elektra®, proponga cuatro portafolios y decida cuál es el mejor para invertir.
- II.** En la hoja de divorcios por entidad del archivo Datos de divorcios.xlsx que se encuentra en la página del texto en SALI, hay una base de datos extraída del INEGI de todos los divorcios registrados en la República Mexicana para cada estado de 1985 a 2011. Con esta información:
- Realice un estudio comparativo entre las causas de divorcio: mutuo consentimiento, abandono de hogar sin causa justificada por más de dos o seis meses e incompatibilidad de caracteres. Establezca qué causa tiene mayor variabilidad y con base en qué medida tomó su decisión. Existe dependencia entre pares de causas.
  - Realice un estudio comparativo entre las causas de divorcio 5, 6 y 7. Establezca qué causa tiene mayor variabilidad y con base en qué medida tomó su decisión. Existe dependencia entre pares de causas.

Realice un estudio comparativo sobre la distribución de la cantidad de divorcios entre todos los estados de la República Mexicana y ordene por variabilidad de menor a mayor. Indique qué medida utilizó para tomar su decisión.

# Distribuciones muestrales y teorema del límite central

UNIDAD  
**2**



## Competencia específica a desarrollar

- Identificar e interpretar las diferentes distribuciones de muestreo para la estimación de los parámetros poblacionales.

## ¿Qué sabes?

- ¿Cómo identificas las diferentes distribuciones de muestreo?
- ¿Cómo determinas una distribución normal?
- ¿Cómo utilizas las tablas porcentuales?
- ¿Por qué es útil el teorema central del límite?

## Introducción

En el estudio de la estadística descriptiva (véase unidad 1) se describen y definen indicadores numéricos para una medida central y otra de variabilidad de un conjunto de observaciones, a los que se llamó *estadísticos*, como la media  $\bar{x}$  y la variancia  $s^2$  de la muestra. En todos los casos el conjunto de observaciones se consideró como un conjunto aislado. En la presente unidad vemos que cada uno de los valores de las observaciones en realidad corresponde a diferentes *variables aleatorias*. Por consiguiente, es de vital importancia que antes de iniciar la unidad el lector revise los conceptos de dependencia e independencia entre variables, la covarianza entre dos variables y el coeficiente de correlación.

Con base en lo expuesto, iniciamos esta unidad con el estudio de las distribuciones normal, *t*-Student, *ji* cuadrada y *F*, bases para el desarrollo de la estadística inferencial. Después, tratamos el problema del muestreo, tema que inició en la unidad 1, donde se hace énfasis en su relevancia para llevar a cabo inferencias estadísticas adecuadas. Podemos decir que, en apariencia, el tema es simple; sin embargo, en realidad, es más complicado de lo que parece. Además, tiene una gran importancia en el desarrollo de la estadística inferencial, puesto que un estudioso de la estadística y sus aplicaciones debe tener una buena técnica de muestreo para que las conjeturas que se realicen con respecto a toda la población sean lo más acertadas.

Por otro lado, contar con una buena técnica de muestreo no lo es todo, se requiere determinar con precisión los caracteres que se han de estudiar ya que, si éstos no se eligen de manera adecuada, las inferencias que se obtengan no serán correctas, aun cuando se realicen muestreos adecuados. Por ejemplo, en los primeros resultados sobre el crecimiento de la población, los cambios en el número de habitantes se predecían al calcular la diferencia entre el número de nacimientos y el de fallecimientos en un lapso determinado. Después, los expertos en estudios de la población comprobaron que la tasa de crecimiento depende solo del número de nacimientos, sin que el número de defunciones tenga importancia. Así, el futuro crecimiento de la población empezó a calcularse con base en el número anual de nacimientos por cada 1 000 habitantes. Sin embargo, pronto se dieron cuenta de que las predicciones obtenidas que utilizaban este método no daban resultados correctos. Los estadísticos comprobaron que hay otros factores que limitan el crecimiento de la población. Dado que el número de posibles nacimientos depende del número de mujeres y no solo del total de la población, además como las mujeres solo tienen hijos durante una parte de su vida, concluyeron que en la predicción del crecimiento de la población debe utilizarse el número de niños nacidos vivos por cada 1 000 mujeres en edad de procrear. Después, observaron que el valor obtenido al emplear este dato mejora al combinarlo con el dato del porcentaje de mujeres sin descendencia. Por tanto, puede decirse que la diferencia entre nacimientos y fallecimientos solo es útil para indicar el crecimiento de una población en un determinado periodo del pasado. Así, el número de nacimientos por cada 1 000 habitantes solo expresa la tasa de crecimiento en el mismo periodo; de manera similar el número de nacimientos por cada 1 000 mujeres en edad de procrear solo sirve para predecir el número de habitantes en el futuro.

A lo largo de esta unidad veremos con detalle el concepto de muestra aleatoria y las implicaciones teóricas que éste tiene. También estudiamos los estadísticos más comunes para la media, diferencia de medias, varianza y proporciones. Debido a su importancia, también analizamos por separado las distribuciones muestrales de la suma y media para la distribución normal y la Bernoulli.

En las últimas secciones de la unidad revisamos un resultado más general que los vistos hasta este momento, el cual se aplica a la suma y promedio de las variables aleatorias de una muestra *grande*. Es decir, realizamos un análisis detallado de uno de los teoremas de mayor trascendencia en la aplicación de la estadística, el teorema central del límite, el cual justifica que cualquier muestra aleatoria de tamaño  $n$  (grande) tomada de cualquier población con media  $\mu$  y varianza  $\sigma^2$  finitas, con estadístico media  $\bar{X}$  o el estadístico suma tendrá una distribución aproximadamente normal con media  $\mu$  y varianza  $\sigma^2/n$ .

## 2.1 Modelo normal

En esta sección veremos uno de los modelos continuos con mayor aplicación en la probabilidad y la estadística, el *modelo normal*. Esta distribución fue descubierta por Carl Friedrich Gauss, por lo que en algunos trabajos se le

conoce como ley de probabilidad de Gauss, según la cual una magnitud sufre la influencia de numerosas causas de variación, todas muy pequeñas e independientes entre sí, de manera que los resultados se acumulan alrededor de la media, se distribuyen de forma simétrica a su alrededor con una frecuencia que disminuye con rapidez al alejarse del centro. Por tanto, la curva que asemeja este comportamiento tiene forma de campana, misma que constituye la representación gráfica de una distribución de esta clase de distribuciones (véanse figuras 2.1, 2.2 y 2.3).

Entonces, en un modelo normal ¿cómo es la distribución de la variable alrededor de su media?

Sea  $X$  una variable aleatoria continua. Se dice que  $X$  tiene una **distribución normal** o de Gauss, con parámetros  $\mu$  y  $\sigma$  (positivo) en todos los reales cuando su función de densidad de probabilidad (FDP) es:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \text{ en } x \in (-\infty, \infty)$$

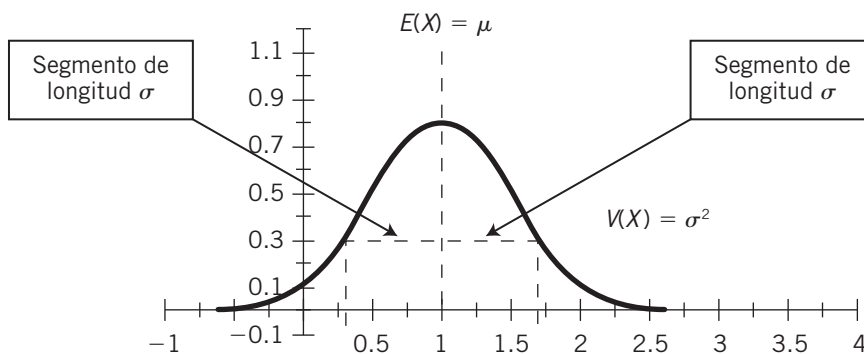
En cuestión de notación, tenemos que la clase de variables aleatorias con distribución normal y parámetros  $\mu, \sigma^2$  suele denotarse por  $N(\mu, \sigma^2)$ .

Como se mencionó, los modelos con distribución normal se caracterizan por la forma de *la gráfica de su función de densidad*. La gráfica de la distribución normal tiene forma de campana, como la que se observa en la figura 2.1.

Carl Friedrich Gauss nació en Brunswick, en 1777, y murió en Gotinga, en 1855. Matemático, astrónomo y físico alemán, autor de una gran cantidad de trabajos acerca de mecánica celeste, geodesia, magnetismo, electromagnetismo y óptica. Su concepción moderna de la naturaleza abstracta de las matemáticas le permitió ampliar el campo de los números. Fue el primero en descubrir la geometría hiperbólica no euclidiana.

La comprobación de que la función anterior es, en efecto, una función de densidad de probabilidad no es tan sencilla, se requiere del cálculo de variables complejas, en especial de una función que tiene gran auge: la *función gamma*, la cual se define como:

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx$$



**Figura 2.1** Gráfica de la FDP (función de densidad de probabilidad) de una variable aleatoria continua  $X$ , con distribución normal, media  $\mu$  y varianza  $\sigma^2$ .

En la figura 2.1 se aprecia que la recta  $x = \mu$  es el eje de simetría de la función, mientras que en los valores  $x = \mu - \sigma$  y  $x = \mu + \sigma$  se tienen los puntos de inflexión de la gráfica de la función. ¡Compruebe esto último mediante cálculo!

El *modelo normal* tiene gran aplicación en diferentes áreas y es una de las distribuciones con mayor auge en el estudio de la probabilidad y la estadística; la dimensión de su importancia radica en el *teorema del límite central*, que se estudia más adelante.

### Teorema 2.1

Si  $X$  es una variable aleatoria continua distribuida de manera normal en  $(-\infty, \infty)$  y  $f(x)$  es su función de densidad de probabilidad, entonces:

a)  $E(X) = \mu$

b)  $V(X) = \sigma^2$

## Cálculo de probabilidades

Como se recordará, la distribución de Gauss tiene gran importancia en el estudio de las probabilidades y la estadística; por consiguiente, es fundamental hacer un análisis detallado sobre su comportamiento para el cálculo de probabilidades. De los cursos de cálculo, se sabe que la integral de la función:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

no se puede resolver con funciones elementales; por tanto, cuando es definida solo podemos aproximar sus valores mediante alguno de los métodos numéricos, entre los que sobresalen: *método del trapecio*, *Simpson 1/3* y *cuadratura de Gauss*.

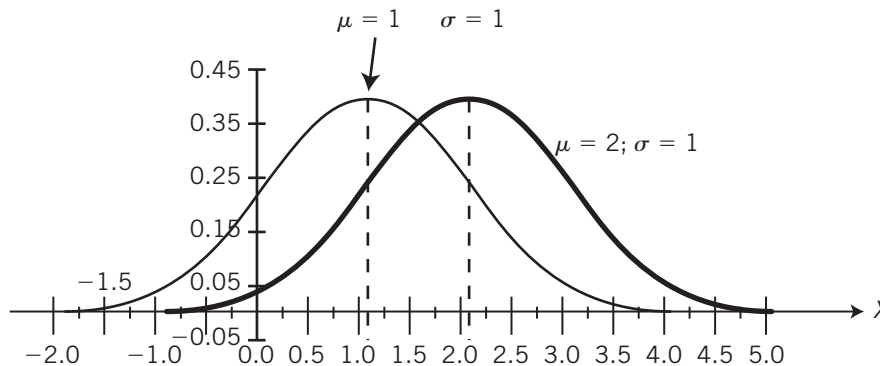
El problema de la estandarización se resuelve con el cambio de variable aleatoria:

$$Z = \frac{X - \mu}{\sigma}$$

que se conoce como la estandarización de la variable  $X$  a unidades en  $Z$ .

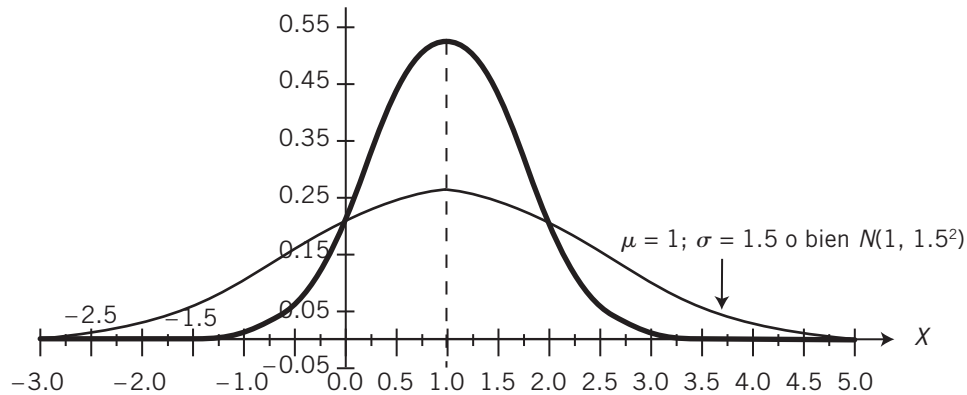
De lo anterior, se puede concluir que el cálculo de probabilidades resulta en extremo engorroso para este tipo de distribuciones, pero debido a su importancia se cuenta con tablas y programas para calcular las probabilidades. Desde luego, como es de suponerse se requiere de algún método con el que no se tenga la necesidad de resolver integrales para diferentes valores de  $\mu$  y  $\sigma$ . La solución a este problema se le conoce como la estandarización de la variable normal.

La fórmula en  $Z$  es una *regla de transformación*, puesto que en la estandarización  $X - \mu$ , representa un desplazamiento del eje de las ordenadas (véase figura 2.2). Mientras que la división entre la desviación estándar influye en la amplitud de la función (véase figura 2.3).



**Figura 2.2** Gráficas de la distribución normal con la misma desviación estándar, pero diferente valor esperado.

En las gráficas de la figura 2.2 se aprecia que ambas son iguales y solo cambia la posición del eje de las ordenadas; mientras que en las gráficas de la figura 2.3 cambia su amplitud; a mayor variancia menor amplitud.

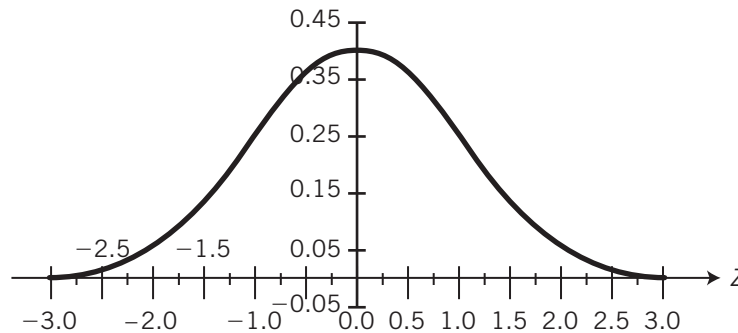


**Figura 2.3** Gráficas de la distribución normal con el mismo valor esperado, pero diferente desviación estándar.

Cuando se realiza la estandarización resulta que:

$$E(Z) = 0 \text{ y } V(Z) = 1, \text{ o bien } N(0, 1).$$

La gráfica se representa en la figura 2.4.



**Figura 2.4** Gráfica de la distribución normal estándar.

La integral para la función acumulada de la variable aleatoria  $Z$ , es decir la distribución normal en su forma estándar se calcula y representa por:

$$F(z_0) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_0} e^{-\frac{z^2}{2}} dz = \Phi(z_0)$$

Para el cálculo de probabilidades se emplean las propiedades siguientes de la distribución y la tabla de la normal estándar.

## Propiedades de la distribución normal estándar

Debido a que la distribución normal estándar juega un papel muy importante en la probabilidad y la estadística, el cálculo de sus probabilidades no es tan simple, se requiere el uso de tablas (como se verá más adelante) y algunas propiedades de la distribución para poder efectuar los cálculos.

- a) **Propiedad de simetría.** La función  $f(z)$  es simétrica con respecto al eje de las ordenadas. Es decir,  $P(Z < -Z_0) = P(Z > Z_0)$ .



- b) **Propiedad del complemento.** En los casos de  $P(Z > Z_0)$  se puede emplear la simetría, inciso a), o el complemento. Es decir,  $P(Z > Z_0) = 1 - P(Z \leq Z_0)$ .
- c)  $P(-1 < Z < 1) = 0.6827$
- d)  $P(-2 < Z < 2) = 0.9545$
- e) La suma de probabilidades fuera del intervalo  $(-4, 4)$ , no puede ser mayor a 0.0001, es decir, valen cero.

## Uso de tablas de la función acumulada

Como se mencionó antes, el uso de tablas o de algún programa para el cálculo de probabilidades es fundamental en la solución de los ejercicios. Por tanto, para homogeneizar el uso de las tablas que empleamos en esta unidad, se muestra con base en las tablas que se incluyen en el anexo de la página del libro en SALI y tienen la presentación que se observa en la figura 2.5 y la tabla 2.1.

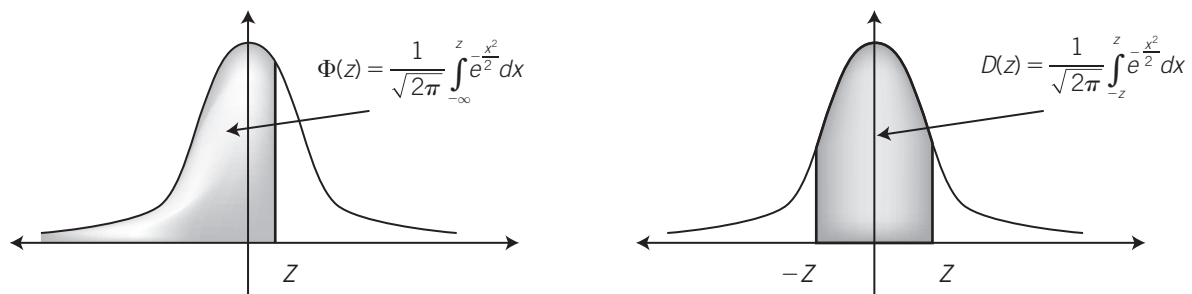


Figura 2.5 Función acumulada de la distribución normal estándar.

Tabla 2.1

$z$	$\Phi(-z)$	$\Phi(z)$	$D(z)$	$z$	$\Phi(-z)$	$\Phi(z)$	$D(z)$
<b>0.04</b>	0.4840	0.5160	0.0319	<b>0.44</b>	0.3300	0.6700	0.3401
<b>0.05</b>	0.4801	0.5199	0.0399	<b>0.45</b>	0.3264	0.6736	0.3473
<b>0.06</b>	0.4761	0.5239	0.0478	<b>0.46</b>	0.3228	0.6772	0.3545
$z$	$\Phi(-z)$	$\Phi(z)$	$D(z)$	$z$	$\Phi(-z)$	$\Phi(z)$	$D(z)$
<b>0.84</b>	0.2005	0.7995	0.5991	<b>1.24</b>	0.1075	0.8925	0.7850
<b>0.85</b>	0.1977	0.8023	0.6047	<b>1.25</b>	0.1056	<b>0.8944</b>	0.7887
<b>0.86</b>	<b>0.1949</b>	0.8051	0.6102	<b>1.26</b>	0.1038	0.8962	0.7923

Como se puede observar en la tabla 2.1, la *función acumulada* se representa por medio de la función  $\Phi(z)$ . Por comodidad del cálculo de probabilidades en intervalos simétricos, en las tablas se tiene otra función:

$$D(z_0) = \frac{1}{\sqrt{2\pi}} \int_{-z_0}^{z_0} e^{-\frac{z^2}{2}} dz = \Phi(z_0) - \Phi(-z_0)$$

En las tablas, los bloques están divididos en cuatro columnas:

- La *primera* corresponde a valores de  $Z$  que varían de centésima en centésima, desde 0 hasta 3.59.
- La *segunda* contiene los valores de la función acumulada hasta valores negativos de  $Z$ .
- La *tercera* tiene los valores de la función acumulada hasta valores positivos de  $Z$ .
- La *cuarta* presenta valores de probabilidades en intervalos simétricos con extremos  $-Z$  y  $Z$ .



Por tanto, el cálculo de probabilidades con base en estas funciones y las propiedades anteriores se puede efectuar de la siguiente forma:

1.  $P(Z < Z_0) = \Phi(Z_0)$
2.  $P(Z > Z_0) = P(Z < -Z_0) = \Phi(-Z_0)$
3.  $P(-Z_0 < Z < Z_0) = D(Z_0)$
4.  $P(a < Z < b) = \Phi(b) - \Phi(a)$

En los siguientes ejemplos empleamos ambas funciones:  $\Phi(z)$  y  $D(z)$ .

### Ejemplos 2.1 Función acumulada

Sea  $Z$  una variable aleatoria continua con distribución normal estándar, calcule cada una de las probabilidades siguientes.

1.  $P(Z < 1.25) = \Phi(1.25) = 0.8944$

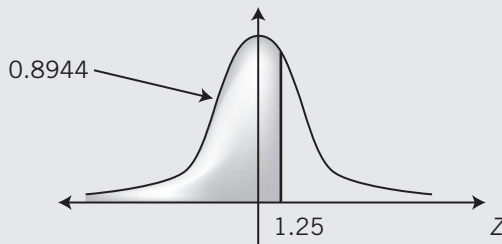


Figura 2.6

2.  $P(Z < -0.86) = \Phi(-0.86) = 0.1949$

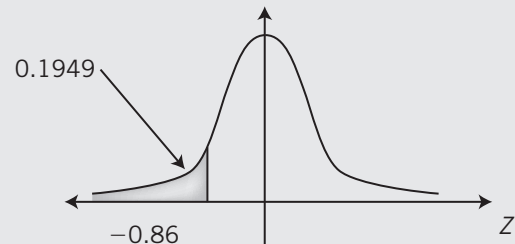


Figura 2.7

3.  $P(Z < -1.03) = 1 - P(Z \leq -1.03) = 1 - \Phi(-1.03)$   
 $= 1 - 0.1515 = 0.8485$ , o  
 $P(Z > -1.03) = P(Z < 1.03) = \Phi(1.03) = 0.8485$

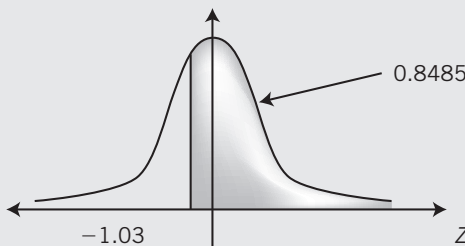


Figura 2.8

4.  $P(-2.97 < Z < 2.97) = D(2.97) = 0.9970$

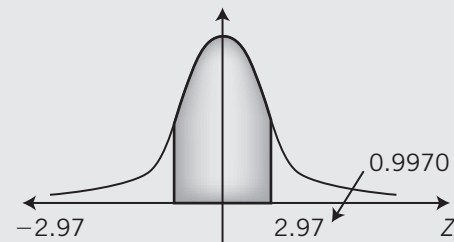


Figura 2.9

5.  $P(-0.57 < Z < 0.57) = D(0.57) = 0.4313$

6.  $P(-0.67 < Z < 1.24) = \Phi(1.24) - \Phi(-0.67) = 0.8925 - 0.2514 = 0.6411$

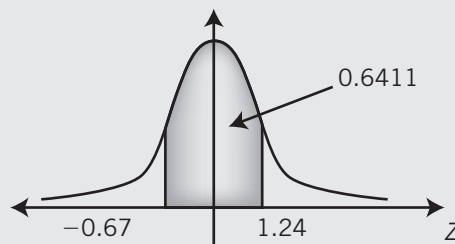


Figura 2.10

7.  $P(0.06 < Z < 3.04) = \Phi(3.04) - \Phi(0.06) = 0.9988 - 0.5239 = 0.4749$   
 8.  $P(Z < -4.5) = \Phi(-4.5) \approx 0$   
 9.  $P(Z < 5) = \Phi(5) \approx 1$   
 10.  $P(0.06 < Z < 5.1) = \Phi(5.1) - \Phi(0.06) \approx 1 - 0.5239 = 0.4761$

En los siguientes ejemplos,  $X$  es una variable aleatoria continua con distribución normal. Calcule las probabilidades indicadas.

11. Si  $E(X) = 4$  y  $V(X) = 9$ ; calcule la probabilidad  $P(X \geq 7)$

### Solución

En este caso, primero realizamos la estandarización de la variable  $X$  y después empleamos las tablas de la distribución normal estándar. Así:

$$\begin{aligned} P(X \geq 7) &= P\left(\frac{X - 4}{\sqrt{9}} \geq \frac{7 - 4}{3}\right) = P(Z \geq 1) \\ &= P(Z \leq -1) = \Phi(-1) = 0.1587 \end{aligned}$$

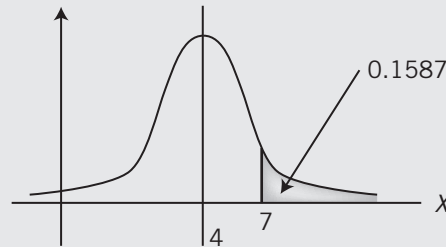


Figura 2.11

12. Si  $E(X) = 3$ , calcule  $P(-1 < X < 5)$ .

### Solución

En este ejemplo, primero realizamos la estandarización y después empleamos las tablas de la distribución normal estándar.

$$\begin{aligned} P(-1 < X < 5) &= P\left(\frac{-1 - 3}{2.5} < \frac{X - 3}{2.5} < \frac{5 - 3}{2.5}\right) = P(-1.6 < Z < 0.8) = \Phi(0.8) - \Phi(-1.6) \\ &= 0.7881 - 0.0548 = 0.7333. \end{aligned}$$

13. El peso de los estudiantes hombres del Instituto Tecnológico de Iguala se distribuye por lo regular con un valor promedio de 70.5 kg y una desviación estándar de 5.3. Si los estudiantes que pesan más de 85 kg son convocados para formar parte del equipo de futbol americano que representará a la escuela, determine el porcentaje de alumnos que podrán ser convocados.

### Solución

Sea la variable aleatoria  $X$ : *peso de los estudiantes hombres del Instituto Tecnológico de Iguala*:

$$P(X > 85) = P\left(\frac{X - 70.5}{5.3} > \frac{85 - 70.5}{5.3}\right) = P(Z > 2.74) = \Phi(-2.74) = 0.0031 = 0.31\%$$

14. Suponga que  $X$  representa la resistencia a la ruptura de una cuerda con un promedio de 100 y una desviación estándar de 4. Cada alambre para cuerda produce una utilidad de \$25, si  $X > 95$ . En caso contrario, la cuerda se tiene que utilizar con otro propósito y se obtiene una utilidad de \$10 por alambre. Encuentre la utilidad esperada por alambre.

**Solución**

Primero, calculamos las probabilidades:

$$P(X \leq 95) = P\left(\frac{X - 100}{4} \leq \frac{95 - 100}{4}\right) = P(Z \leq -1.25) = \Phi(-1.25) = 0.1056$$

$$P(X > 95) = 1 - P(X \leq 95) = 1 - 0.1056 = 0.8944$$

El valor esperado estará dado por:

$$\text{Ganancia esperada} = P(X > 95) \times 25 + P(X \leq 95) \times 10 = 0.8944 \times 25 + 0.1056 \times 10 = \$23.416$$

**Uso de tablas porcentuales**

Con frecuencia, al resolver problemas se deben hacer conclusiones con respecto a la variable aleatoria en estudio. Para tal efecto, es común tener que encontrar los valores de la variable con los cuales se obtienen las probabilidades establecidas (pueden estar dadas en porcentajes). Por otra parte, en lo que concierne a las variables aleatorias con distribución normal, se emplean las tablas porcentuales de la distribución normal, cuya presentación se muestra en la figura 2.12 y la tabla 2.2.

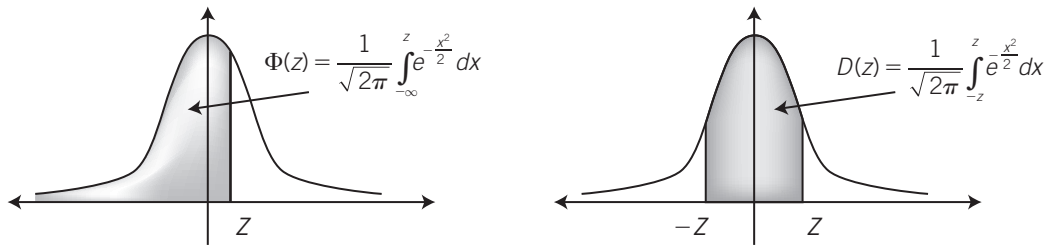


Figura 2.12

Tabla 2.2

%	$z(\Phi)$	$z(D)$	%	$z(\Phi)$	$z(D)$	%	$z(\Phi)$	$z(D)$
<b>0.6</b>	-2.512	0.008	<b>5.6</b>	-1.589	0.070	<b>10.6</b>	-1.248	0.133
<b>0.7</b>	-2.457	0.009	<b>5.7</b>	-1.580	0.071	<b>10.7</b>	-1.243	0.135
<b>0.8</b>	-2.409	0.010	<b>5.8</b>	-1.572	0.073	<b>10.8</b>	<b>-1.237</b>	0.136
<b>0.9</b>	-2.366	0.011	<b>5.9</b>	-1.563	0.074	<b>10.9</b>	-1.232	0.137
%	$z(\Phi)$	$z(D)$	%	$z(\Phi)$	$z(D)$	%	$z(\Phi)$	$z(D)$
<b>15.6</b>	-1.011	0.197	<b>20.6</b>	-0.820	0.261	<b>25.6</b>	-0.656	0.327
<b>15.7</b>	-1.007	0.198	<b>20.7</b>	-0.817	0.262	<b>25.7</b>	-0.653	0.328
<b>15.8</b>	-1.003	0.199	<b>20.8</b>	-0.813	0.264	<b>25.8</b>	-0.650	0.329
<b>15.9</b>	-0.999	0.201	<b>20.9</b>	-0.810	0.265	<b>25.9</b>	-0.646	0.331

En las tablas, los bloques están divididos en tres columnas.

- *La primera* corresponde a las probabilidades dadas en porcentajes y varía en décimas de porcentaje, desde 0.0 hasta 99.9.
- *La segunda* corresponde a los valores de Z, cuya función acumulada proporciona el porcentaje de la primera columna.

- La tercera corresponde a los valores de  $Z$ , con intervalos simétricos (extremos  $-Z$  y  $Z$ ), de manera que la probabilidad en este intervalo es igual al porcentaje de la primera columna.

Veamos los siguientes ejemplos sobre el uso de estas tablas.

### Ejemplos 2.2 Tablas porcentuales de la distribución normal

1. Encuentre el valor de  $z_0$ , tal  $P(Z < z_0) = 0.108$

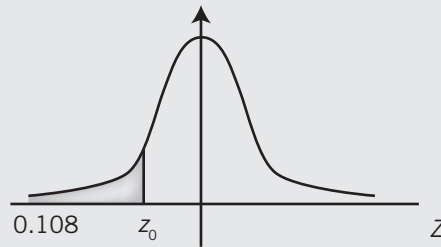


Figura 2.13

#### Solución

La probabilidad que se indica es igual a 10.8%; por tanto, al buscar en las tablas porcentuales 0.8%, tenemos:

$$z_0 = Z(\Phi) = -1.237; \text{ esto es: } P(Z < -1.237) = 0.108$$

2. Encuentre el valor  $z_0$ , tal que  $P(Z \geq z_0) = 5\%$

Como las tablas porcentuales nos muestran los valores para la función acumulada de menos infinito hasta el valor indicado, tenemos que emplear la propiedad del complemento. Es decir:

$$P(Z \geq z_0) = 1 - P(Z < z_0) = 5\%$$

De donde necesitamos  $P(Z < z_0) = 95\%$

$$z_0 = Z(\Phi) = 1.645$$

Esto es:

$$P(Z \geq 1.645) = 0.05$$

Este ejercicio también se puede resolver si se emplea la propiedad de simetría:

$$P(Z \geq z_0) = P(Z \leq -z_0) = 5\%$$

De donde:

$$-z_0 = -1.645$$

es decir:

$$z_0 = 1.645$$

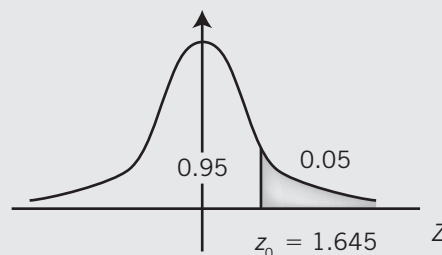


Figura 2.14

3. Si  $E(X) = 4$  y  $V(X) = 9$ ; calcule el valor  $x_0$ , tal que  $P(X \leq x_0) = 75\%$

**Solución**

En este ejemplo, primero realizamos la estandarización y después empleamos las tablas porcentuales de la distribución normal estándar.

$$P(X \leq x_0) = P\left(\frac{X - 4}{3} \leq \frac{x_0 - 4}{3}\right) = P(Z \leq z_0) = 75\%$$

De donde tenemos:

$$z_0 = 0.674$$

Por otro lado:

$$z_0 = \frac{x_0 - 4}{3}$$

Al despejar  $x_0$ , tenemos:

$$x_0 = 4 + 3z_0 = 4 + 3(0.674) = 6.022$$

$$P(X \leq 6.022) = P(Z \leq 0.674) = 75\%$$

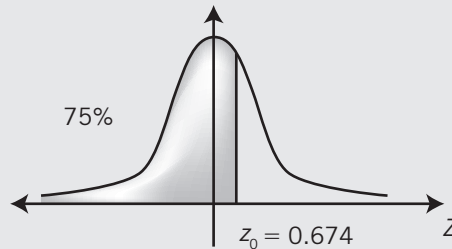


Figura 2.15

4. La variable aleatoria  $X$  representa la vida promedio de cierto aparato electrónico; tiene una distribución aproximada normal, con media  $\mu = 3.5$  años y desviación estándar  $\sigma = 1.5$  años. Si el fabricante solo desea reparar los aparatos que fabrica en su periodo de garantía, ¿cuál tendría que ser el periodo de garantía?

**Solución**

Como  $X$  representa a la vida promedio de los aparatos y 10% la probabilidad de que el aparato dure menos que el periodo establecido,  $x_0$ , tenemos:

$$P(X \leq x_0) = P\left(\frac{X - 3.5}{1.5} \leq \frac{x_0 - 3.5}{1.5}\right) = P(Z \leq z_0) = 0.10$$

De donde:

$$z_0 = \frac{x_0 - 3.5}{1.5}$$

Por tanto, al despejar  $x_0$ , resulta:

$$x_0 = 3.5 + 1.5z_0$$

De las tablas porcentuales de la distribución normal estándar resulta:

$$z_0 = -1.282$$

Por último, el periodo de garantía es:

$$x_0 = 3.5 + 1.5(-1.282) = 1.577 \text{ años}$$

## Ejercicios 2.1

---

1. La administración de una empresa maquiladora quiere calcular los costos de reparación anual de cierta máquina, para lo cual lleva a cabo un estudio en el que obtiene que los costos de reparación anual se comportan de forma normal con media \$400 000 y desviación estándar de \$50 000.
  - a) Calcule la probabilidad de que los costos de reparación para este año estén entre \$300 000 y \$500 000.
  - b) ¿Abajo de qué costo se encuentra el presupuesto para la reparación anual de las máquinas en 10% de los casos?
2. Una compañía de servicios paga a sus empleados un salario promedio de \$10 por hora con una desviación estándar de \$1. Si los salarios tienen una distribución normal.
  - a) ¿Qué porcentaje de los trabajadores recibe salario entre \$9 y \$11 por hora?
  - b) ¿Mayor de qué cantidad es 5% de los salarios más altos?
3. El peso de los estudiantes hombres del Instituto Tecnológico de Piedras Negras se distribuye normalmente con un valor promedio de 75.5 kg y una variancia de 24. Si los estudiantes que pesan más de 85 kg serán convocados para formar parte del equipo de futbol americano que representará a la escuela, determine el porcentaje de alumnos que podrán ser convocados.
4. Ciertos tipos de baterías para automóvil tienen un tiempo de vida normalmente distribuido con media 1 200 días y desviación estándar igual a 100 días.
  - a) ¿Cuántas de las 3 000 baterías que se venderán durarán más de 1 300 días?
  - b) ¿Por cuánto tiempo se deben garantizar las baterías si el fabricante quiere reemplazar solo 10% de las baterías vendidas?
5. Si la calificación promedio de un grupo es de 6.43, con una desviación estándar de 1.91, y se supone que la distribución de las calificaciones es normal, calcule la probabilidad de que en el siguiente examen un alumno pase. (Nota: la calificación mínima aprobatoria es 6.)
6. El diámetro de los pernos de una fábrica tiene una distribución normal con una media de 950 milímetros y una desviación estándar de 10 mm.
  - a) ¿Cuál es la probabilidad de que un perno escogido al azar tenga un diámetro entre 947 y 958 mm?
  - b) ¿Cuánto debe valer  $c$  para que un perno escogido al azar tenga un diámetro menor que  $c$  con una probabilidad de 0.90?
7. Se supone que los resultados de un examen tienen una distribución normal con una media de 78 y una variancia de 36.
  - a) ¿Cuál es la probabilidad de que una persona que presenta un examen obtenga una calificación mayor a 72?
  - b) ¿Cuál debe ser la calificación mínima aprobatoria si el examinador pretende que solo 28% de los estudiantes apruebe?
8. En un aserradero se cortan árboles en trozos de 4 m de longitud en promedio, con desviación estándar de 0.23 m, y las longitudes se distribuyen en forma aproximadamente normal.
  - a) Si se elige un lote de 500 trozos, ¿cuántos de estos pernos se espera que tengan una longitud mayor a 4.12 m?
  - b) Si se eligen nueve trozos, ¿cuál es la probabilidad de que cuatro tengan una longitud mayor de 4.05 m?
9. Los pesos de un número grande de perros miniatura de lana están distribuidos aproximadamente en forma normal con una media de 8 kg y una desviación estándar de 0.9 kg. Encuentre la fracción de los perros de lana con pesos:
  - a) Arriba de 9.5 kg.

- b) Menos de 8.6 kg.  
c) Entre 7.3 y 9.1 kg.
10. Una fábrica produce pistones cuyos diámetros se encuentran distribuidos en forma normal con un diámetro promedio de 5 cm y una desviación estándar de 0.001 cm. Para que un pistón sea útil su diámetro debe encontrarse entre 4.998 y 5.002 cm. Si el diámetro del pistón es menor de 4.998, se desecha; y si es mayor de 5.002 se puede reprocesar. Si en la fábrica se producen cada mes 20 000 pistones:
- a) ¿Cuántos serán útiles?  
b) ¿Cuántos serán desechados?  
c) ¿Cuántos necesitan ser reprocesados?
11. El tiempo necesario para armar cierta unidad es una variable aleatoria distribuida normalmente con  $\mu = 30$  minutos,  $\sigma = 2$  minutos. Determine el tiempo de armado de manera que la probabilidad de excederlo sea de 0.02.

## 2.2 Distribución ji cuadrada

La función de densidad de una variable aleatoria continua con distribución ji cuadrada y parámetro  $\nu$  está definida por:

$$f(x; \nu) = \begin{cases} \frac{1}{2^{\frac{\nu}{2}} \Gamma\left(\frac{\nu}{2}\right)} \left[ x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}} \right] & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Al parámetro  $\nu$  se le conoce como *grados de libertad*, y está muy relacionado con el tamaño de la muestra. Esta distribución se representa por  $\chi^2$  (ji cuadrada). En la figura 2.16 se muestran algunas gráficas de la ji cuadrada para diferentes valores del parámetro  $\nu$ .

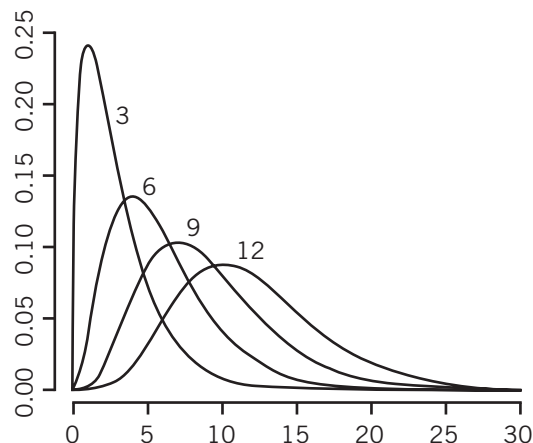


Figura 2.16 Función de densidad de la distribución  $\chi^2$  con 3, 6, 9 y 12 grados de libertad.

### Uso de tablas de la distribución ji cuadrada

Las tablas de la distribución ji cuadrada que se usan en el texto sirven para calcular los valores de la distribución para ciertas probabilidades, se muestran en la página del libro en SALI y tienen la presentación que se muestra en la figura 2.17 y en la tabla 2.3.

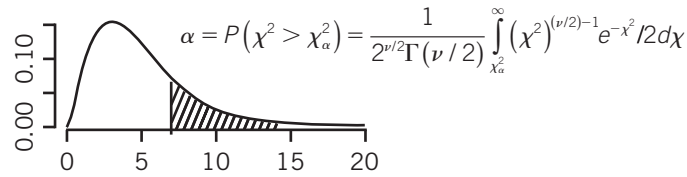


Figura 2.17

Tabla 2.3

$\gamma\lambda = \nu$	Valores de $\alpha$							
	0.005	0.995	0.010	0.990	0.015	0.985	0.020	0.980
<b>8</b>	21.9549	1.3444	20.0902	1.6465	18.9738	1.8603	18.1682	2.0325
<b>9</b>	23.5893	1.7349	21.6660	2.0879	20.5125	2.3348	19.6790	2.5324
<b>10</b>	25.1881	2.1558	23.2093	2.5582	22.0206	2.8372	21.1608	3.0591

Como se puede observar, la tabla 2.3 muestra los valores de la distribución *ji* cuadrada con los cuales el área derecha bajo la curva es igual  $\alpha$ . Esto es, si la variable aleatoria  $X$  tiene distribución *ji* cuadrada con  $\nu$  grados de libertad, entonces:

$$P(X_\nu > k) = \alpha = 1 - P(X_\nu \leq k) = 1 - F_{X_\nu}(k)$$

denota la probabilidad de que  $X$  (con  $\nu$  grados de libertad) sea mayor al valor  $k$  y  $F$  su función de distribución acumulada.

Las tablas están conformadas de la siguiente manera. En la *primera columna* se muestran los grados de libertad de la variable. Después, se forman parejas de columnas que en la parte de arriba muestran el valor de la probabilidad. Por comodidad, para estudios posteriores sobre intervalos de confianza y pruebas de hipótesis, se localizan los valores de la probabilidad y su complemento juntos.

### Ejemplos 2.3 Distribución *ji* cuadrada

Encuentre el valor correspondiente a la probabilidad indicada de una distribución *ji* cuadrada.

1.  $P(X_8 > k) = 0.99$

#### Solución

La probabilidad es de cola derecha, por tanto, se busca en las tablas el valor  $\alpha = 0.99$ . Luego, en la columna de grados de libertad, se localiza el 8 y el cruce de fila y columna es el valor buscado (véase tabla 2.4).

Tabla 2.4

$\gamma\lambda = \nu$	Valores de $\alpha$							
	0.005	0.995	0.010	0.990	0.015	0.985	0.020	0.980
<b>8</b>	21.9549	1.3444	20.0902	<b>1.6465</b>	18.9738	1.8603	18.1682	2.0325
<b>9</b>	23.5893	1.7349	21.6660	2.0879	20.5125	2.3348	19.6790	2.5324
<b>10</b>	25.1881	2.1558	23.2093	2.5582	22.0206	2.8372	21.1608	3.0591

Así,  $k = 1.6465$ , es decir,  $P(X_8 > 1.6465) = 0.99$

2.  $P(X_9 < k) = 0.98$



**Solución**

La probabilidad es de cola izquierda y las tablas dan las probabilidades de cola derecha, por lo que se utiliza el complemento. Luego:

$$P(X_9 < k) = 1 - P(X_9 \geq k) = 0.98 \text{ esto es } P(X_9 \geq k) = 0.02$$

Por último, buscamos en las tablas el valor de  $\alpha = 0.02$  y en la columna de grados de libertad localizamos el valor 9; el cruce entre la fila y la columna es el valor buscado (véase tabla 2.5).

**Tabla 2.5**

$\gamma\lambda = \nu$	Valores de $\alpha$							
	0.005	0.995	0.010	0.990	0.015	0.985	0.020	0.980
<b>8</b>	21.9549	1.3444	20.0902	1.6465	18.9738	1.8603	18.1682	2.0325
<b>9</b>	23.5893	1.7349	21.6660	2.0879	20.5125	2.3348	<b>19.6790</b>	2.5324
<b>10</b>	25.1881	2.1558	23.2093	2.5582	22.0206	2.8372	21.1608	3.0591

Así,  $k = 19.6790$ ; es decir,  $P(X_9 < 19.6790) = 0.98$  o  $P(X_9 \geq 19.6790) = 0.02$

**Ejercicios 2.2**

1. Sea  $X$  una variable aleatoria con distribución  $ji$  cuadrada con ocho grados de libertad, calcule el valor de  $k$ , tal que  $P(X > k) = 0.02$ .
2. Sea  $X$  una variable aleatoria con distribución  $ji$  cuadrada con 10 grados de libertad, calcule el valor de  $k$ , tal que  $P(X < k) = 0.10$ .
3. Calcule la mediana de una variable aleatoria  $ji$  cuadrada con dos grados de libertad. *Sugerencia:* La mediana es el valor medio, es decir, el valor  $m$  en el que  $P(X > m) = P(X < m)$ .
4. Calcule el valor esperado y la varianza para una variable aleatoria con distribución  $ji$  cuadrada.
5. Sea  $X$  una variable aleatoria con distribución  $ji$  cuadrada y  $\nu = 4$ , encuentre una expresión para la función de distribución acumulada.

**2.3 Distribución t-Student**

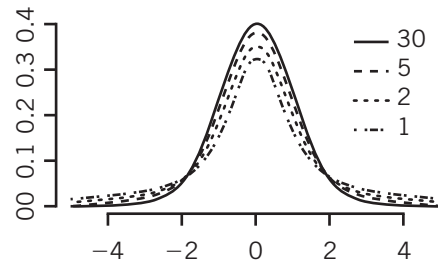
Una variable aleatoria continua  $X$  tiene una distribución de probabilidad t-Student, también se puede nombrar como  $t$  de Student, si su distribución se asemeja a la de un modelo normal. De hecho, la distribución t-Student, al igual que la distribución normal, es simétrica y tiene forma de campana. La diferencia entre la distribución normal y la t-Student reside en que esta última a menos grados de libertad tiene colas más pesadas que la normal. Es decir, las probabilidades en las colas son más pesadas que la normal, por consiguiente, a menos grados de libertad la distribución t-Student es más chata que la normal. Con respecto a los grados de libertad, la distribución t-Student coincide de manera asintótica con la distribución normal. En las unidades 3, 4 y 6 se verá que la distribución  $t$  varía más, debido a que sus fluctuaciones dependen tanto del valor esperado, como de la variancia muestrales.

Esta distribución de probabilidad se publicó por primera vez en 1908, por el irlandés W. S. Gosset. En esa época Gosset trabajaba en una cervecería irlandesa que desaprobaba la publicación de trabajos de investigación, por lo que Gosset publicó su trabajo con el seudónimo "Student". Por este motivo, a esta distribución se le asignó el nombre de t-Student.

La función de densidad de una variable aleatoria continua con distribución t-Student y parámetro  $\nu$  está definida por:

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

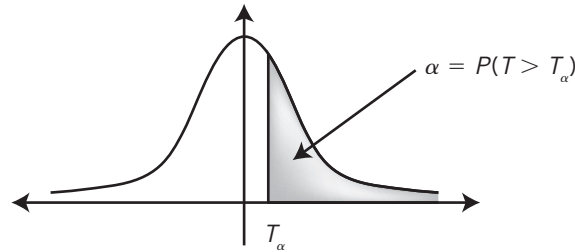
Al parámetro  $\nu$  se le conoce como *grados de libertad* y está muy relacionado con el tamaño de las muestras. Como vemos en la figura 2.18, mientras mayor sea el número de grados de libertad, más se asemejará la distribución t-Student a la normal.



**Figura 2.18** La función de densidad de los modelos t-Student, para 1, 2, 5 y 30 grados de libertad y la distribución normal.

## Uso de tablas de la distribución t-Student

Las tablas de la distribución t-Student que se usan en el texto sirven para calcular los valores de la variable para ciertas probabilidades; éstas se encuentran en la página del libro en SALI y tienen la presentación que se muestra en la figura 2.19 y en la tabla 2.6.



**Figura 2.19**

**Tabla 2.6**

$\gamma\lambda = \nu$	0.0035	0.0040	0.0045	0.0050	0.0055	0.0060	0.0065	0.0070	0.0075
<b>11</b>	3.306	3.231	3.165	3.106	3.052	3.004	2.959	2.917	2.879
<b>12</b>	3.247	3.175	3.111	3.055	3.003	2.956	2.913	2.873	2.836
<b>13</b>	3.198	3.128	3.067	3.012	2.963	2.917	2.876	2.837	2.801
<b>14</b>	3.157	3.089	3.030	2.977	2.929	2.885	2.844	2.807	2.771
<b>15</b>	3.122	3.056	2.998	2.947	2.900	2.857	2.817	2.780	2.746

Como se puede observar, las tablas muestran los valores de la distribución t-Student para diferentes probabilidades, las cuales se denotan por  $\alpha$ . Esto es, si la variable aleatoria  $X$  tiene distribución t-Student con  $\nu$  grados de libertad, entonces  $P(X_\nu > k) = \alpha = 1 - P(X_\nu \leq k) = 1 - F_{X_\nu}(k)$ , denota la probabilidad de que  $X$  (con  $\nu$  grados de libertad) sea mayor al valor  $k$ .

Las tablas están conformadas de la siguiente manera. En la *primera columna* se muestran los grados de libertad,  $\nu$ , de la variable y en la *primera fila* los valores de las probabilidades  $\alpha$ , mientras que en los cruces de grados de libertad y probabilidades se muestran los valores de la variable  $k$  que cumplen:

$$P(X_\nu > k) = \alpha \text{ o } F_{X_\nu}(k) = 1 - \alpha$$

### Ejemplos 2.4 Distribución t-Student

Encuentre el valor correspondiente de la probabilidad indicada para una distribución t-Student.

1.  $P(X_{12} > k) = 0.004$

#### Solución

Como la probabilidad es de cola derecha, en las tablas se busca el valor de  $\alpha = 0.004$ . Luego, en la columna de grados de libertad se localiza el 12 y el cruce de fila y columna es el valor buscado (véase tabla 2.7).

Tabla 2.7

$\gamma\lambda = \nu$	0.0035	0.0040	0.0045	0.0050	0.0055	0.0060	0.0065	0.0070	0.0075
11	3.306	3.231	3.165	3.106	3.052	3.004	2.959	2.917	2.879
12	3.247	<b>3.175</b>	3.111	3.055	3.003	2.956	2.913	2.873	2.836
13	3.198	3.128	3.067	3.012	2.963	2.917	2.876	2.837	2.801
14	3.157	3.089	3.030	2.977	2.929	2.885	2.844	2.807	2.771
15	3.122	3.056	2.998	2.947	2.900	2.857	2.817	2.780	2.746

Así,  $k = 3.175$ , es decir,  $P(X_{12} > 3.175) = 0.004$

2.  $P(X_{14} < k) = 0.007$

#### Solución

Como la probabilidad es de cola izquierda y las tablas dan las probabilidades de cola derecha, se utiliza la simetría. De tal forma que:

$$P(X_{14} < k) = P(X_{14} \geq -k) = 0.007$$

Ahora, se busca en las tablas el valor  $\alpha = 0.007$ ; luego se localiza el 14 en la columna de grados de libertad y el cruce de fila y columna es el valor deseado (véase tabla 2.8).

Tabla 2.8

$\gamma\lambda = \nu$	0.0035	0.0040	0.0045	0.0050	0.0055	0.0060	0.0065	0.0070	0.0075
11	3.306	3.231	3.165	3.106	3.052	3.004	2.959	2.917	2.879
12	3.247	3.175	3.111	3.055	3.003	2.956	2.913	2.873	2.836
13	3.198	3.128	3.067	3.012	2.963	2.917	2.876	2.837	2.801
14	3.157	3.089	3.030	2.977	2.929	2.885	2.844	<b>2.807</b>	2.771
15	3.122	3.056	2.998	2.947	2.900	2.857	2.817	2.780	2.746

Así,  $-k = 2.807$ , luego  $k = -2.807$ . Es decir,  $P(X_{14} < -2.807) = 0.007$

## Ejercicios 2.3

1. Sea  $X$  una variable aleatoria con distribución t-Student con 15 grados de libertad, calcule  $t_\alpha$ , tal que  $P(T < t_\alpha) = 0.95$ .

- Sea  $X$  una variable aleatoria con distribución t-Student con 19 grados de libertad, calcule  $t_\alpha$ , tal que  $P(T > t_\alpha) = 0.90$ .
- Con las tablas de la t-Student determine el valor de  $\int_2^\infty \frac{dx}{(12 + x^2)^2}$
- Sea  $T$  una variable aleatoria con distribución t-Student y  $\nu = 1$ , encuentre una expresión para la función de distribución acumulada.
- Sea  $T$  una variable aleatoria con distribución t-Student y  $\nu = 3$ , encuentre una expresión para la función de distribución acumulada.

## 2.4 Distribución $F$

Una variable aleatoria continua  $X$  tiene una distribución de probabilidad  $F$ , cuando su función de densidad se puede representar como se muestra en la definición.

La función de densidad de una variable aleatoria continua con distribución  $F$  y parámetros  $\nu_1$  y  $\nu_2$  está definida por:

$$f(x) = \frac{\Gamma\left(\frac{\nu_1 + \nu_2}{2}\right) \left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}}}{\Gamma\left(\frac{\nu_1}{2}\right) \Gamma\left(\frac{\nu_2}{2}\right)} \cdot x^{\frac{\nu_1}{2} - 1} \left[1 + \frac{\nu_1}{\nu_2} x\right]^{-\frac{\nu_1 + \nu_2}{2}}, \quad x > 0$$

A los parámetros  $\nu_1$  y  $\nu_2$  se les conoce como *grados de libertad* de la distribución  $F$  del numerador y denominador, respectivamente. Este tipo de distribuciones se emplea con mucha frecuencia en estadística, cuando se trabaja con la razón entre variancias.

### Uso de tablas de la distribución $F$

Las tablas de la distribución  $F$  que se usan en el texto sirven para calcular los valores de la distribución para ciertas probabilidades; se pueden encontrar en el portal del libro en SALI y tienen la presentación que se observa en la figura 2.20 y la tabla 2.9.

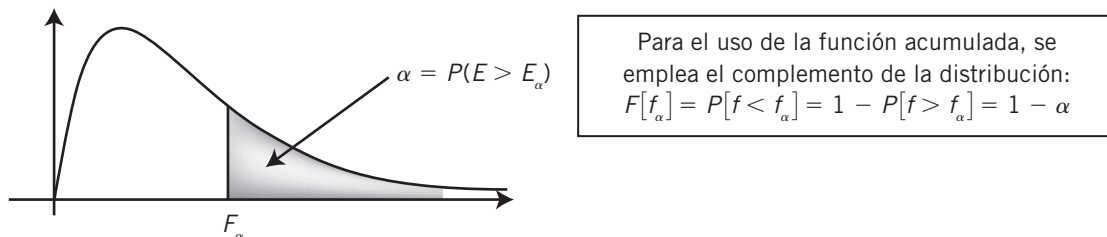


Figura 2.20

Tabla 2.9 Grados de libertad del numerador  $\nu_1$  (valor de  $\alpha = 0.005$ )

$gl = \nu_2$	Grados de libertad del numerador $\nu_1$ (valor de $\alpha = 0.005$ )									
	1	2	3	4	5	6	7	8	9	10
6	18.635	14.544	12.917	12.028	11.464	11.073	10.786	10.566	10.391	10.250
7	16.235	12.404	10.883	10.050	9.522	9.155	8.885	8.678	8.514	8.380
8	14.688	11.043	9.597	8.805	8.302	7.952	7.694	7.496	7.339	7.211

Como se puede observar, la tabla muestra los valores de la distribución  $F$  para diferentes probabilidades a la derecha, las cuales se denotan por  $\alpha$ ; a diferencia de las otras dos distribuciones para cada valor de  $\alpha$ , se tienen dos páginas de valores de la variable según sean los grados de libertad del 1 al 30 y después para valores saltados. Esto es, si la variable aleatoria  $X$  tiene distribución  $F$  con  $\nu_1$  grados de libertad en el numerador y  $\nu_2$  grados de libertad en el denominador, entonces  $P(X(\nu_1, \nu_2) > k) = \alpha = 1 - P(X(\nu_1, \nu_2) \leq k) = 1 - F_{X(\nu_1, \nu_2)}(k)$ , denota la probabilidad de que  $X$  (con  $\nu_1$  y  $\nu_2$  grados de libertad en el numerador y denominador, respectivamente) sea mayor al valor  $k$ , con  $F$  la función de distribución acumulada de  $X$ .

Las tablas están formadas de la siguiente manera: en la primera columna se muestran los grados de libertad del denominador y en la primera fila, el valor de la probabilidad al que corresponden las tablas (los valores disponibles son: 0.005, 0.01, 0.02, 0.025, 0.05, 0.10), mientras que en la segunda fila se muestran los grados de libertad del numerador.

### Ejemplos 2.5 Distribución F

1. Encuentre el valor de  $k$ , para que  $P(X(5, 7) > k) = 0.005$ , donde  $X$  tiene una distribución  $F$ .

#### Solución

Como la probabilidad es de cola derecha, se busca en las tablas el valor de  $\alpha = 0.005$ ; luego, en la columna de grados de libertad del numerador se localiza el 5, después se encuentra el cruce con la fila correspondiente a los 7 grados de libertad del denominador, y el valor de la intersección es el valor buscado (véase tabla 2.10).

Tabla 2.10

$gl = \nu_2$	Grados de libertad del numerador $\nu_1$ (valor de $\alpha = 0.005$ )									
	1	2	3	4	5	6	7	8	9	10
6	18.635	14.544	12.917	12.028	11.464	11.073	10.786	10.566	10.391	10.250
7	16.235	12.404	10.883	10.050	<b>9.522</b>	9.155	8.885	8.678	8.514	8.380
8	14.688	11.043	9.597	8.805	8.302	7.952	7.694	7.496	7.339	7.211

Así,  $k = 9.522$ , es decir,  $P(X(5, 7) > 9.522) = 0.005$

Para calcular valores de la distribución acumulada, tales que sus complementos sean algunos de los valores de tablas (0.005, 0.01, 0.02, 0.025, 0.05, 0.10), se utiliza el siguiente resultado.

#### ◆ Proposición

Suponga que  $X(n, m)$  representa a una variable aleatoria con distribución  $F$ , con  $n$  y  $m$  grados de libertad del numerador y denominador, respectivamente, si:

$$\alpha = P(X(n, m) < f_\alpha(n, m)), \text{ entonces } f_{1-\alpha}(m, n) = \frac{1}{f_\alpha(n, m)}$$

En donde  $f_\alpha(n, m)$  representa el valor de la variable  $F$  con  $n$  y  $m$  grados de libertad en el numerador y denominador, respectivamente, cuya distribución acumulada tenga una probabilidad  $\alpha$ , de igual manera  $f_{1-\alpha}(m, n)$ .

2. Encuentre los valores de  $k$ , para que  $P(X(6, 7) < k) = 0.995$ , en donde  $X$  tiene una distribución  $F$ .

#### Solución

Como la probabilidad es de cola izquierda y su complemento  $1 - 0.995 = 0.005$  está en tablas, se puede utilizar el resultado anterior, con el que se obtiene:

$$P(X(6, 7) < k) = 0.995 \Rightarrow P\left(X(7, 6) > \frac{1}{k}\right) = 0.005$$

Ahora, en las tablas se busca el valor de  $\alpha = 0.005$ ; luego se localiza el 7 en la columna de grados de libertad del numerador, después se encuentra el cruce con la fila correspondiente a los 6 grados de libertad del denominador, y el valor de la intersección es el que se busca (véase tabla 2.11).

**Tabla 2.11** Grados de libertad del numerador  $\nu_1$  (valor de  $\alpha = 0.005$ )

$gl = \nu_2$	Grados de libertad del numerador $\nu_1$ (valor de $\alpha = 0.005$ )									
	1	2	3	4	5	6	7	8	9	10
6	18.635	14.544	12.917	12.028	11.464	11.073	<b>10.786</b>	10.566	10.391	10.250
7	16.235	12.404	10.883	10.050	9.522	9.155	8.885	8.678	8.514	8.380
8	14.688	11.043	9.597	8.805	8.302	7.952	7.694	7.496	7.339	7.211

Así,  $\frac{1}{k} = 10.786$ , luego  $k = \frac{1}{10.786} = 0.093$ . Es decir  $P(X(6, 7) < 0.093) = 0.995$

## Ejercicios 2.4

1. Sea  $X$  una variable aleatoria con distribución  $F$  con ocho y 20 grados de libertad, en el numerador y denominador, calcule el valor de  $k$ , tal que  $P(X(8, 20) > k) = 0.01$ .
2. Sea  $X$  una variable aleatoria con distribución  $F$  con 15 y 7 grados de libertad, en el numerador y denominador, calcule el valor de  $k$ , tal que  $P(X(15, 7) > k) = 0.976$ .
3. Calcule la mediana de una variable aleatoria  $F$  con ocho grados de libertad, tanto en el numerador como en el denominador. *Sugerencia.* La mediana es el valor medio, es decir, el valor  $m$  en el que  $P(X > m) = P(X < m)$ .
4. En una distribución  $F$  encuentre una expresión para la función de distribución acumulada en el caso en que  $\nu_1 = \nu_2 = 4$ .
5. En una distribución  $F$  encuentre una expresión para la función de distribución acumulada en el caso en que  $\nu_1 = \nu_2 = 6$ .
6. En una distribución  $F$  calcule el valor esperado en el caso en que  $\nu_1 = \nu_2 = 4$ .
7. En una distribución  $F$  calcule el valor esperado en el caso en que  $\nu_1 = \nu_2 = 3$ .
8. En una distribución  $F$  calcule el valor esperado y la varianza en el caso en que  $\nu_1 = \nu_2 = 5$ .

## 2.5 Muestra aleatoria

En la unidad previa mencionamos que la estadística descriptiva es la parte de la estadística que analiza, estudia y describe la totalidad de individuos de una población o muestra. Además, también definimos, entre otros conceptos, población y muestra, parámetro y estadístico, respectivamente. Ahora, aquí vemos que los parámetros y estadísticos son la base para el desarrollo de la *estadística inferencial*.

La **estadística inferencial** es la parte de la estadística que trabaja con muestras, a partir de las cuales pretende inferir aspectos relevantes de toda la población.

De esta definición se aprecia que en el estudio de la estadística inferencial es fundamental dar respuesta a ciertas preguntas cuyo estudio requiere de conocimientos en probabilidad y matemáticas. Por ejemplo, ¿cómo seleccionar la muestra?, ¿cómo realizar la inferencia?, ¿qué grado de confianza se puede tener en ésta? Para responder la primera pregunta, en la unidad 1 se menciona que la materia prima de la estadística son los conjuntos de números obte-

nidos al *contar* o *medir* elementos de algún fenómeno en estudio. Por esta razón, debemos tener especial cuidado para garantizar que la información sea completa y correcta. Entonces, el primer problema para los estadísticos reside en determinar qué información y en qué cantidad se habrá de reunir, ya que con base en ésta se establece la confiabilidad de los resultados. Por otra parte, en el caso de la segunda y tercera preguntas en las unidades 3 y 4 se verán técnicas para llevar a cabo inferencias sobre las que podremos establecer su grado de confianza.

Cuando es imposible o poco práctico analizar todo el conjunto de observaciones que constituyen a la población, pero se quiere llegar a conclusiones con respecto a cierta medida de ésta, se acostumbra usar el *muestreo aleatorio simple*. Ya hemos visto antes que este tipo de muestreo se caracteriza porque cualquier muestra de tamaño  $n$  de la población tiene la misma probabilidad de ser seleccionada que cualquier otra muestra del mismo tamaño. Es decir, con el *muestreo aleatorio* se elimina cualquier problema en el que se sobreestime o subestime de forma consciente o inconsciente alguna característica de la población. Así, las observaciones que se realizan deben ser *independientes* y *aleatorias*.

Se dice que las variables aleatorias  $X_1, X_2, \dots, X_n$  obtenidas del proceso de muestreo de una población forman una **muestra aleatoria simple** de tamaño  $n$ , si son *independientes* y *tienen la misma distribución de probabilidad que toda la población*.

Aunque la definición de muestra aleatoria parece simple, podemos observar que involucra varios conceptos de la teoría de probabilidades. En este sentido, las variables  $X_1, X_2, \dots, X_n$  constituyen una muestra aleatoria cuando cumple:

- Las variables  $X_1, X_2, \dots, X_n$  son independientes.
- Las variables  $X_1, X_2, \dots, X_n$  tienen la misma distribución.

En la unidad 1 nos referimos a  $x_1, x_2, \dots, x_n$  como una muestra, la cual se evitó nombrarla muestra aleatoria, ya que de acuerdo con la definición anterior  $x_i$  no es más que un valor de la variable  $X_i$ , para  $i = 1, 2, \dots, n$ , las cuales sí pueden formar una muestra aleatoria. Entonces, de aquí surge la pregunta, ¿cuál es el nombre más apropiado para  $x_1, x_2, \dots, x_n$ ?

En estadística, a las variables aleatorias  $X_1, X_2, \dots, X_n$  que forman una muestra aleatoria en forma simplificada se les conocen como variables *IID*, que significa **variables independientes e idénticamente distribuidas**.

Sea  $X_1, X_2, \dots, X_n$  una muestra aleatoria simple de tamaño  $n$  obtenida de una población, se le denomina **realización** de la muestra a los valores  $x_1, x_2, \dots, x_n$ , donde  $x_i$  es un valor de la variable  $X_i$ , para  $i = 1, 2, \dots, n$ .

Finalizamos esta sección formulando algunas preguntas comunes en la estadística inferencial, las cuales se responderán en las unidades siguientes.

- ¿Qué se puede decir acerca del parámetro en estudio? (véanse unidades 3 y 4).
- ¿Cuál es un valor factible para el parámetro? Pregunta que surge en la estimación puntual (véase unidad 3).
- ¿Cuál es un rango de valores posibles del parámetro? Pregunta que surge en los intervalos de confianza (véase unidad 4).
- ¿El parámetro resulta consistente con algún valor propuesto? Pregunta que surge en las pruebas de hipótesis (véase unidad 4).

## 2.6 Estadísticas importantes

Uno de los objetivos de la estadística inferencial consiste en obtener información de los parámetros. Por ejemplo:

- Suponga que la administración de una empresa que fabrica línea blanca quiere tener conocimiento acerca de la vida promedio de cierto modelo de refrigeradores.

En este caso se tendrían que observar a todos los refrigeradores en cuanto a su duración; como es de esperarse, este estudio sería muy costoso. Por tanto, se procede a realizar una inferencia con respecto a la vida promedio de dicho modelo de refrigeradores.

- Suponga que el gerente de una industria que fabrica focos quiere tener conocimiento acerca del promedio y la variabilidad de la vida de los focos de 100 watts que produce.

Al igual que en el caso anterior, si se quisiera conocer la vida promedio de dichos focos se tendría que probar todos para conocer su duración, pero como es de esperarse, este estudio sería muy laborioso y costoso. Por tanto, se procede a realizar una inferencia con respecto a la vida promedio de todos los focos de 100 watts que produce.

- En futuras elecciones se quisiera tener conocimiento sobre las tendencias para candidatos a la presidencia de la república.

En vísperas de elecciones, conocer las tendencias de la población hacia un cierto candidato juega un papel muy importante; como es de esperarse, no es posible estudiar a toda la población votante, hasta el día de las elecciones.

En este momento es fundamental diferenciar entre las variables aleatorias introducidas para las muestras y sus realizaciones. Suponga que  $X_1, X_2, \dots, X_n$  representan  $n$  variables aleatorias de la muestra y que  $x_1$  representa un valor correspondiente a la variable  $X_1$ , de igual manera,  $x_2$  representa el valor correspondiente a la variable  $X_2$ , y así de manera sucesiva. Por tanto, cuando se hace referencia a los valores de *una sola realización*  $x_1, x_2, \dots, x_n$  sus medidas como la media, mediana y variancia estarán dadas por las siguientes fórmulas (véase unidad 1).

$$\text{Media } \bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{k=1}^n x_k$$

$$\text{Variancia insesgada o muestral } s_{n-1}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \sum_{i=1}^n x_i^2 - \frac{n}{n-1} \bar{x}^2 = \frac{n}{n-1} s_n^2$$

Por ejemplo, si elegimos una realización de una muestra para cuatro refrigeradores que duraron 4.0, 4.1, 5.0 y 3.8 años, su vida promedio es:

$$\bar{x}_1 = \frac{4 + 4.1 + 5 + 3.8}{4} = 4.225 \text{ años.}$$

Se elige una segunda realización de la muestra de cuatro refrigeradores, pero ahora sus duraciones son: 5.2, 6.4, 7.0 y 5.9 años, entonces su duración promedio es:

$$\bar{x}_2 = \frac{5.2 + 6.4 + 7.0 + 5.9}{4} = 6.125 \text{ años.}$$

Como se puede observar, la media muestral varía de realización en realización. Por tanto, se concluye que  $\bar{x}$  no es más que un valor de una variable  $\bar{X}$ , a la que se le da el nombre de *estadística*, de la que en general tenemos la siguiente definición.

Se llama **estadística** o **estadístico** a cualquier función que se obtenga de las variables aleatorias correspondientes a una muestra aleatoria, pero que *no contenga algún parámetro*.

Veamos algunos ejemplos para ilustrar qué es un estadístico.



### Ejemplo 2.6 Estadístico

Sean  $X_1, X_2, \dots, X_n$  variables aleatorias con función de densidad  $f(x, \theta)$  donde  $\theta$  es el parámetro de la distribución. En este caso, los incisos a)-d) son ejemplos de estadísticas,  $T(\mathbf{X})$ , mientras que los incisos e) y f) no son estadísticas por depender del parámetro.

$$a) T(\mathbf{X}) = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad e) T(\mathbf{X}) = \sum_{i=1}^n (X_i - \theta)^2$$

$$b) T(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^n X_i^3 \quad f) T(\mathbf{X}) = \theta + \sum_{i=1}^n X_i^3$$

$$c) T(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^6$$

$$d) T(\mathbf{X}) = \sum_{i=1}^n (X_i - 5)^2$$

Cabe aclarar que, por tradición, en los diferentes textos de la materia por desgracia también se conoce como *estadística* a algún valor particular de la disciplina de *estadística*. Por ejemplo, en la duración promedio de los cuatro refrigeradores a las medias  $\bar{x}_1$  y  $\bar{x}_2$  se le conoce como *estadísticas*. Pero no son más que valores particulares de la *disciplina* que se obtienen de las realizaciones de las variables aleatorias  $X_1, X_2, X_3$  y  $X_4$  y que se calculan por medio de  $\bar{X} = \frac{X_1 + X_2 + X_3 + X_4}{4}$ .

A continuación, se muestra un resumen de las estadísticas más comunes utilizadas en la estadística inferencial en su parte metodológica, las cuales son analizadas junto con otras estadísticas en el texto.

## Media

Sea una muestra aleatoria  $X_1, \dots, X_n$ , entonces la estadística media está dada por:

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{1}{n} \sum_{k=1}^n X_k$$

## Diferencia de medias

Sean dos muestras aleatorias  $X_1, X_2, \dots, X_n$  y  $Y_1, Y_2, \dots, Y_n$  independientes, entonces la estadística de la diferencia está dada  $\bar{X} - \bar{Y}$ .

## Varianza insesgada o muestral

Sea una muestra aleatoria  $X_1, X_2, \dots, X_n$ , entonces la estadística varianza muestral está dada por:

$$S_{n-1}^2 = \frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X})^2 = \frac{1}{n-1} \sum_{k=1}^n X_k^2 - \frac{n}{n-1} \bar{X}^2$$

## Proporciones

Sea una muestra aleatoria  $X_1, X_2, \dots, X_n$  de distribuciones de Bernoulli, entonces la estadística para las proporciones está dada por  $\bar{X} = \frac{T}{n}$ , en donde  $T = \sum_{i=1}^n X_i$  y tiene distribución binomial.

## Media y varianza de la media muestral

La estadística que más se emplea en la práctica es la media muestral y se usa para llevar a cabo inferencias con respecto al parámetro media,  $\mu$ . Por eso, es conveniente que se estudien dos propiedades sencillas de la estadística  $\bar{X}$  que resultan con bastante frecuencia en las inferencias para una muestra aleatoria.

### Teorema 2.2

Sea  $X_1, X_2, \dots, X_n$  una muestra aleatoria tomada con reemplazo de una distribución con valor medio  $\mu$  y varianza finita  $\sigma^2$ , entonces:

$$E(\bar{X}) = \mu_{\bar{X}} = \mu \text{ y } V(\bar{X}) = \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$$

De tal forma que si  $X = X_1 + \dots + X_n$ , entonces  $E(X) = n\mu$   $V(X) = n\sigma^2$ .

Observe que en las condiciones del teorema no se indica la distribución de las variables, lo único que se pide es que sean *IID*.

## Media y varianza de una diferencia de medias

Otra de las estadísticas de gran uso se refiere a la diferencia de medias para muestras independientes, en esta situación también se puede formular un teorema similar al 2.2, en el que no se indica la distribución de las variables, pero podemos llegar a un resultado general para calcular la media y varianza correspondientes a la estadística de la diferencia de medias.

### Teorema 2.3

Sean  $X_1, X_2, \dots, X_{n_1}$  y  $Y_1, Y_2, \dots, Y_{n_2}$  dos muestras aleatorias independientes de dos poblaciones con  $n_1$  observaciones de la población 1 y  $n_2$  observaciones de la población 2; si las medias y varianzas (finitas) de las poblaciones 1 y 2 son  $(\mu_1, \sigma_1^2)$  y  $(\mu_2, \sigma_2^2)$ , respectivamente, entonces:

$$E(\bar{X} - \bar{Y}) = \mu_1 - \mu_2 \text{ y } V(\bar{X} - \bar{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

El resultado se deduce en forma inmediata de la definición de muestra aleatoria y las propiedades del valor esperado y la varianza para variables independientes (véase teorema 2.2).

## 2.7 Distribuciones muestrales asociadas a la normal

En la sección anterior estudiamos las estadísticas más comunes y las fórmulas para el cálculo de la media y la varianza. Además, hablamos de la importancia de una estadística para realizar inferencias, ahora lo que hace falta es tener más conocimiento sobre el comportamiento de dichas estadísticas.

En primera instancia, recordemos que una estadística es una *variable aleatoria* que depende solo de la muestra aleatoria en estudio, por tanto, debe tener una distribución de probabilidad que describa su comportamiento.

Se llama **distribución muestral** a la distribución de probabilidad de la estadística en estudio.

### Ejemplos 2.7 Distribución muestral

1. La distribución muestral de  $\bar{X}$  se llama *distribución muestral de la media*.
2. La distribución muestral de  $S^2$  es la *distribución muestral de la variancia*.

En este momento podemos mencionar que una de las razones principales por la que se estudian las distribuciones de probabilidad consiste en poder determinar y evaluar las propiedades de las distribuciones de las estadísticas en estudio, ya que proporcionan las bases para poder llevar a cabo mejores inferencias sobre los parámetros de la distribución.

Desde un punto de vista práctico, la distribución muestral representa un *modelo teórico* para el histograma de frecuencias relativas o absolutas que se obtienen con los valores correspondientes a las realizaciones. Esto se debe a que la forma de la distribución muestral teórica de un estadístico depende de la distribución de las variables de la muestra. Esto último es una de las razones para el estudio de los histogramas de frecuencias que se tratan en la unidad 1.

A continuación, se presentan algunas de las distribuciones muestrales más comunes que se usan en el estudio de la estadística inferencial. Iniciamos con el caso de mayor aplicación; esto es, para variables que tienen distribución normal. Las siguientes secciones pueden parecer bastante teóricas para los estudiantes de administración y de las diferentes ingenierías, sin embargo, resultan útiles para entender algunas situaciones más complejas que las revisadas en las unidades subsecuentes. Por ejemplo, al tener variables aleatorias con la misma distribución con frecuencia se piensa que su suma o producto debe tener la misma distribución, pero veremos que no siempre es así.

## Sumas, promedios y combinaciones lineales de variables aleatorias normales con la misma media y varianza

El siguiente teorema se usa cuando se tiene una muestra aleatoria de **variables normales** e indica qué tipo de distribución tendrá la suma o el promedio.

### Teorema 2.4

Sean  $X_1, X_2, \dots, X_n$  las variables de la muestra aleatoria de una distribución normal con parámetros  $\mu$  y  $\sigma^2$  (finita), entonces:

a) La estadística suma  $T = \sum_{i=1}^n X_i$  es normal con media  $n\mu$  y varianza  $n\sigma^2$ .

Luego,  $Z = \frac{T - E(T)}{\sqrt{V(T)}} = \frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}}$  tiene una **distribución normal estándar**.

b) La estadística media  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  es normal con media  $\mu$  y varianza  $\frac{\sigma^2}{n}$ .

Luego,  $Z = \frac{\bar{X} - E(\bar{X})}{\sqrt{V(\bar{X})}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \left( \frac{\bar{X} - \mu}{\sigma} \right) \sqrt{n}$  tiene una **distribución normal estándar**.

A continuación, se muestran algunos ejemplos para ilustrar los cálculos del teorema anterior.

### Ejemplo 2.8 Teorema 2.4

1. Cierta tipo de tornillos que se fabrica tiene una distribución normal con un diámetro promedio de 10 mm y una desviación estándar de 1 mm. ¿Cuál es la probabilidad de que una muestra aleatoria de 10 tornillos tenga un diámetro promedio menor o igual a 10.05 mm?

**Solución**

Sean  $X_1, X_2, \dots, X_{10}$  las variables aleatorias que representan los diámetros en milímetros de las 10 variables aleatorias para los 10 tornillos. Por las condiciones del problema, cada una tiene distribución normal, con  $\mu = 10$  y  $\sigma = 1$  y el tamaño de la muestra es de  $n = 10$ . Por otro lado, la probabilidad que se pide está dada por:

$$P(\bar{X} \leq 10.05)$$

De manera que se puede utilizar el resultado del inciso *b)* del teorema 2.4. Es decir,  $\bar{X}$  tiene una distribución normal, por tanto, para el cálculo de la probabilidad se requiere la media y la desviación estándar de la variable  $\bar{X}$ . Así:

$$\begin{aligned} P(\bar{X} \leq 10.05) &= P\left(\frac{\bar{X} - E(\bar{X})}{\sqrt{V(\bar{X})}} \leq \frac{10.05 - E(\bar{X})}{\sqrt{V(\bar{X})}}\right) = P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq \frac{10.05 - 10}{1/\sqrt{10}}\right) = P(Z \leq 0.16) = \Phi(0.16) \\ &= 0.5636 \end{aligned}$$

2. Suponga que el peso de una población es una variable aleatoria normal con  $\mu = 72$  kg y  $\sigma^2 = 16$  kg<sup>2</sup>. Si siete personas suben a un ascensor que tiene una capacidad máxima de 500 kg. ¿Cuál es la probabilidad de que el ascensor no funcione?

**Solución**

Sean  $X_1, X_2, \dots, X_7$  las variables aleatorias que representan los pesos de las personas en kilogramos. Por condiciones del problema, cada una tiene una distribución normal, con  $\mu = 72$  kg y  $\sigma^2 = 16$  kg<sup>2</sup>, con una muestra de tamaño  $n = 7$ . Por otro lado, la probabilidad que se pide para que no funcione el elevador es que la suma de los pesos de las siete personas sea mayor a 500 kilogramos, así:

$$P\left(\sum_{i=1}^7 X_i > 500\right)$$

Se puede utilizar el resultado *a)* del teorema 2.4, es decir,  $T = \sum_{i=1}^7 X_i$  tiene distribución normal. Por tanto, para el cálculo de la probabilidad se requiere la media y desviación estándar de la estadística  $T = \sum_{i=1}^7 X_i$ . Del teorema 2.4 inciso *a)*,  $E(T) = n\mu$  y  $\sqrt{V(T)} = \sigma\sqrt{n}$ . Entonces:

$$\begin{aligned} P\left(\sum_{i=1}^7 X_i > 500\right) &= P\left(\frac{T - n\mu}{\sigma\sqrt{n}} > \frac{500 - 7 \times 72}{4 \times \sqrt{7}}\right) = P(Z > -0.38) = 1 - P(Z \leq -0.38) = \Phi(0.38) \\ &= 0.6480 \end{aligned}$$

3. Un guardabosque que estudia los efectos de la fertilización en ciertos bosques de pino se interesa en estimar el área fertilizada promedio de la base de éstos. Al estudiar las áreas de las bases de árboles similares durante muchos años descubrió que estas mediciones (en pulgadas cuadradas) tienen una distribución normal con desviación estándar de alrededor de 5 in<sup>2</sup>. Si el guardabosque selecciona una muestra de 16 árboles, encuentre la probabilidad de que la media de la muestra se desvíe máximo 2 in<sup>2</sup> de la media de la población.

**Solución**

Sean  $X_1, X_2, \dots, X_{16}$  las variables aleatorias que representan las áreas de las bases de los 16 árboles. Por condiciones del problema, las variables tienen una distribución normal, con  $\sigma = 5$  y tamaño de muestra  $n = 16$ . Por otro lado, la probabilidad que se pide es:

$$P(|\bar{X} - \mu| \leq 2)$$

El valor absoluto se debe a que no se especifica qué es mayor, si la media muestral o la poblacional, y solo se menciona que exista una **desviación** entre éstas. De tal forma que se puede utilizar el resultado *b)* del teorema 2.4. Es decir,  $\bar{X}$  tiene una distribución normal, para el cálculo de la probabilidad requerida se puede estandarizar, para esto se necesita la desviación estándar de  $\bar{X}$ :

$$P(|\bar{X} - \mu| \leq 2) = P\left(\left|\frac{\bar{X} - E(\bar{X})}{\sqrt{V(\bar{X})}}\right| \leq \frac{2}{5/\sqrt{16}}\right) = P(|Z| \leq 1.60) = D(1.60) = 0.8904$$

## Cálculo del tamaño de la muestra en distribuciones normales

En ocasiones se quiere conocer el tamaño de la muestra si se parte de la probabilidad y la distribución normal de las variables. En estos casos se utilizan las tablas porcentuales de la normal y las fórmulas del teorema 2.4, de donde se despeja el tamaño de la muestra.

### Ejemplo 2.9 Tamaño mínimo de la muestra

El diámetro promedio de cierto tipo de tornillos que se fabrica es de 10 mm y tiene una desviación estándar de 2 mm. Suponga que la distribución de los diámetros de los tornillos es normal.

- ¿Cuál es el tamaño mínimo de la muestra que debe seleccionarse para que el promedio de los diámetros sea menor a 9.5 mm con una probabilidad menor a 0.05?
- ¿Cuál es el tamaño mínimo de la muestra que debe seleccionarse para que el promedio de los diámetros sea mayor a 11 mm con una probabilidad máxima de 0.05?

#### Solución

Sean  $X_1, X_2, \dots, X_n$ , las variables aleatorias que representan los diámetros en milímetros de los tornillos. Por condiciones del problema las variables tienen distribución normal con  $\mu = 10$  y  $\sigma = 2$ .

- Por otro lado, la probabilidad que se indica es  $P(\bar{X} < 9.5) < 0.05$ .

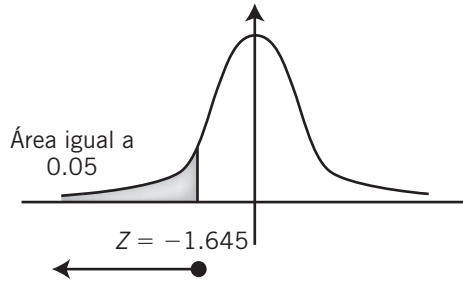
De tal forma que se puede utilizar el resultado *b)* del teorema 2.4 y despejar al tamaño de la muestra. Luego,  $\bar{X}$  tiene una distribución normal, por tanto, para el cálculo de la probabilidad se requiere la media y la desviación estándar de la variable  $\bar{X}$ :

$$P(\bar{X} \leq 9.50) = P\left(\frac{\bar{X} - E(\bar{X})}{\sqrt{V(\bar{X})}} \leq \frac{9.50 - E(\bar{X})}{\sqrt{V(\bar{X})}}\right) = P\left(Z \leq \frac{9.50 - 10}{2/\sqrt{n}}\right) = P(Z \leq -0.25\sqrt{n}) < 0.05$$

De las tablas porcentuales para un valor  $Z(0.05) = -1.645$ , resulta que  $P(Z < -1.645) = 0.05$ . Así, se obtiene que  $-0.25\sqrt{n} \leq -1.645$ .

## Explicación de la desigualdad anterior

En la figura 2.21 se puede apreciar que el valor buscado de  $Z$  tiene que ser menor a  $-1.645$  para que la probabilidad (área bajo la curva) sea menor a 0.05. De manera que si se despeja el tamaño de la muestra y se cambia el signo, la desigualdad se invierte  $0.25\sqrt{n} \geq 1.645$ . Por último  $n \geq \left(\frac{1.645}{0.25}\right)^2 \cong 43.3$ . Entonces, el tamaño adecuado de la muestra se puede considerar  $n = 44$ .



**Figura 2.21** Representación del área de integración inciso a).

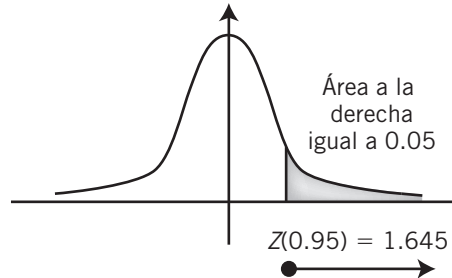
b) De forma similar al inciso a)  $P(\bar{X} > 11) \leq 0.05$ :

$$P(\bar{X} > 11) = P\left(\frac{\bar{X} - E(\bar{X})}{V(\bar{X})} > \frac{11 - E(\bar{X})}{V(\bar{X})}\right) = P\left(Z > \frac{11 - 10}{2/\sqrt{n}}\right) = P(Z > 0.5\sqrt{n}) \leq 0.05$$

Para poder usar las tablas porcentuales es necesario emplear la simetría de la distribución normal  $P(Z < -0.5\sqrt{n}) \leq 0.05$ . Otra forma sería por el complemento, es decir,  $P(Z \leq 0.5\sqrt{n}) > 1 - 0.05 = 0.95$ .

### Explicación de la desigualdad anterior

En la figura 2.22 se aprecia que el valor buscado de  $Z$  tiene que ser mayor a 1.645 para que la probabilidad (área bajo la curva) sea menor a 0.05, lo cual equivale a que el área izquierda de la gráfica sea mayor a 0.95.



**Figura 2.22** Representación del área de integración inciso b).

Así, de esta forma se obtiene el valor  $Z(0.95) = 1.645$ . Entonces,  $0.5\sqrt{n} \geq 1.645$ , de donde se despeja el tamaño de la muestra:

$$n \geq \left(\frac{1.645}{0.5}\right)^2 \cong 10.8$$

Por último, el tamaño adecuado de la muestra, se puede considerar  $n = 11$ .

### Fórmulas para el tamaño mínimo de muestra en distribuciones normales

En el ejemplo 2.9, el cálculo del tamaño mínimo de muestra requerido se resuelve de manera formal, resultando un poco laborioso su cálculo. Sin embargo, en la práctica se puede encontrar el tamaño mínimo de muestra mediante el empleo del teorema 2.5.

#### Teorema 2.5

Sea  $X_1, X_2, \dots, X_n$  una muestra aleatoria de variables también aleatorias con distribución normal y parámetros  $\mu$  y  $\sigma^2$ , entonces el **tamaño mínimo de la muestra** para que:

- a)  $P(\bar{X} \leq x_0) < \alpha$  condición  $x_0 < \mu$  y  $\alpha < 0.50$   
 b)  $P(\bar{X} > x_0) > \alpha$  condición  $x_0 < \mu$  y  $\alpha > 0.50$   
 c)  $P(\bar{X} \leq x_0) > \alpha$  condición  $x_0 > \mu$  y  $\alpha > 0.50$   
 d)  $P(\bar{X} > x_0) < \alpha$  condición  $x_0 > \mu$  y  $\alpha < 0.50$

se obtiene con la fórmula (en donde  $[q]$  representa la parte entera de  $q$ ,  $[3.6] = 3$ ):

$$n \geq \left( \frac{\sigma \Phi^{-1}(\alpha)}{x_0 - \mu} \right)^2 \text{ y se elige } n = \left[ \left( \frac{\sigma \Phi^{-1}(\alpha)}{x_0 - \mu} \right)^2 \right] + 1$$

- a)  $P(|\bar{X} - \mu| < d_0) > \alpha$  condición  $\alpha > 0.50$  o su complemento b).

$$n \geq \left( \frac{\sigma \Phi^{-1}((1-\alpha)/2)}{d_0} \right)^2 \text{ y se elige } n = \left[ \left( \frac{\sigma \Phi^{-1}((1-\alpha)/2)}{d_0} \right)^2 \right] + 1$$

- c)  $P(|\bar{X} - \mu| > d_0) < \alpha$  condición  $\alpha < 0.50$  o su complemento d).

$$n \geq \left( \frac{\sigma \Phi^{-1}(\alpha/2)}{d_0} \right)^2 \text{ y se elige } n = \left[ \left( \frac{\sigma \Phi^{-1}(\alpha/2)}{d_0} \right)^2 \right] + 1$$

El resultado a) con b)  
son complementos.  
El resultado c) con d)  
son complementos.

### Ejemplo 2.10 Teorema 2.5

Resuelva el ejemplo 2.9, utilice el teorema 2.5.

#### Solución

- a)  $P(\bar{X} \leq 9.50) < 0.05$ ; en este caso  $x_0 = 9.50$ ,  $\mu = 10$ , se cumple  $x_0 < \mu$  y  $\alpha = 0.05 < 0.50$ , entonces estamos en el inciso a) con  $\sigma = 2$ ,  $\Phi^{-1}(\alpha) = -1.645$ :

$$n \geq \left( \frac{\sigma \Phi^{-1}(\alpha)}{x_0 - \mu} \right)^2 = \left( \frac{2(-1.645)}{9.5 - 10} \right)^2 \cong 43.3$$

El **tamaño mínimo** de la muestra que debe elegirse es  $n = [43.3] + 1 = 43 + 1 = 44$ .

- b)  $P(\bar{X} > 11) < 0.05$ , en este caso  $x_0 = 11$ ,  $\mu = 10$  se cumple  $x_0 > \mu$  y  $\alpha = 0.05 < 0.50$ , entonces estamos en el inciso d) con  $\sigma = 2$ ,  $\Phi^{-1}(\alpha) = -1.645$ .

$$n \geq \left( \frac{\sigma \Phi^{-1}(\alpha)}{x_0 - \mu} \right)^2 = \left( \frac{2(-1.645)}{11 - 10} \right)^2 \cong 10.82$$

Entonces, el **tamaño mínimo** de la muestra que debe elegirse es  $n = [10.82] + 1 = 10 + 1 = 11$ .

### Ejercicios 2.5

- Un vendedor de automóviles sospecha que su margen de beneficios promedio por auto vendido está por debajo del promedio nacional de \$700. Se admite que el margen de beneficios por auto vendido tiene una desviación estándar de \$30. Una muestra aleatoria de 10 autos vendidos da un margen de beneficios  $\bar{X}$ . Suponga que la población se distribuye de manera normal.
  - Calcule la probabilidad de que  $\bar{X}$  sea inferior a \$680.
  - Calcule la probabilidad de que  $\bar{X}$  difiera de la media poblacional en menos de \$20.
  - En estas condiciones, y si se supone desconocido  $n$ , encuentre el tamaño mínimo de la muestra requerido para que la media muestral sea inferior a \$680 con una probabilidad máxima de 0.05.

2. El peso de ciertos paquetes de azúcar es una variable aleatoria normal con una media de 48 g y una desviación estándar de 5 g. Se toman al azar 21 paquetes de azúcar para cubrir una necesidad de 1 kg. ¿Cuál es la probabilidad de que no alcance el azúcar?
3. Se sabe que el tiempo promedio de reacción a un estímulo auditivo es una variable aleatoria con distribución normal con  $\mu = 0.15$  s y  $\sigma = 0.03$  s para personas de oído normal. Encuentre el tamaño mínimo de la muestra que se debe tomar si se requiere con 95% de seguridad que el tiempo medio de reacción muestral sea menor a 0.153 segundos.
4. La cantidad promedio que se gasta una empresa durante un año en servicios médicos por cada empleado fue de \$2575 con una desviación estándar de \$325. Suponga una población normal.
  - a) ¿Cuál es la probabilidad de que a partir de una muestra aleatoria de 20 empleados se observe una media muestral comprendida entre \$2500 y \$2700?
  - b) Si el gerente de la empresa dice que ha realizado un estudio estadístico y obtuvo que el porcentaje de empleados en los que gasta más de \$2 500 al año en servicios médicos es mayor a 90%. ¿Cuál será el tamaño mínimo de la muestra aleatoria que debió de haber considerado el gerente para llevar a cabo el estudio estadístico?
5. Para determinar la calidad de diferentes tipos de secadoras de cabello que se venden en el mercado, una empresa ha realizado un estudio intensivo de este tipo de producto. El estudio reportó que la duración de los secadores de una marca determinada tiene una distribución normal con una media de 1 200 horas y una desviación estándar de 100 horas.
  - a) Se elige una muestra aleatoria de tamaño 25, ¿cuál es la probabilidad de que el tiempo promedio de duración de la muestra esté entre 1 150 y 1 250 horas?
  - b) Si el fabricante afirma que puede probar de manera estadística que más de 90% de sus secadoras duran en promedio al menos 1 180 horas, ¿cuál es el tamaño mínimo de la muestra que debe el fabricante elegir para que pueda corroborar su afirmación?
6. Una planta industrial fabrica bombillas de luz cuya duración tiene una distribución normal con una media de 780 horas y una desviación estándar de 50 horas.
  - a) Calcule la probabilidad de que al seleccionar una muestra aleatoria de 25 bombillas tengan una duración promedio mayor a 800 horas. ¿Qué puede concluir?
  - b) Si un grupo de compradores afirma que puede probar de manera estadística que menos de 20% de las bombillas duran en promedio menos de 770 horas, ¿cuál es el tamaño mínimo de la muestra que deben elegir los compradores para que puedan corroborar su afirmación?
7. Con base en el problema de las bombillas.
  - a) Suponga que no conoce la media poblacional y la desviación estándar de la vida de las bombillas es de 50 horas, por lo que se toma una muestra de 30 bombillas. Calcule la probabilidad de que la vida promedio muestral se desvíe de la verdadera media en menos de 20 horas.
  - b) Suponga que no conoce ni la media ni la varianza poblacional y se toma una muestra de 30 bombillas. Calcule la probabilidad de que la vida promedio muestral se desvíe de la verdadera media en menos de 0.5 veces la desviación estándar poblacional.
8. Una máquina envasa sal en bolsas con  $X$  kg de peso; después se introducen en cajas que contienen 20 bolsas cada una. Si  $X$  es una variable aleatoria distribuida de manera normal con media de 2 kg y desviación estándar de 0.20 kg:
  - a) ¿Qué porcentajes de cajas llenas se espera que pesen entre 41 y 42 kg?
  - b) Después de 41.5 kg la caja, el producto pierde \$3 por cada kilogramo de más. ¿Cuánto se espera que pierda en el siguiente embarque de 10 000 cajas?



9. Se ha hecho un estudio en la Ciudad de México sobre las ganancias de los taxistas. Se estima que la ganancia media de los taxistas es de \$650 diarios con una desviación estándar de \$100 y distribución normal. De ser ciertas estas estimaciones y el vehículo pertenece al chofer.

- a) ¿Cuál sería la probabilidad de que el chofer gane al mes (25 días de trabajo) más de \$17 500?  
 b) ¿Cuál sería la probabilidad de que el chofer gane al mes (25 días de trabajo) menos de \$14 500?

## Diferencia de medias de distribuciones normales

Los teoremas anteriores se formularon para sumas y promedios de variables aleatorias con distribución normal, pero todas proceden de una misma muestra.

Otro problema que resulta en las aplicaciones de la estadística se refiere a la comparación de varias poblaciones. Por ejemplo, cuando se tienen dos máquinas o líneas de producción de un mismo producto, puede ser de interés conocer cuál es más eficiente. Esta comparación se lleva a cabo de una manera sencilla cuando conocemos la distribución de la diferencia de las estadísticas medias correspondientes. En el caso de dos poblaciones normales se pueden tomar muestras independientes y la distribución también resulta normal.

A continuación, se formula otro teorema para la diferencia de medias de distribuciones normales.

### Teorema 2.6

Sean  $X_1, X_2, \dots, X_{n_1}$  y  $Y_1, Y_2, \dots, Y_{n_2}$  dos muestras aleatorias independientes de dos poblaciones normales con  $n_1$  observaciones de la población 1 y  $n_2$  observaciones de la población 2, si las medias y varianzas (finitas) de las poblaciones 1 y 2 son  $(\mu_1, \sigma_1^2)$  y  $(\mu_2, \sigma_2^2)$ , respectivamente, entonces  $\bar{X} - \bar{Y}$

tiene distribución normal con media  $\mu_1 - \mu_2$  y varianza  $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$ . Luego,

$Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$  tiene una distribución **normal estándar**.

La distribución  $Z$  resulta útil para hacer inferencias acerca de la **media** ( $\mu$ ) y **diferencia de medias** ( $\mu_1 - \mu_2$ ) de poblaciones con distribuciones normales cuando se conocen las varianzas poblacionales.

### Ejemplo 2.11 Diferencia de medias

Dos fabricantes de cables ( $A$  y  $B$ ) afirman que su producto tiene una resistencia distribuida en forma normal con promedio a la rotura de 4 500 y 4 300 lb con varianzas de 900 y 800 lb, respectivamente. Si se seleccionan 20 cables del fabricante  $A$  y 15 cables del fabricante  $B$ , ¿cuál es la probabilidad de que la media a la resistencia de la rotura de los cables del fabricante  $A$  sea al menos 220 lb mayor que la resistencia promedio de los cables del fabricante  $B$ ?

#### Solución

Sean  $X_1, X_2, \dots, X_{20}$  las variables aleatorias que representan la resistencia a la rotura de los cables del fabricante  $A$  y  $Y_1, Y_2, \dots, Y_{15}$  las variables aleatorias que representan la resistencia a la rotura de los cables del fabricante  $B$ . Por condiciones del problema, las variables tienen una distribución normal, con  $\mu_1 = 4500$ ,  $\sigma_1^2 = 900$  y  $\mu_2 = 4300$ ,  $\sigma_2^2 = 800$ . Por otro lado, la probabilidad que se pide es:

$$P(\bar{X} \geq \bar{Y} + 220) = P(\bar{X} - \bar{Y} \geq 220)$$

Con base en el teorema 2.6, resulta:

$$P\left(\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \geq \frac{220 - (4500 - 4300)}{\sqrt{\frac{900}{20} + \frac{800}{15}}}\right) = P(Z \geq 2.02) = \Phi(-2.02) = 0.0217$$

Significa que es muy poco probable que la resistencia media a la rotura de los cables del fabricante de cable tipo A sea al menos de 220 lb mayor que la resistencia promedio de los cables del tipo B.

## Cálculo del tamaño de la muestra para diferencia de medias

El cálculo del tamaño mínimo de muestra deseado para que se cumplan ciertas condiciones se complica por tener dos tamaños de muestra. Pero, si se contempla la sugerencia de que ambos tamaños de muestra sean iguales, el problema se reduce a uno similar a los vistos, donde los casos son los mismos que en el teorema 2.5, y donde consideramos que  $\mu_1 > \mu_2$  y  $D_0 > 0$ :

- a)  $P(\bar{X}_1 - \bar{X}_2 \leq D_0) < \alpha$  condición  $D_0 < \mu_1 - \mu_2$  y  $\alpha < 0.50$
- b)  $P(\bar{X}_1 - \bar{X}_2 > D_0) > \alpha$  condición  $D_0 < \mu_1 - \mu_2$  y  $\alpha > 0.50$
- c)  $P(\bar{X}_1 - \bar{X}_2 \leq D_0) > \alpha$  condición  $D_0 > \mu_1 - \mu_2$  y  $\alpha > 0.50$
- d)  $P(\bar{X}_1 - \bar{X}_2 > D_0) < \alpha$  condición  $D_0 > \mu_1 - \mu_2$  y  $\alpha < 0.50$

Se obtiene con la fórmula (en donde  $[q]$  representa la parte entera de  $q$ ,  $[3.6] = 3$ ), por ejemplo:

$$n \geq \left( \frac{\sqrt{\sigma_1^2 + \sigma_2^2} \Phi^{-1}(\alpha)}{D_0 - (\mu_1 - \mu_2)} \right)^2 \text{ y se elige } n = \left\lceil \left( \frac{\sqrt{\sigma_1^2 + \sigma_2^2} \Phi^{-1}(\alpha)}{D_0 - (\mu_1 - \mu_2)} \right)^2 \right\rceil + 1$$

- e)  $P(|\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)| \leq D_0) > \alpha$  condición  $\alpha > 0.50$  o su complemento:

$$n \geq \left( \frac{\sqrt{\sigma_1^2 + \sigma_2^2} \Phi^{-1}((1 - \alpha)/2)}{D_0} \right)^2 \text{ y se elige } n = \left\lceil \left( \frac{\sqrt{\sigma_1^2 + \sigma_2^2} \Phi^{-1}((1 - \alpha)/2)}{D_0} \right)^2 \right\rceil + 1$$

### Ejemplo 2.12 Tamaño de la muestra para diferencia de medias

En el ejemplo 2.11 suponga que el fabricante A afirma que de manera estadística su producto tiene en más de 90% de los casos una resistencia promedio a la rotura de al menos 190 lb mayor que la resistencia promedio de los cables del fabricante B. ¿Cuál es el tamaño mínimo de las muestras que debe tomar el fabricante A para corroborar su afirmación de manera estadística?

#### Solución

Por condiciones del problema tenemos  $P(\bar{X} - \bar{Y} \geq 190) > 0.90$ , donde  $D_0 = 190$  y  $\mu_1 = 4500$  y  $\mu_2 = 4300$ , entonces  $D_0 < \mu_1 - \mu_2$  y  $\alpha > 0.50$  estamos en el caso b):

$$n = \left\lceil \left( \frac{\sqrt{\sigma_1^2 + \sigma_2^2} \Phi^{-1}(\alpha)}{D_0 - (\mu_1 - \mu_2)} \right)^2 \right\rceil + 1 = \left\lceil \left( \frac{\sqrt{900 + 800} \Phi^{-1}(0.90)}{190 - (4500 - 4300)} \right)^2 \right\rceil + 1 = 28$$

Es decir, el fabricante A debe tomar las muestras de tamaños  $n_1 = n_2 = n = 28$ .

## Ejercicios 2.6

### Distribución de diferencia de medias de variables normales

En los siguientes ejercicios, cuando se pida calcular el tamaño de las muestras, se supondrán tamaños iguales  $n_1 = n_2 = n$ .

- Una compañía transnacional instituyó un programa de seguridad en el trabajo para reducir el tiempo perdido debido a accidentes. La empresa interesada en comprar el programa de seguridad toma la decisión de hacer un comparativo entre los tiempos perdidos antes y después de implantar el programa y establece que si al comparar 10 semanas la probabilidad de reducir al menos 14 horas en promedio es mayor a 0.80, entonces comprará dicho programa. Suponga que la empresa interesada sabe que la distribución del tiempo perdido actual por semana tiene una distribución  $N(108, 12^2)$ , mientras que la compañía asegura que la distribución semanal del tiempo perdido al implementar el programa tiene una distribución  $N(91, 14^2)$ .
  - Justifique de manera estadística si la empresa comprará el programa de seguridad.
  - Para vender el programa de seguridad, el gerente de mercadotecnia de la compañía asegura a las empresas interesadas que su programa reduce los tiempos promedio perdidos en más de 10 horas a la semana en más de 90% de los casos. ¿Cuál debe ser el tamaño mínimo de semanas que debe seleccionar el gerente para corroborar su afirmación?
- La variable de interés en una investigación es el puntaje obtenido por dos poblaciones de alumnos de último año de una prueba de rendimiento en matemáticas. Los investigadores suponen que los puntajes de las dos poblaciones estaban distribuidos normalmente con medias:  $\mu_1 = 50$ ,  $\mu_2 = 40$  y varianzas,  $\sigma_1^2 = 25$  y  $\sigma_2^2 = 49$ . Una muestra aleatoria de 10 alumnos se selecciona de la primera población y otra de forma independiente a la primera de 12 alumnos de la segunda población.
  - ¿Cuál es la probabilidad de que la diferencia entre las medias muestrales 1 menos la 2 esté comprendida entre 5 y 15 puntos?
  - ¿Cuál debe ser el tamaño mínimo de las muestras para que menos de 5% del promedio de alumnos de la muestra uno tenga una ventaja máxima de siete puntos sobre el promedio de los alumnos de la muestra 2?
- Se diseña un experimento para probar cuál de los operarios  $A$  o  $B$  obtiene el trabajo para manejar una nueva máquina. Para esto, se toma el tiempo de 20 pruebas que involucran la realización de cierto trabajo en la máquina para cada operario. Si los promedios de las muestras para las 20 pruebas difieren en menos de un segundo, se considera que el experimento termina en empate; en caso contrario, el operario con la media más pequeña obtiene el trabajo. Se supone que las varianzas de los tiempos para cada operario son de  $3.5 \text{ s}^2$ . Suponga que las poblaciones se distribuyen en forma normal.
  - Si ambos operarios son igual de hábiles, ¿cuál es la probabilidad de un empate?
  - ¿Cuál es el tamaño de muestra mínimo que se requiere para que la probabilidad de empate sea al menos de 0.98?
- Suponga que se realiza un estudio acerca del rendimiento de las empresas WM y TX, para lo cual se toman dos muestras aleatorias independientes de tamaño 12 de los rendimientos diarios para cada una de las empresas. Si se supone que los rendimientos tienen distribución normal con desviación estándar de 0.00776 y 0.00717, respectivamente:
  - Calcule la probabilidad de que la diferencia muestral de los rendimientos sea distinta de la verdadera diferencia en menos de 0.0035.
  - ¿Cuál es el tamaño de muestra mínimo que se requiere para que la probabilidad calculada en *a)* sea al menos de 0.90?
- Según un estudio de renta-habitación en departamentos de la Ciudad de México, en un área  $A$  el costo tiene una distribución normal con valor medio de \$3750 y desviación estándar de \$270 y en un área  $B$  el costo también tiene una distribución normal con valor medio de \$3900 y una desviación estándar de \$250. Para llevar a

cabo una validación del estudio, se piensa en tomar muestras independientes una en cada área de tamaños de 25 y 20 departamentos, respectivamente. Si la probabilidad de que la diferencia de medias  $\bar{X}_B - \bar{X}_A$  esté entre 0 y \$200, sea mayor a 90%, entonces se considerará válido el estudio. Si supone que no hay diferencia alguna entre las dos áreas respecto a los gastos mensuales de los departamentos y que las rentas mensuales tienen un comportamiento normal, ¿el estudio será considerado válido?

6. Los cinescopios para receptores de televisión de un fabricante  $A$  tienen una vida con distribución normal con media 6.5 años y una desviación estándar de 0.9 años; en tanto que la vida de los cinescopios de un fabricante  $B$  también se distribuyen en forma normal, pero con una media de 6.0 años y una desviación estándar de 0.8 años.
  - a) ¿Cuál es la probabilidad de que en una muestra de 26 cinescopios del fabricante  $A$  tenga una vida promedio que sea al menos un año mayor que la vida promedio de una muestra de 20 cinescopios del fabricante  $B$ ?
  - b) El fabricante  $A$  afirma que en al menos 80% de los casos sus cinescopios duran en promedio muestral más de tres meses que el del fabricante  $B$ . ¿Cuál debe ser el tamaño mínimo de la muestra que debe considerar el fabricante  $A$  para corroborar su afirmación de manera estadística?
7. El flujo de agua a través de los suelos depende, entre otras cosas, de la porosidad (la proporción del volumen de huecos) del suelo. Para la comparación de dos tipos de suelo arenoso se registran 30 mediciones respecto a la porosidad del suelo  $A$  y 22 mediciones para el suelo  $B$ . Suponga que las varianzas poblacionales son  $\sigma_A^2 = 0.01$  y  $\sigma_B^2 = 0.02$ . Calcule la probabilidad de que el valor de la diferencia entre las medias muestrales se aleje a lo más en 0.05 unidades de la diferencia entre las medias de las poblaciones. Suponga que ambas poblaciones se distribuyen en forma normal.

## 2.8 Distribuciones de Bernoulli

En el estudio de las variables aleatorias se dice que una variable aleatoria  $X$  tiene una distribución de Bernoulli cuando toma solo dos valores, 0 y 1. Además cumple que  $P(X = 1) = p$  y  $P(X = 0) = 1 - p$ . Una pregunta que surge con frecuencia se refiere a la distribución de la suma y promedio de dichas variables.

### Distribución de la suma de variables de Bernoulli (binomial)

Otra distribución de variables, además de la normal, que tiene gran uso es la de Bernoulli, debido a que se utiliza en situaciones donde se realizan encuestas y las respuestas solo pueden tomar dos valores. Por ejemplo, una persona ve un programa de televisión o no lo ve, una persona es o no fumadora, etcétera. En general, una variable aleatoria de este tipo por sí sola no es de mucha ayuda, pero cuando se analiza una muestra aleatoria de estas variables puede ser de mucho interés conocer el comportamiento de la suma de variables.

#### Teorema 2.7

Sean  $X_1, X_2, \dots, X_n$  las variables de una muestra aleatoria de una distribución de Bernoulli con parámetro  $p$ , entonces  $T = X_1 + X_2 + \dots + X_n$  es binomial con parámetros  $n$  y  $p$ .

#### Ejemplo 2.13 Distribución de Bernoulli

En el área metropolitana 15% de los adolescentes ha tenido algún contacto con la policía (desde amonestaciones hasta arrestos) por motivos de delincuencia juvenil. Se selecciona una muestra aleatoria de 20 adolescentes de esta área. ¿Cuál es la probabilidad de que entre ocho y 10, de la muestra, respectivamente, hayan tenido contacto con la policía por estas causas?

**Solución**

Sean  $X_1, X_2, \dots, X_{20}$  las variables aleatorias que representan si un adolescente ha tenido algún contacto con la policía. Por condiciones del problema, las variables tienen una distribución de Bernoulli, puesto que los valores de la variable son cero cuando el adolescente no ha tenido problemas con la policía y uno en caso contrario. Luego,  $T = X_1 + X_2 + \dots + X_{20}$  tiene una distribución binomial con  $p = 0.15$  y  $n = 20$  y la probabilidad que se pide  $P(8 \leq X \leq 10)$ . Con base en el teorema 2.7, resulta:

$$\begin{aligned} P(8 \leq T \leq 10) &= P(T = 8) + P(T = 9) + P(T = 10) \\ &= C_8^{20}(0.15)^8(0.85)^{12} + C_9^{20}(0.15)^9(0.85)^{11} + C_{10}^{20}(0.15)^{10}(0.85)^{10} \\ &= 0.0059 \end{aligned}$$

Esto significa que es muy poco probable que entre ocho y 10 adolescentes de los 20 muestreados haya tenido problemas con la policía.

**Media y varianza de una proporción**

Se ha visto que las distribuciones muestrales de una suma de variables de Bernoulli con probabilidades de éxito  $p$  resulta ser una distribución binomial con parámetros  $n$  y  $p$ . De tal forma que es de interés conocer el comportamiento del promedio de sus resultados favorables, lo que origina la siguiente definición.

Sea  $X_1, X_2, \dots, X_n$  una muestra aleatoria de una distribución de Bernoulli con parámetro  $p$  (probabilidad de éxito) y la estadística  $T = X_1 + X_2 + \dots + X_n$ , que representa la cantidad de éxitos de la muestra, entonces se denomina **proporción** a  $\hat{p} = \bar{X} = \frac{T}{n}$ .

Luego, el valor esperado y la varianza de una proporción están dados por el siguiente teorema.

**Teorema 2.8**

Sean  $X_1, X_2, \dots, X_n$  una muestra aleatoria de una distribución de Bernoulli con parámetro  $p$ , y sea  $\hat{p} = \bar{X}$ ; entonces  $E(\hat{p}) = p$  y  $V(\hat{p}) = \frac{p(1-p)}{n}$ . Además,  $\bar{X}$  tiene una distribución discreta con valores  $0, 1/n, 2/n, \dots, 1$  cuyas probabilidades respectivas son las de la binomial  $C_0^n p^0(1-p)^n, C_1^n p^1(1-p)^{n-1}, \dots, C_n^n p^n(1-p)^0$ .

**Ejemplo 2.14 Valor esperado y varianza de la proporción**

En un área metropolitana, 18% de los adolescentes ha tenido algún contacto con la policía (desde amonestaciones hasta arrestos) por motivos de delincuencia juvenil. Se selecciona una muestra aleatoria de 100 adolescentes de esta área. ¿Cuál es el valor esperado y la varianza de la proporción de adolescentes que ha tenido algún contacto con la policía para muestras aleatorias de tamaño 100?

**Solución**

Sean  $X_1, X_2, \dots, X_{100}$ , las variables aleatorias que representan si un adolescente ha tenido algún contacto con la policía. Por condiciones del problema, las variables tienen una distribución de Bernoulli, puesto que los valores de la variable son cero para el caso que no ha tenido problemas el adolescente con la policía y uno para el caso contrario. Luego,  $X = X_1 + X_2 + \dots + X_{100}$  la proporción es  $\hat{p} = \frac{X}{100}$  del teorema 2.8, resulta  $E(\hat{p}) = p = 0.18$  y  $V(\hat{p}) = \frac{p(1-p)}{n} = 0.001476$ .

## Media y varianza de una diferencia de proporciones

De forma similar a la diferencia de medias, en ocasiones resulta interesante comparar las proporciones de dos muestras. Por ejemplo, las proporciones de ventas de dos artículos, las proporciones de artículos buenos producidos por dos máquinas, etcétera.

### Teorema 2.9

Sean  $X_1, X_2, \dots, X_{n_1}$  y  $Y_1, Y_2, \dots, Y_{n_2}$  dos muestras aleatorias independientes tomadas de dos poblaciones de Bernoulli con parámetros  $p_1$  y  $p_2$  y proporciones  $\hat{p}_1 = \bar{X}$  y  $\hat{p}_2 = \bar{Y}$  respectivamente, entonces:

$$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2 \quad \text{y} \quad V(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

### Ejercicios 2.7

#### Distribución de sumas de variables de Bernoulli y proporciones

- Suponga que en cierta población de mujeres embarazadas, 90% de las que inician su tercer trimestre de embarazo ha tenido algún cuidado prenatal. Si se selecciona una muestra aleatoria de 15 mujeres de esta población:
  - ¿Cuál es la probabilidad de que el número de mujeres en la muestra que han tenido algún cuidado prenatal sea máximo de 13?
  - ¿Cuál es la probabilidad de que el número de mujeres en la muestra que han tenido algún cuidado prenatal sea mayor a 9?
  - Si se selecciona una muestra aleatoria de 200 mujeres de esta población, ¿cuál es la varianza de la proporción para las muestras de tamaño de 200 mujeres que han tenido algún cuidado prenatal?
- Se sabe que el medicamento estándar usado para tratar cierta enfermedad ha resultado efectivo en un lapso de tres días en 75% de los casos en que se empleó. Al evaluar la efectividad de un nuevo medicamento para tratar la misma enfermedad, los médicos del laboratorio donde se produce el medicamento aseguran que es 10% más efectivo que el estándar. Con base en la afirmación de los médicos, se administró a 10 pacientes que padecían la enfermedad.
  - ¿Cuál es la probabilidad de observar al menos tres pacientes que se recuperen de la muestra?
  - ¿Cuál es la probabilidad de observar entre tres y ocho pacientes que se recuperen de la muestra?
  - ¿Cuál es la varianza del nuevo medicamento para muestras aleatorias de tamaño 150?
- Un estudio llevado a cabo en un distrito de la ciudad dio como resultado que 70% de los trabajadores habían cambiado de empleo al menos una vez en su vida. Suponga que seleccionamos una muestra aleatoria de 20 trabajadores de este distrito.
  - ¿Cuál es la probabilidad de que entre seis y 10 hayan cambiado de empleo al menos una vez en su vida?
  - ¿Cuál es la probabilidad de que a lo más la mitad de la muestra haya cambiado de empleo al menos una vez en su vida?
- En un proceso de llenado de botellas de refresco 10% no se llenan por completo. Para este proceso se selecciona al azar una muestra de 30 botellas.
  - ¿Cuál es la probabilidad de que menos de seis botellas de la muestra no estén completamente llenas?
  - ¿Cuál es la varianza de la proporción de muestras aleatorias de tamaño 225 para botellas que estén casi llenas?

## 2.9 Teorema central del límite media y suma muestral

En las secciones anteriores estudiamos las distribuciones muestrales para la suma y promedio de distribuciones normales. Pero, en general, para cualquier distribución el problema es un poco complejo. Sin embargo, existe un método asintótico que sirve para determinar probabilidades de la suma o promedio de variables provenientes de una muestra aleatoria de cualquier distribución, el único requisito es que tengan media y varianzas finitas. A continuación, formalizamos este resultado asintótico.

### Teorema central del límite para la media de variables

Iniciamos esta sección con el estudio de la aplicación del teorema para la media, el cual afirma que la **distribución límite de la distribución muestral para la estadística media es normal**. Para mayor precisión, a continuación, vemos el teorema.

#### Teorema 2.10

##### Teorema central del límite (TCL) para la media

Sean  $X_1, X_2, \dots, X_n$  las variables de una muestra aleatoria de una distribución con valor medio  $\mu$  y variancia finita  $\sigma^2$ , entonces la forma límite de la distribución de la variable:

$$Z_n = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

tiene una distribución normal estándar cuando  $n$  se hace infinita.

Ahora bien, ¿para qué tipo de distribución de la muestra se aplica el teorema central del límite?

En la formulación de este teorema se puede apreciar que se aplica al *promedio* de variables, por consiguiente, también a la suma de las variables. Entonces, surge la pregunta: ¿existe alguna restricción para poder aplicarlo?

En la formulación del teorema central del límite podemos apreciar que solo se exige que el valor esperado y la varianza de la distribución de la variable aleatoria existan y sean finitos.

En ese sentido, en la formulación del teorema central del límite se habla de un tamaño de muestra infinita; pero, en la práctica ¿qué tamaño de muestra puede dar una buena aproximación?

Sabemos que un tamaño de muestra infinito es teórico y en la práctica nunca lo vamos a tener, entonces ¿para qué nos sirve el teorema central del límite? Se ha probado que a partir de muestras *grandes* de tamaño 30 la aplicación de este teorema da buenas aproximaciones, razón por la que en la mayoría de los libros metodológicos se aplica en muestras de tamaños mayores o iguales a 30.

Ahora vamos a revisar algunos ejemplos que muestren la aplicación de este resultado.

#### Ejemplos 2.15 Teorema central del límite

1. Se fabrica un cierto tipo de tornillos con un diámetro de 10 mm y una desviación estándar de 1 mm. ¿Cuál es la probabilidad de que una muestra aleatoria de 400 tornillos tenga un diámetro promedio menor o igual a 10.05 mm?

##### Solución

Sean  $X_1, X_2, \dots, X_{400}$  las variables aleatorias que representan los diámetros en milímetros de los 400 tornillos para  $\mu = 10$ ,  $\sigma = 1$  y  $n = 400$ . La probabilidad que se pide está representada por:

$$P(\bar{X} \leq 10.05)$$



Puesto que no se conoce la distribución de las variables aleatorias, no podemos conocer la distribución de la  $\bar{X}$ , por tanto, no es posible calcular la probabilidad pedida. Sin embargo, la población tiene media y varianza finita; además, el tamaño de la muestra es grande. De este modo, podemos aproximar la probabilidad de aplicar el teorema 2.10 a la variable  $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$  y tener en cuenta que  $Z$  se aproximaría a una distribución normal estándar para tamaños de muestra grandes:

$$P(\bar{X} \leq 10.05) = P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq \frac{10.05 - 10}{1/\sqrt{400}}\right) = P(Z \leq 1) \cong 0.8413$$

Ahora bien, ¿por qué no se utilizó el teorema 2.4 para calcular la probabilidad?

Nótese que en esta situación no podemos utilizar el teorema 2.4 porque en éste se pide que las variables provengan de una población normal, pero en este problema no conocemos la distribución de las variables.

2. En el ejemplo anterior, ¿cuál será el tamaño de muestra mínimo que debe elegirse, si se desea que el diámetro promedio de una muestra de tornillos se diferencie a lo más en 0.10 mm del verdadero diámetro con una probabilidad mayor o igual a 0.95?

### Solución

Sean  $X_1, \dots, X_n$  las variables aleatorias que representan los diámetros en milímetros de los  $n$  tornillos, para  $\mu = 10$  y  $\sigma = 1$ . Para determinar el tamaño de la muestra conocemos la probabilidad:

$$P(|\bar{X} - \mu| \leq 0.10) \geq 0.95$$

Puesto que no se conoce la distribución de los datos se puede utilizar una aproximación con el teorema central del límite:

$$P(|\bar{X} - \mu| \leq 0.10) = P\left(|Z| \leq \frac{0.10}{1/\sqrt{n}}\right) = P(|Z| \leq 0.10\sqrt{n}) \geq 0.95$$

De las tablas porcentuales de la normal se tiene que  $P(|Z| \leq 1.96) = 0.95$ . Entonces:

$$0.10\sqrt{n} \geq 1.96 \Rightarrow n \geq \left(\frac{1.96}{0.10}\right)^2 = 384.16, n \geq 385$$

También se puede aplicar la fórmula para el tamaño de la muestra del inciso e) del teorema 2.5:

$$n = \left\lceil \left[ \left( \frac{\sigma \Phi^{-1}((1-\alpha)/2)}{d_0} \right)^2 \right] \right\rceil + 1 = \left\lceil \left[ \left( \frac{\Phi^{-1}(0.025)}{0.10} \right)^2 \right] \right\rceil + 1 = 385$$

Por otro lado, ¿el teorema central del límite solo se aplica a la distribución muestral media?

La formulación del teorema central del límite dada en el teorema 2.10 es la clásica, pero vemos que es fácil poder ampliarla a la distribución muestral de la suma; véanse los siguientes teoremas.

## Teorema central del límite suma de variables

Otra aplicación del teorema central del límite se refiere a la suma de variables aleatorias, para lo cual se aplica el

resultado del teorema 2.10 con  $E\{T\} = E\left\{\sum_{i=1}^n X_i\right\} = n\mu$  y  $V\{T\} = V\left\{\sum_{i=1}^n X_i\right\} = n\sigma^2$ .



**Teorema 2.11****Teorema central del límite para la suma de variables**

Sean  $X_1, X_2, \dots, X_n$  las variables de una muestra aleatoria de una distribución con valor medio  $\mu$  y variancia

finita  $\sigma^2$ , entonces la forma límite de la distribución de la estadística  $T = \sum_{i=1}^n X_i$ :

$$Z = \frac{T - \mu_T}{\sigma_T} = \frac{T - n\mu}{\sqrt{n}\sigma}$$

tiene una distribución normal estándar cuando  $n$  se hace infinita.

Note que los teoremas 2.10 y 2.11 en realidad son el mismo al dividir numerador y denominador entre  $n$ . Ahora, vamos a revisar un ejemplo en el que se ilustre el uso del teorema 2.11.

**Ejemplo 2.16 Variables aleatorias**

Suponga que el peso de los paquetes de café de cierto tipo tiene una media de 1 kg y desviación estándar de 0.05 kg. Si en una caja se colocan 64 de esos paquetes, ¿cuál es la probabilidad de que el peso total de los paquetes esté entre 63 y 64.4 kg?

**Solución**

Sean  $X_1, X_2, \dots, X_{64}$  las variables aleatorias que representan el peso de los 64 paquetes de café, con  $\mu = 1$ ,  $\sigma = 0.05$  y  $n = 64$ . La probabilidad que se pide es con respecto a la suma de los pesos de los 64 paquetes. Es decir:

$$P\left(63 \leq \sum_{i=1}^{64} X_i \leq 64.4\right)$$

Puesto que no se conoce la distribución del peso de los paquetes, pero la muestra es grande, se puede emplear el teorema central del límite 2.11, para la suma de variables. Donde:

$$\begin{aligned} P\left(63 \leq \sum_{i=1}^{64} X_i \leq 64.4\right) &= P(63 \leq T \leq 64.4) = P\left(\frac{63 - 1 \times 64}{\sqrt{64} \times 0.05} \leq \frac{T - n\mu}{\sqrt{n}\sigma} \leq \frac{64.4 - 1 \times 64}{\sqrt{64} \times 0.05}\right) \\ &= P(-2.5 \leq Z \leq 1) \cong \Phi(1) - \Phi(-2.5) \\ &= 0.8413 - 0.0062 \\ &= 0.8351 \end{aligned}$$

**Ejercicios 2.8****Teorema central del límite para medias y totales**

1. Los tiempos de atención para los clientes de un banco son variables aleatorias *IID* con una media de 1.5 minutos y varianza de 1.0 minuto cuadrado. Debido a las demandas de servicio en el banco se plantea lo siguiente:
  - a) El gerente quiere realizar un estudio estadístico para tomar la decisión de contratar a un cajero más. Decide que si la probabilidad de atender a 85 clientes en menos de dos horas es menor a 0.25, entonces contratará otro cajero. ¿Qué hará el gerente?
  - b) El gerente asegura que el tiempo promedio en atender 80 clientes es menor a 1.7 minutos en al menos 95% de los casos. ¿Estadísticamente esto será cierto?

- c) ¿Cuál es el tamaño mínimo de muestra que debe considerar el gerente para que el tiempo promedio de espera no sobrepase 1.4 minutos con una probabilidad máxima de 0.10?
2. Suponga que los pesos de bultos de azúcar tienen una media de  $\mu = 50$  y desviación estandar  $\sigma = 0.75$  kg. Si en una camioneta se cargan 60 bultos, determinar lo que se pide.
- a) ¿Cuál es la probabilidad de que el peso total de esos bultos sea mayor a 3 010 kg?
- b) ¿Cuál es la probabilidad de que el peso promedio de los 60 bultos sea menor a 49.75 kg?
- c) ¿Cuál es la probabilidad de que el peso promedio de los 60 bultos se encuentre entre 49.5 y 50?
- d) El productor de los bultos de azúcar afirma que su producto tiene en promedio una cantidad mayor a 49.9 kg de azúcar en más de 95% de los bultos. ¿Qué tamaño mínimo de la muestra debe tomar para justificar estadísticamente su afirmación?
3. La cantidad promedio que gastó una empresa durante un año en servicios médicos por cada empleado fue de \$2 575 con una desviación estándar de \$325.
- a) ¿Cuál es la probabilidad de que a partir de una muestra aleatoria de 70 empleados se observe una media muestral comprendida entre \$2 500 y \$2 700?
- b) ¿Cuál es la probabilidad de que a partir de una muestra aleatoria de 100 empleados se observe una media muestral menor a \$2 600?
- c) El gerente de la empresa asegura que gasta un promedio mínimo anual de \$2 600 en servicios médicos en al menos 90% de sus empleados. ¿Qué tamaño mínimo de la muestra debe elegir el gerente para confirmar su afirmación de manera estadística?
- d) El líder sindical de la empresa afirma que la compañía gasta un promedio mínimo anual de \$2 550 en servicios médicos en al menos 25% de sus empleados. ¿Qué tamaño mínimo de la muestra debe elegir el líder sindical para confirmar de manera estadística su afirmación?
4. El número de años de conducir de cierto grupo de camioneros tiene  $\sigma = 6$  años.
- a) ¿Cuál es la probabilidad de que una muestra aleatoria de 150 de estos camioneros produzca una media muestral que se desvíe de su media poblacional en menos de 0.75 años?
- b) Calcule la probabilidad de que la media muestral de 150 camioneros sea más pequeña que la poblacional en un año.
- c) Se desea confirmar estadísticamente que, al tomar una muestra aleatoria de camioneros, ellos tendrán en promedio menos de un año más de experiencia que la media poblacional en más de 95% de los casos. ¿Qué tamaño mínimo de la muestra debe elegir para la confirmación estadística?
5. Una planta industrial fabrica bombillas de luz cuya duración es una variable aleatoria con una media de 780 horas y una desviación estándar de 50 horas.
- a) Calcule la probabilidad de que al seleccionar una muestra aleatoria de 60 focos tengan una duración promedio mayor a 800 horas.
- b) ¿Cuál es el tamaño mínimo de muestra que debe seleccionar para que con una probabilidad máxima de 0.01 la media muestral sea menor a 770 horas?
6. Se ha hecho un estudio en la Ciudad de México acerca de las ganancias de los taxistas. Se estima que su ganancia media es de \$850 diarios, con una desviación estándar de \$200. Si fueran ciertas estas estimaciones y el vehículo perteneciera al chofer.
- a) ¿Cuál sería la probabilidad de que el chofer gane en promedio en medio año (25 días de trabajo al mes) entre \$800 y \$900 diarios?
- b) ¿Cuál es el tamaño mínimo de muestra que debe elegirse para que con una probabilidad mínima de 0.90 la ganancia media de la muestra se encuentre entre \$800 y \$900 diarios?
7. Se sabe que el tiempo promedio de reacción a un estímulo auditivo es una variable aleatoria con  $\mu = 0.15$  y  $\sigma = 0.04$  s para personas que oyen con normalidad.

- a) ¿Qué tamaño mínimo de la muestra se tomará si se requiere al menos 95% de seguridad para que el tiempo medio de reacción muestral sea menor de 0.153 segundos?
- b) ¿Cuál es la probabilidad de que el tiempo promedio de reacción de 90 personas con oído normal sea mayor a 0.16?
8. Una máquina envasa sal en bolsas con  $X$  kg, después se introducen en cajas que contienen 100 bolsas cada una, si  $X$  es una variable aleatoria con una media de 0.5 kg y desviación estándar 0.05 kg.
- a) ¿Qué porcentajes de cajas llenas se espera que pesen entre 50.5 y 51 kg?
- b) El productor pierde \$20 por cada caja que tenga más de 50.5 kg. ¿Cuánto se espera que pierda en el siguiente embarque de 10000 cajas?
9. Suponga que los eslabones para cadenas de bicicletas tienen longitudes distribuidas alrededor de la media de  $\mu = 0.50$  cm y  $\sigma = 0.04$  cm. Los modelos del fabricante requieren cadenas de 100 eslabones que posean una longitud entre 49 y 50 cm, ¿qué proporción de estas satisfacen los requisitos del fabricante?
10. El tiempo tomado por una persona para llenar un formato de empleo es de 8 minutos con una desviación estándar de 2.5 minutos. Si a las oficinas llegan 50 personas para llenar la solicitud de empleo, ¿cuál es la probabilidad de que la suma de sus tiempos de llenado de las 50 personas no sobrepase siete horas?

## 2.10 Teorema central del límite para diferencia de medias

Otra aplicación del teorema central del límite se refiere a la diferencia de medias, para lo cual se aplica que  $E\{\bar{X} - \bar{Y}\} = \mu_1 - \mu_2$  y  $V\{\bar{X} - \bar{Y}\} = \sigma_1^2/n_1 + \sigma_2^2/n_2$ .

### Teorema 2.12

#### Teorema central del límite para diferencia de medias

Sean  $X_1, X_2, \dots, X_{n_1}$  y  $Y_1, Y_2, \dots, Y_{n_2}$  dos muestras aleatorias independientes de dos poblaciones con  $n_1$  observaciones de la población 1 y  $n_2$  observaciones de la población 2, por otro lado sean  $(\mu_1, \sigma_1^2)$  y  $(\mu_2, \sigma_2^2)$  las medias y varianzas finitas de las poblaciones 1 y 2, respectivamente, entonces la forma límite de la distribución de la variable:

$$Z = \frac{(\bar{X} - \bar{Y}) - \mu_{\bar{X} - \bar{Y}}}{\sigma_{\bar{X} - \bar{Y}}} = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Tiene una distribución normal estándar cuando  $n_1$  y  $n_2$  se hacen infinitas.

En los siguientes ejemplos se muestra una aplicación del teorema 2.12.

### Ejemplos 2.17 Teorema central del límite para diferencia de medias

1. Las calificaciones de los exámenes de admisión a licenciatura en la Ciudad de México tienen una media de 64 con una varianza de 200, pero se desconoce su distribución. De igual manera, las calificaciones de los aspirantes a ingresar a licenciatura en Monterrey tienen una media de 60 con una varianza de 100. Calcule la probabilidad de que al tomar dos muestras aleatorias de tamaños 100 y 50 de aspirantes a licenciatura de la Ciudad de México y Monterrey, respectivamente, el promedio de los aspirantes de Monterrey sea mayor a los de la Ciudad de México.

**Solución**

Sean  $X_1, X_2, \dots, X_{100}$  las variables aleatorias que representan las calificaciones de los 100 aspirantes a ingresar a licenciatura en la Ciudad de México; de la misma forma, sean  $Y_1, \dots, Y_{50}$  las variables aleatorias que representan las calificaciones de los 50 aspirantes a ingresar a licenciatura en Monterrey; así, se conoce:

$$\mu_D = 64, \sigma_D^2 = 200 \text{ y } n_D = 100, \mu_M = 60, \sigma_M^2 = 100 \text{ y } n_M = 50$$

La probabilidad pedida es:

$$P(\bar{Y} > \bar{X}) = P(\bar{Y} - \bar{X} > 0)$$

Puesto que no se conoce la distribución de los datos y las muestras son grandes, podemos utilizar una aproximación con el teorema central del límite:

$$P(\bar{Y} - \bar{X} > 0) = P\left(\frac{(\bar{Y} - \bar{X}) - (\mu_M - \mu_D)}{\sqrt{\frac{\sigma_M^2}{n_M} + \frac{\sigma_D^2}{n_D}}} > \frac{0 - (60 - 64)}{\sqrt{\frac{100}{50} + \frac{200}{100}}}\right) = P(Z > 2) \cong 0.0228$$

Como se puede apreciar, es poco probable que en muestras aleatorias de tamaños 50 y 100 para Monterrey y la Ciudad de México, respectivamente, el promedio de los aspirantes en Monterrey sea mayor a los de la Ciudad de México.

2. Si en las condiciones del problema anterior se toman muestras aleatorias del mismo tamaño,  $n$ , calcule cuál debe ser el tamaño mínimo de  $n$  para que la probabilidad de que el promedio de los aspirantes a licenciatura en Monterrey sea mayor a la de los aspirantes de la Ciudad de México y sea menor a 0.05.

**Solución**

En las condiciones anteriores, se tiene  $P(\bar{Y} - \bar{X} > 0) < 0.05$ .

Luego,

$$P(\bar{Y} - \bar{X} > 0) = P\left(\frac{(\bar{Y} - \bar{X}) - (\mu_M - \mu_D)}{\sqrt{\frac{\sigma_M^2}{n} + \frac{\sigma_D^2}{n}}} > \frac{0 - (60 - 64)}{\sqrt{\frac{100}{n} + \frac{200}{n}}}\right) = P\left(Z > \frac{4}{\sqrt{300}}\sqrt{n}\right) < 0.05$$

Si se utilizan las tablas porcentuales de la distribución normal resulta  $P(Z > 1.6449) < 0.05$ .

Luego,

$$\frac{4}{\sqrt{300}}\sqrt{n} \geq 1.6449, \text{ al final } n \geq \left(\frac{1.6449\sqrt{300}}{4}\right)^2 = 50.73$$

Pero  $n$  natural, entonces  $n \geq 51$ .

Para resolver el problema también se pueden utilizar las fórmulas de la sección diferencia de medias de distribuciones normales:

$$n = \left\lceil \left[ \frac{\sqrt{\sigma_1^2 + \sigma_2^2} \Phi^{-1}(1 - \alpha)}{D_0 - (\mu_1 - \mu_2)} \right]^2 \right\rceil + 1 = \left\lceil \left[ \frac{\sqrt{100 + 200} \Phi^{-1}(0.95)}{0 - (60 - 64)} \right]^2 \right\rceil + 1 = 51$$

## Ejercicios 2.9

### Teorema central del límite para la diferencia de medias

1. Los empleados de gobierno suelen pedir días de incapacidad en un año, lo que ocasiona problemas en los servicios que atienden. Después de hacer una revisión de los historiales sobre incapacidades de los empleados, se notó que es posible dividirlos en dos clases, lo que depende de la antigüedad. Una clase son los empleados que tienen menos de 10 años de servicio y la otra los que tienen 10 o más. Si se obtiene para los de menos de 10 años una media de 8.8 días con una varianza de 36, en la otra clase fue la media de 9.6 días con una varianza de 80.
  - a) Para validar estos resultados se decide considerar dos muestras aleatorias independientes de 100 empleados cada una, si la probabilidad de que la diferencia entre las medias muestrales sea distinta de la diferencia de las medias poblacional por más de dos días de incapacidad es menor a 0.20, entonces se validan los resultados. Pero, ¿serán validados los resultados?
  - b) Con base en las muestras anteriores, ¿cuál es la probabilidad de que en promedio la muestra de la clase, 10 años o más pida menos de 1.5 días más que la muestra de la otra clase al año?
  - c) ¿Cuál será el tamaño mínimo de las muestras que deberá elegirse para que en promedio la muestra de la clase 10 años o más pida menos de 1.5 días más que la muestra de la otra clase en al menos 95% de los casos al año?
2. Una compañía farmacéutica lanzó un producto  $A$  para bajar de peso similar al producto existente  $B$ . En la farmacéutica se han hecho los estudios correspondientes entre una población de personas con características de edad y estatura similares; a una mitad se administró el producto  $A$  y a la mitad  $B$  la otra mitad. El producto  $A$  tuvo una media de 4.5 kg con una desviación estándar de 1.25 kg de reducción de peso en un mes, mientras que el producto  $B$  ayudó a reducir el peso mensual con una media de 4 kg y desviación estándar de 1.5 kg.
  - a) El gerente de mercadotecnia de la farmacéutica asegura que el producto  $A$  rebaja en un mes en promedio más de 0.25 kg que el producto  $B$  en más de 95% de las veces. Para probar de forma estadística su afirmación, el gerente elige dos muestras aleatorias independientes de tamaño 60 de cada población estudiada y realiza la comparación de diferencias muestrales. Explique por qué en estas condiciones no resultará válida la afirmación del gerente, justifique su respuesta.
  - b) En el inciso anterior, ¿cuál sería el tamaño mínimo de las muestras que debería de haber considerado el gerente para que su afirmación sea válida estadísticamente?
  - c) Calcule la probabilidad de que la diferencia de medias muestrales se desvíe en menos de 0.25 kg de la verdadera diferencia poblacional. Considere las condiciones del inciso a).
3. Sea el puntaje obtenido por dos poblaciones de alumnos de una secundaria de una prueba de rendimiento en física. Suponga que los puntajes de las dos poblaciones están distribuidos con medias  $\mu_1 = 50$ ,  $\mu_2 = 40$  y varianzas,  $\sigma_1^2 = 25$  y  $\sigma_2^2 = 49$ . Se consideran muestras aleatorias independientes de 40 y 45 alumnos de la primera y segunda población.
  - a) ¿Cuál es la probabilidad de que la diferencia entre la media muestral 1 menos la 2 esté comprendida entre siete y 14 puntos inclusive?
  - b) ¿Cuál debe ser el tamaño mínimo de las muestras para que la probabilidad de que el promedio de alumnos de la muestra 1 tenga una ventaja máxima de 9 puntos sobre el promedio de los alumnos de la muestra 2, sea menor a 10%?
4. Se diseña un experimento para probar quién de los operarios  $A$  o  $B$  obtiene el trabajo para manejar una nueva máquina. Se toma el tiempo de 100 pruebas que involucran la realización de cierto trabajo en la máquina para cada operario. Si la diferencia de los promedios de las muestras para las 100 pruebas difieren en menos de medio minuto de la diferencia de las medias poblacionales, se considera que el experimento termina en empate; en caso contrario, el operario con la media más pequeña obtiene el trabajo. Se supone que las varianzas de los tiempos para cada operario son de  $3.5 \text{ m}^2$  y  $3 \text{ m}^2$ .

- a) ¿Cuál es la probabilidad de un empate?
- b) ¿Cuál es el tamaño de muestra mínimo que se requiere para que la probabilidad de empate sea al menos de 0.98?
5. Según un estudio de renta-habitación de departamentos en la Ciudad de México, en la delegación Iztapalapa,  $A$ , el costo medio es de \$3 500, con una desviación estándar de \$500, y en la delegación Iztacalco,  $B$ , el costo medio es de \$3 900, con una desviación estándar de \$600. Suponga válido el estudio y seleccione dos muestras aleatorias e independientes una en cada área de tamaños 70 y 60 departamentos, respectivamente.
- a) Calcule la probabilidad de que la diferencia de medias  $\bar{X}_B - \bar{X}_A$  esté entre 100 y 500.
- b) Calcule la probabilidad de que la diferencia de medias  $\bar{X}_B - \bar{X}_A$  sea menor a 250.
- c) ¿Al menos cuántos departamentos se tendrían que muestrear para que  $\bar{X}_B - \bar{X}_A$  sea al menos \$300 en más de 95% de los casos?

## 2.11 Teorema central del límite para proporciones

Otra aplicación del teorema central del límite se refiere a las variables aleatorias con distribución de Bernoulli o las proporciones; para esto recordemos que  $E(\hat{p}) = p$  y  $V(\hat{p}) = p(1 - p)/n$ .

### Teorema 2.13

#### Teorema central del límite para proporciones

Sean  $X_1, X_2, \dots, X_n$  una muestra aleatoria de Bernoulli con probabilidad de éxito  $p$ ; por otro lado, sea la proporción de éxitos de la muestra  $\hat{p} = \bar{X}$ . Entonces, la variable:

$$Z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

tiene una distribución normal estándar cuando  $n$  se hace infinita.

En el siguiente ejemplo se muestra una aplicación del teorema 2.13.

### Ejemplo 2.18 Teorema central del límite para proporciones

La confiabilidad de un fusible eléctrico corresponde a la probabilidad de que uno de éstos, seleccionado al azar de la línea de producción, funcione de manera adecuada bajo las condiciones de diseño. Si se sabe que su confiabilidad es de 98%, calcule la probabilidad de que en otra muestra de 1 000 se contengan al menos 27 defectuosos.

#### Solución

Sean  $X_1, X_2, \dots, X_{1000}$  las variables aleatorias que representan el funcionamiento de un fusible (correcto e incorrecto), de manera que la proporción de fusibles que trabajan de manera correcta es  $\hat{p} = \bar{X}$ . Por otro lado, debido a que las variables se refieren a fusibles que funcionan bien, la condición: *al menos 27 de 1 000 fusibles sean defectuosos*, es equivalente a decir que: *a lo más 973 de 1 000 fusibles son buenos*. De tal forma que la probabilidad que se pide es:

$$P\left(\bar{X} \leq \frac{973}{1000}\right)$$

Luego, del teorema 2.13 con  $p = 0.98$  y  $\hat{p} = \bar{X} = \frac{973}{1000} = 0.973$ , tenemos que:

$$P\left(\bar{X} \leq \frac{973}{1000}\right) = P\left(\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq \frac{0.973 - 0.98}{\sqrt{\frac{0.98(1-0.98)}{1000}}}\right) = P(Z \leq -1.58) \cong 0.0571$$

## Teorema central del límite para diferencia de proporciones

Otra aplicación del teorema central del límite se refiere a la diferencia de proporciones, para esto recordamos que:

$$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2 \text{ y } V(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}.$$

### Teorema 2.14

#### Teorema central del límite para diferencia de proporciones

Sean  $X_1, X_2, \dots, X_{n_1}$  y  $Y_1, Y_2, \dots, Y_{n_2}$  dos muestras aleatorias independientes tomadas de dos poblaciones con distribución de Bernoulli y parámetros  $p_1$  y  $p_2$ ; por otro lado,  $\hat{p}_1 = \bar{X}$  y  $\hat{p}_2 = \bar{Y}$  son las proporciones respectivas; entonces, la forma límite de la distribución de la variable:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - \mu_{\hat{p}_1 - \hat{p}_2}}{\sigma_{\hat{p}_1 - \hat{p}_2}} = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

tiene una distribución normal estándar cuando  $n_1$  y  $n_2$  se hacen infinitas.

En el siguiente ejemplo se muestra una aplicación del teorema 2.14.

### Ejemplo 2.19 Teorema central del límite para diferencia de proporciones

Con respecto a las elecciones para presidente en México de 2012, se tomó una muestra aleatoria de 200 personas y resultó que 60 de ellas estuvieron a favor del candidato Andrés Manuel López Obrador, mientras que, en una segunda muestra aleatoria de 300 personas, resultó que 80 estaban a favor del candidato Enrique Peña Nieto. Al suponer que los porcentajes de simpatizantes para López Obrador y Peña Nieto son 36% y 22%, calcule en qué porcentaje de muestras independientes de 200 y 300 personas, se cumple que  $\hat{p}_L - \hat{p}_S > 0.20$ .

#### Solución

Sean  $X_1, X_2, \dots, X_{200}$  las variables aleatorias que representan a las personas entrevistadas para saber si son simpatizantes de López Obrador y  $Y_1, \dots, Y_{300}$  las variables aleatorias que representan las personas entrevistadas para saber si son simpatizantes de Peña Nieto. De tal forma que las proporciones muestrales de simpatizantes de López Obrador y Peña Nieto son  $\hat{p}_L = \bar{X}$  y  $\hat{p}_S = \bar{Y}$ . Así, la probabilidad que se pide es  $P(\hat{p}_L - \hat{p}_S > 0.20)$ .

Puesto que las muestras son grandes se emplea el teorema 2.14, para la diferencia de proporciones, con  $\hat{p}_L$  y  $\hat{p}_S$ , de manera que:

$$\begin{aligned}
 P(\hat{p}_L - \hat{p}_S > 0.20) &= P\left(\frac{(\hat{p}_L - \hat{p}_S) - (p_L - p_S)}{\sqrt{\frac{p_L(1-p_L)}{n_L} + \frac{p_S(1-p_S)}{n_S}}} > \frac{0.20 - (0.36 - 0.22)}{\sqrt{\frac{0.36(0.64)}{200} + \frac{0.22(0.78)}{300}}}\right) \\
 &= P(Z > 1.445) \\
 &= 0.0742
 \end{aligned}$$

En 7.42% de los casos de muestras independientes de tamaño 200 y 300, respectivamente para López Obrador y Peña Nieto, resulta  $\hat{p}_L - \hat{p}_S > 0.20$ .

## Cálculo del tamaño mínimo de muestra para proporciones de muestras grandes

En el caso de proporciones, el tamaño mínimo de muestra que cumpla ciertas condiciones se puede determinar mediante la aproximación de la normal, pues las proporciones son promedios. Entonces:

$$P(\hat{p} \leq p_0) < \alpha \Rightarrow P\left(\frac{\hat{p} - E(\hat{p})}{\sqrt{V(\hat{p})}} \leq \frac{p_0 - p}{\sqrt{p(1-p)/n}}\right) \cong P\left(Z \leq \sqrt{n} \frac{p_0 - p}{\sqrt{p(1-p)}}\right) < \alpha$$

Para los casos:

- a)  $P(\hat{p} \leq p_0) < \alpha$ , condición  $p_0 < p$  y  $\alpha \leq 0.50$
- b)  $P(\hat{p} > p_0) > \alpha$ , condición  $p_0 < p$  y  $\alpha > 0.50$
- c)  $P(\hat{p} \leq p_0) > \alpha$ , condición  $p_0 > p$  y  $\alpha > 0.50$
- d)  $P(\hat{p} > p_0) < \alpha$ , condición  $p_0 > p$  y  $\alpha \leq 0.50$

El resultado a) con b) son complementos.  
El resultado c) con d) son complementos.

Se utiliza el tamaño mínimo de muestra para  $[q]$  parte entera de  $q$ :

$$n \geq p(1-p) \left(\frac{\Phi^{-1}(\alpha)}{p_0 - p}\right)^2 \text{ o } n = \left\lceil p(1-p) \left(\frac{\Phi^{-1}(\alpha)}{p_0 - p}\right)^2 \right\rceil + 1$$

- e)  $P(|\hat{p} - p| < p_0) > \alpha$  condición  $\alpha > 0.50$  o su complemento:

$$n \geq p(1-p) \left(\frac{\Phi^{-1}((1-\alpha)/2)}{p_0}\right)^2 \text{ y } n = \left\lceil p(1-p) \left(\frac{\Phi^{-1}((1-\alpha)/2)}{p_0}\right)^2 \right\rceil + 1$$

De manera similar, para la diferencia de proporciones, con la consideración de que las muestras deben ser de tamaños iguales, se obtiene el de la muestra mínimo para los casos:

- a)  $P(\hat{p}_1 - \hat{p}_2 \leq p_0) < \alpha$ , condición  $p_0 < p_1 - p_2$  y  $\alpha \leq 0.5$
- b)  $P(\hat{p}_1 - \hat{p}_2 > p_0) > \alpha$ , condición  $p_0 < p_1 - p_2$  y  $\alpha > 0.5$
- c)  $P(\hat{p}_1 - \hat{p}_2 \leq p_0) > \alpha$ , condición  $p_0 > p_1 - p_2$  y  $\alpha > 0.5$
- d)  $P(\hat{p}_1 - \hat{p}_2 \geq p_0) < \alpha$ , condición  $p_0 > p_1 - p_2$  y  $\alpha \leq 0.5$

Se utiliza el tamaño mínimo de muestra para  $[q]$  parte entera de  $q$ :

$$n_1 = n_2 = n = \left\lceil (p_1(1-p_1) + p_2(1-p_2)) \left(\frac{\Phi^{-1}(\alpha)}{p_0 - (p_1 - p_2)}\right)^2 \right\rceil + 1$$



e)  $P(|\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)| \leq p_0) > \alpha$  condición  $\alpha > 0.50$  o su complemento:

$$n_1 = n_2 = n = \left[ (p_1(1-p_1) + p_2(1-p_2)) \left( \frac{\Phi^{-1}((1-\alpha)/2)}{D_0} \right)^2 \right] + 1$$

A continuación, se presentan dos ejemplos que muestran el uso de estos resultados.

### Ejemplo 2.20 Cálculo del tamaño mínimo de muestra

Conforme las condiciones del ejemplo, ¿cuál es el tamaño mínimo de la muestra que debe elegirse para que la proporción muestral de fusibles buenos sea mayor a 0.973, con una probabilidad mayor a 95%. Considere que se conserva la confiabilidad de 98% de los fusibles en toda la población.

#### Solución

Tenemos los valores  $p_0 = 0.973$ ,  $\alpha = 0.95$  y  $p = 0.98$ ; por tanto, el caso b) se ajusta al problema para el tamaño de muestra, con  $P(\hat{p} > p_0) > \alpha$ , condición  $p_0 < p$  y  $\alpha > 0.5$  y la fórmula:

$$n = \left[ p(1-p) \left( \frac{\Phi^{-1}(1-\alpha)}{p_0 - p} \right)^2 \right] + 1 = \left[ 0.98(0.02) \left( \frac{\Phi^{-1}(0.05)}{0.973 - 0.980} \right)^2 \right] + 1 = 1083$$

Es decir, el tamaño mínimo de la muestra es  $n = 1083$  para  $P(\hat{p} \geq 0.973 | p = 0.98) > 0.95$ .

### Ejemplo 2.21 Tamaño mínimo de la muestra

Conforme a las condiciones del ejemplo 2.19, ¿cuál es el tamaño mínimo de la muestra que debe elegirse para que la diferencia de proporciones muestrales de los candidatos  $\hat{p}_L - \hat{p}_S$  sea menor a 0.20, con una probabilidad mayor a 95%. Al suponer que los porcentajes de simpatizantes para los candidatos López Obrador y Peña Nieto son 36% y 22%, respectivamente.

#### Solución

Tenemos los valores  $p_0 = 0.20$ ,  $\alpha = 0.95$ ,  $p_L = 0.36$  y  $p_S = 0.22$ ; por tanto, el caso c) se ajusta al problema para el tamaño de muestra, con  $P(\hat{p}_L - \hat{p}_S \leq p_0) > \alpha$ , para  $p_0 > p_L - p_S$  y  $\alpha > 0.5$  y la fórmula:

$$n = \left[ (p_1(1-p_1) + p_2(1-p_2)) \left( \frac{\Phi^{-1}(\alpha)}{p_0 - (p_1 - p_2)} \right)^2 \right] + 1$$

$$= \left[ (0.36(0.64) + 0.22(0.78)) \left( \frac{\Phi^{-1}(0.95)}{0.2 - (0.36 - 0.22)} \right)^2 \right] + 1 = 303$$

El tamaño mínimo de la muestra es  $n = 303$  para que  $P(\hat{p}_L - \hat{p}_S < 0.2 | p_L = 0.36, p_S = 0.22) > 0.95$ .

## Ejercicios 2.10

### Teorema central del límite para proporciones y diferencia de proporciones

- Las autoridades de la Ciudad de México realizaron un estudio sobre robos en esa ciudad, el cual afirma que 15% de la población ha sufrido algún tipo de robo.
  - Se selecciona una muestra aleatoria de 100 ciudadanos de la Ciudad de México. ¿Cuál es la probabilidad de que entre 10 y 20% de personas haya sufrido algún tipo de robo?

- b) Un representante del gobierno de la ciudad afirma que la proporción muestral de robos es menor a 17% con una probabilidad mayor a 0.95. ¿Cuál tendría que ser el tamaño mínimo de la muestra que deba elegir el representante para que su afirmación resulte válida estadísticamente?
2. En cierto proceso industrial, la proporción de artículos defectuosos es de 5%. El gerente de calidad asegura que la proporción muestral de defectuosos se desvía de la proporción poblacional en menos de 3% con una probabilidad mayor a 0.90.
- a) ¿Será válida la afirmación del gerente si toma una muestra  $n = 120$ ? Justifique su respuesta.
- b) ¿Cuál debe ser el tamaño mínimo de la muestra aleatoria de artículos que se debe seleccionar de la producción, para que la proporción de artículos defectuosos en la muestra sea menor de 6% con una probabilidad mayor a 0.98?
3. En la Ciudad de México y área metropolitana se ha introducido un nuevo refresco de cola; según los estudios, este tiene 20% de consumidores de la población de esta zona geográfica. El gerente de mercadotecnia de la empresa refresquera afirma que la proporción de bebedores de una muestra aleatoria de 150 ciudadanos de dicha área es mayor a 15% con una probabilidad mayor a 0.98.
- a) ¿Es válida la afirmación? Justifique su respuesta.
- b) Calcule el tamaño mínimo de la muestra que hace válida la afirmación del gerente.
4. Debido al alto índice delictivo en la Ciudad de México se hizo un estudio referente a las tiendas pequeñas que podrían aceptar tarjetas de débito en el pago de los consumidores, de lo cual se encontró que 58% lo aceptarían. Calcule la probabilidad de que en una muestra aleatoria de 80 tiendas:
- a) 48 o más estén dispuestas a aceptar tarjetas de débito.
- b) entre 44 y 60 tiendas estén dispuestas a aceptar las tarjetas de débito.
- c) Los bancos interesados en contribuir con los servicios de las tarjetas aseguran que en estas muestras menos de 43 tiendas estarían de acuerdo con una probabilidad menor a 5%. ¿Cuál es el tamaño mínimo que tendrían que elegir de muestra para corroborar esta afirmación?
5. Suponga que se sabe que en cierta población de mujeres, 90% de las que cursan su tercer trimestre de embarazo han tenido algún cuidado prenatal. Si se selecciona una muestra aleatoria de 200 mujeres de esta población.
- a) Calcule la probabilidad de que la proporción de la muestra se diferencie de la proporción poblacional máximo en 5%.
- b) Calcule la probabilidad de que la proporción de la muestra se diferencie de la proporción poblacional al menos en 2%.
6. En un proceso de llenado de botellas de soda, 8% del total no se llenan por completo. Un comprador potencial del producto decide rechazar un gran lote de compra, si al seccionar una muestra de 225 botellas el porcentaje de botellas que no están llenas por completo es mayor a 7% con una probabilidad superior a 85%.
- a) ¿En estas condiciones comprarán el lote? Justifique su respuesta.
- b) ¿Cuál es el tamaño mínimo de la muestra para que no se compre el lote?
7. Los propietarios de una empresa que fabrica baterías para linterna han concluido, con datos históricos, que 96% de las baterías que producen resultan buenas. La condición para pasar la certificación de calidad de las baterías consiste en elegir una muestra aleatoria de 300 baterías y si al menos 95% resultan buenas con una probabilidad mínima de 0.90, entonces se dice que la producción queda certificada y que está dentro de los estándares de producción.
- a) ¿En estas condiciones la producción de baterías estará dentro de los estándares de producción?
- b) Determine el tamaño mínimo de la muestra que se requiere para que estadísticamente las baterías queden dentro de los estándares de producción.

8. Un fabricante de insecticidas en presentación de aerosol desea comparar dos nuevos productos. En el experimento se emplean dos habitaciones del mismo tamaño, cada una con una muestra de 1 000 moscas. En una habitación se rocía el insecticida  $A$  y en la otra el insecticida  $B$  en igual cantidad, se supone que los insecticidas son efectivos en 85 y 76%, respectivamente.
- ¿Cuál es la probabilidad de que la proporción de moscas muertas de la muestra con el insecticida  $A$  sea mayor a la proporción de las moscas muertas en la muestra con el insecticida  $B$  al menos en 14%?
  - Encuentre el tamaño mínimo de muestra que se tiene que considerar para que la probabilidad en el inciso  $a$ ) sea menor a 0.001.
9. Suponga que en un estudio acerca del uso de sistemas de información computarizados se observa que 62.5% de los gobiernos de las ciudades y 48.1% de los gobiernos estatales usan estos sistemas para el manejo de registro de personal.
- Calcule la probabilidad de que al seleccionar muestras independientes de tamaño 200, la diferencia en las proporciones muestrales de gobiernos de las ciudades y de los estados que los usan sea menor a 10%.
  - Encuentre el tamaño mínimo de muestra que se tiene que considerar para que la probabilidad en el inciso  $a$ ) sea al menos de 0.90.
10. Dos productores de pasta dental,  $A$  y  $B$ , afirman que 50 y 30% de la población, respectivamente, utilizan su pasta dental. Se consideran dos muestras independientes de la población de tamaño 300 cada una; a las personas de la muestra 1 se les pregunta si utilizan la marca  $A$  como pasta dental y a la muestra 2 si utilizan la marca  $B$  como pasta dental.
- Si en la muestra 1, 140 personas contestan de manera afirmativa y 104 en la 2, calcule la probabilidad de que la diferencia de proporciones muestrales sea más grande que la diferencia de estas dos muestras. ¿Corrobora este resultado la diferencia entre las proporciones que afirman los productores?
  - Si se supone que se conservan las proporciones muestrales de  $a$ ), ¿qué tamaño mínimo de muestra debió elegirse para que la probabilidad del inciso  $a$ ) sea mayor a 0.95?
  - ¿Se pueden resolver con el mismo método los incisos  $a$ ) y  $b$ ), si en lugar de considerar dos muestras se estudia una sola y la pregunta se formula como: qué tipo de pasta de dientes usa  $A$  o  $B$  u otra? Explique su respuesta.
11. Los medicamentos genéricos han tenido gran auge en los últimos años por su bajo costo y alta efectividad. Para probarlo, se comparan dos medicamentos semejantes, uno estándar y el otro genérico, cuyos fabricantes afirman que son 80% efectivos en un lapso de tres días en los casos en que se usó. Se consideran dos muestras aleatorias independientes de personas con una misma enfermedad; a la muestra 1, de tamaño 80, se le aplica la medicina estándar, mientras que a los enfermos de la muestra 2, de tamaño 100, se le aplica el medicamento genérico. Calcule la probabilidad de que las proporciones muestrales se desvíen máximo 10%.

## Teorema central del límite para distribuciones discretas

El teorema central del límite se aplica para cualquier tipo de distribución, ya sea continua o discreta, pero en el caso de las variables aleatorias discretas se puede mejorar la aproximación al hacer una corrección, la cual dependerá de los valores de la variable. Por ejemplo, en el caso de que los valores discretos de la variable sean solo enteros consecutivos, la corrección que se hace es de 0.5. La corrección se lleva a cabo debido a que la variable aleatoria es discreta y la aproximación se hace con la distribución normal que es continua. Por tanto, debemos hacer algunos *ajustes* al emplear la distribución normal.

La probabilidad en un punto  $k$  de la distribución discreta es igual al área del rectángulo con base 1 y altura  $P(X = k)$ . Luego, el área del rectángulo la podemos aproximar al área bajo la curva de la distribución normal, desde  $k - 0.5$  hasta  $k + 0.5$ , como se puede apreciar en la figura 2.23.

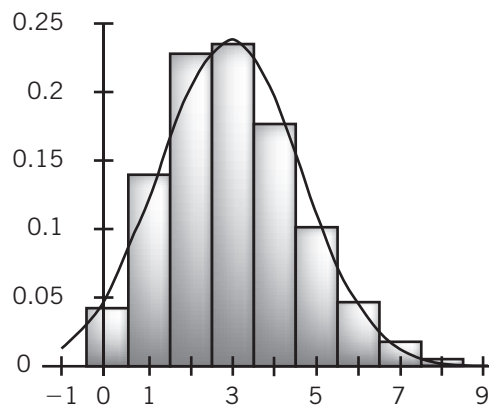


Figura 2.23 Aproximación de la normal a una distribución discreta.

Revisemos un par de ejemplos para el teorema central del límite caso discreto.

### Ejemplo 2.22 Distribuciones discretas

1. Sea  $X_1, X_2, \dots, X_{100}$  una muestra aleatoria de variables con distribución Bernoulli y parámetro  $p = 0.3$  y la estadística  $T = X_1 + X_2 + \dots + X_{100}$ , calcule  $P(T \geq 35)$  con y sin la corrección.

#### Solución

Sean las variables aleatorias  $X_1, \dots, X_{100}$  con distribución Bernoulli y parámetro  $p = 0.3$ , definimos a la estadística  $T = X_1 + X_2 + \dots + X_{100}$ .

a) Sin corrección. Como  $E(X_i) = p$  y  $V(X_i) = p(1 - p)$  para toda  $i$  desde 1 hasta 100, entonces:

$$Z = \frac{T - \mu_X}{\sigma_X} = \frac{T - n\mu}{\sqrt{n}\sigma} = \frac{T - np}{\sqrt{n}\sqrt{p(1-p)}}$$

Así:

$$P(T \geq 35) = P\left(Z \geq \frac{35 - 100(0.3)}{\sqrt{100}\sqrt{0.3(0.7)}}\right) = P(Z \geq 1.09) \cong \Phi(-1.09) = 0.1378$$

b) Con corrección. Con los resultados del inciso anterior:

$$P(T \geq 35) = P\left(Z \geq \frac{35 - 0.5 - 100(0.3)}{\sqrt{100}\sqrt{0.3(0.7)}}\right) = P(Z \geq 0.98) \cong \Phi(-0.98) = 0.1635$$

Por otro lado, si calculamos el valor de la distribución binomial con  $n = 100$  y  $p = 0.3$  con ayuda de algún software resulta  $P(T \geq 35) = 0.162858$ . Luego, el menor error en la aproximación lo da el resultado con la corrección 0.1635 contra el resultado sin la corrección de 0.1378.

2. Suponga que llegan los clientes al centro comercial entre las 10 y 12 a.m. con una razón de ocho personas cada 15 minutos y que tienen una distribución de Poisson. Resuélvalo con el uso de la corrección y compare los resultados.

#### Solución

Si se lleva a cabo el mismo desarrollo del ejemplo anterior, tenemos:

$$P\left\{\sum_{i=1}^{100} X_i > 900\right\} = P\left\{\sum_{i=1}^{100} X_i \geq 901\right\} = P\left\{Z \geq \frac{901 - 800}{\sqrt{100}\sqrt{8}}\right\} \cong 1 - \Phi\left\{\frac{901 - 0.5 - 800}{\sqrt{100}\sqrt{8}}\right\} = 0.0002$$

En general, se puede concluir que el teorema central del límite da buenas aproximaciones para el caso de las distribuciones discretas aun sin la corrección (esto se debe a que el tamaño de las muestras es grande), por lo que el uso de esta última en las aproximaciones queda como opción, según decida la persona que lleva a cabo el cálculo de las probabilidades.

## Distribuciones a las que no se puede aplicar el teorema central del límite

El teorema central del límite se aplica a cualquier distribución discreta o continua con la única condición de que la varianza sea finita. Por tanto, éste no aplica a distribuciones cuya varianza no sea finita. Por ejemplo:

- Distribución Cauchy,  $f(x; \alpha, \beta) = \frac{1}{\pi\beta \left[ 1 + \left( \frac{x - \alpha}{\beta} \right)^2 \right]}$ , con  $x, \alpha \in \mathbb{R}$ ,  $\beta > 0$ .
- Distribución Pareto con parámetro  $\theta$ ,  $f(x; \theta) = \frac{\theta x_0^\theta}{x^{\theta+1}} I_{(x_0, \infty)}(x)$ , con  $x_0 > 0$  y  $\theta \in (0, 2]$ . Si el parámetro toma valores mayores a 2; entonces, si se puede aplicar el teorema central del límite.

## Ejercicios de repaso

### Preguntas de autoevaluación

- Sea  $Z$  la variable de un modelo normal estándar, calcule (sin usar tablas ni calculadora) con cuatro dígitos exactos:
  - $P(Z > 7.23)$
  - $P(Z \leq -3\pi)$
  - $P(-8 < Z < 8)$
  - $P(0 < Z < 12.79)$
- ¿Será cierto que la distribución de la suma y promedio de distribuciones normales es otra normal?
- ¿Es correcto que la distribución de la suma y promedio de distribuciones uniformes es otra uniforme?
- ¿En qué se diferencian los teoremas para el cálculo de probabilidades de una suma de normales y el TCL?
- ¿A partir de qué tamaño de la muestra se aconseja utilizar el TCL?
- Si conocemos la distribución de una variable aleatoria, ¿por qué es necesario utilizar el TCL para el cálculo de probabilidades de la suma o promedio de una muestra aleatoria de estas variables?
- ¿Será verdad que una estadística que proviene de una muestra aleatoria de variables, siempre debe contener los mismos parámetros de éstas?
- ¿A qué se debe el factor de corrección en el TCL para variables aleatorias discretas?

### Ejercicios complementarios con grado de dificultad uno

- Para seleccionar a sus empleados un ejecutivo industrial usa una prueba que tiene una puntuación promedio de 140 puntos y una desviación estándar  $\sigma = 10$ . Suponga que la distribución de las puntuaciones es normal y que una puntuación mínima de 130 le permite al solicitante ser considerado. ¿Cuál es la probabilidad de que, bajo este criterio, el siguiente aspirante sea considerado?
- Una empresa metalúrgica produce rodamientos con un diámetro que tiene una desviación normal, con media de 3.0005 in y desviación estándar de 0.0010 in. Las especificaciones requieren que los diámetros estén en el intervalo de  $3.000 \pm 0.0020$  in. Si los cojinetes cuyos diámetros quedan fuera de ese intervalo se rechazan, ¿qué fracción de la producción total será descalificada?
- El tiempo de ausencia a clases de los niños de primaria en la Ciudad de México en el invierno tiene una distribución normal con media de 80 horas y desviación estándar de 20 horas. ¿Cuál es la probabilidad de que el tiempo de ausencia de los niños a la primaria en el siguiente invierno sea menor a 70 horas?
- El líquido despachado por una máquina de refrescos está distribuido normalmente, con una media de 230 mm y una desviación estándar de 10 mm. Calcule la probabilidad de que el siguiente vaso despachado tenga más de 250 mm.
- Ciertos tipos de baterías para automóvil tienen un tiempo de vida distribuido con media 1200 días y desviación estándar igual a 100 días. ¿Por cuánto tiempo se deben garantizar las baterías si el fabricante quiere reemplazar solo 20% de las baterías vendidas?
- Algunas baterías para automóvil tienen un tiempo de vida distribuido con media 1200 días y desviación estándar igual a 100 días. ¿Cuántas de las 3000 baterías que se venderán durarán más de 1300 días?
- En un aserradero se cortan árboles en trozos de 5 m en promedio con desviación estándar de 0.25 m, las longitudes se distribuyen en forma aproximadamente normal. Si se elige un lote de 200 trozos, ¿cuál será el número esperado de éstos que superen la longitud de 4.60 m?

- 2.17** Suponga que los resultados de los exámenes de Probabilidad y estadística I tienen una distribución aproximadamente normal con  $\mu = 4.5$  puntos y variancia de 1 punto.
- ¿Cuál es la probabilidad de que quien presente el examen obtenga una calificación mayor o igual a 6?
  - ¿Cuál debe ser la calificación mínima aprobatoria, si los profesores sinodales pretenden que solo aprueben 20% de los estudiantes que presenten examen?
- 2.18** Suponga que los resultados de los exámenes de Introducción a la Administración tienen una distribución aproximadamente normal con  $\mu = 7.2$  puntos y variancia de 1.8 puntos. ¿Cuántos de los 480 alumnos que van a presentar el examen de esta asignatura obtendrán una calificación menor a 6?
- 2.19** Se observó durante un largo periodo que la cantidad semanal gastada en el mantenimiento y las reparaciones de cierta fábrica tiene una distribución normal con  $\mu = \$400$  y  $\sigma = \$20$ . ¿De cuánto tendría que ser el presupuesto para reparaciones semanales y mantenimiento, para que la cantidad presupuestada solo sea rebasada con una probabilidad de 0.10?
- 2.20** Cierta proceso de manufactura produce pernos que deben tener un diámetro entre 1.2 y 1.25 in. Se sabe que el diámetro se distribuye normalmente con  $\mu = 1.21$  y  $\sigma = 0.02$ . ¿Qué porcentaje de los pernos está fuera de las especificaciones?
- 2.21** Al probarse a compresión simple los cilindros de concreto, se obtuvieron los resultados: en promedio resistieron  $240 \text{ kg/cm}^2$ , con una desviación estándar de  $30 \text{ kg/cm}^2$ . Suponga que la resistencia a la compresión tiene una distribución normal. ¿Cuál es la probabilidad de que otro cilindro tomado al azar:
- tenga una resistencia mayor de  $330 \text{ kg/cm}^2$ ?
  - su resistencia esté en el intervalo de 210 a  $240 \text{ kg/cm}^2$ ?
- 2.22** Una compañía paga a sus empleados un salario promedio de \$5.25 por hora con una desviación estándar de 50 centavos. Si los salarios tienen aproximadamente una distribución normal.
- ¿Qué porcentaje de los trabajadores recibe salario entre \$4.75 y \$5.69 por hora?
  - ¿Mayor de qué cantidad es 5% de los salarios más altos?
- 2.23** La cantidad promedio que gastó una empresa durante un año en servicios médicos por cada empleado fue de \$2 575 con una desviación estándar de \$325. Suponga una población normal.
- Calcule la probabilidad de que en una muestra aleatoria de 15 empleados la media muestral sea mayor a \$2800.
  - Si los representantes sindicales de la empresa afirman que menos de 15% de los trabajadores de la empresa gastan más de \$2480. ¿Cuál deberá ser el tamaño mínimo de la muestra que tomen los representantes sindicales cuando realicen el estudio estadístico para verificar su afirmación?
- 2.24** Se ha hecho un estudio en la Ciudad de México sobre las ganancias de los taxistas. Se estima que la ganancia media de los taxistas es de \$650 diarios con una desviación estándar de \$100 y distribución normal. Si fueran ciertas estas estimaciones y el vehículo pertenece al chofer:
- ¿Cuál sería la probabilidad de que el chofer gane al mes (25 días de trabajo) entre \$15 000 y \$17 000?
  - ¿Cuál sería la probabilidad de que el chofer gane al mes (25 días de trabajo) más de \$17 550?
- 2.25** Una compañía transnacional instituyó un programa de seguridad en el trabajo para reducir el tiempo perdido debido a accidentes de trabajo. La empresa interesada en comprar el programa de seguridad toma la decisión de hacer un comparativo entre los tiempos perdidos antes y después de implantar el programa y establece que si al comparar 10 semanas la probabilidad de reducir al menos 14 horas en promedio es mayor a 0.80, entonces comprará dicho programa. Suponga que la empresa interesada sabe que la distribución del tiempo perdido actual por semana tiene una distribución  $N(108, 12^2)$  mientras que la compañía asegura que la distribución semanal del tiempo perdido con la implementación del programa tiene una  $N(91, 14^2)$ . Para vender el programa de seguridad, el gerente de mercadotecnia de la compañía asegura a las empresas interesadas que su programa reduce los tiempos promedio perdidos en más de 14 horas a la semana en más de 90% de los casos. ¿Cuál debe ser el tamaño mínimo de semanas que debe seleccionar el gerente para corroborar su afirmación de manera estadística?
- 2.26** La variable de interés en una investigación es el puntaje obtenido por dos poblaciones de alumnos de último año de una prueba de rendimiento en matemáticas. Los investigadores suponen que los puntajes de las dos poblaciones estaban distribuidos normalmente con medias de:  $\mu_1 = 50$ ,  $\mu_2 = 40$  y varianzas:  $\sigma_1^2 = 25$  y  $\sigma_2^2 = 49$ . Una muestra aleatoria de 10 alumnos se selecciona de la primera población y otra de forma independiente a la primera de 10 alumnos de la segunda población. ¿Cuál es la probabilidad de que la diferencia entre las medias muestrales 1 menos la 2 esté comprendida entre cinco y 15 puntos?
- 2.27** Se diseña un experimento para probar cuál de los operarios A o B obtiene el trabajo para manejar una nueva máquina. Se toma el tiempo de 20 pruebas que involucran la realización de cierto trabajo en la máquina para cada operario. Si los promedios de las muestras para las 20 pruebas difieren en menos de un segundo se considera que el experimento termina en empate; en caso contrario, el operario con la media más pequeña obtiene el trabajo. Se supone que las varianzas de los tiempos para cada operario son de  $3.5 \text{ s}^2$ . Suponga que las poblaciones se distribuyen en forma normal. ¿Cuál debe ser el tamaño mínimo de la muestra, de pruebas que se consideren para que la probabilidad de encontrar un ganador sea menor a 0.25?
- 2.28** Los cinescopios para receptores de televisión de un fabricante A tienen una vida con distribución normal con media 6.5 años y una desviación estándar de 0.9 años; en tanto que la vida de los cinescopios de un fabricante B también se distribuyen en forma normal, pero con una media de 6.0 años y una desviación estándar de 0.8 años.



- a) ¿Cuál es la probabilidad de que en una muestra de 20 cinescopios del fabricante  $A$  tenga una vida promedio que sea al menos un año mayor que la vida promedio de una muestra de 26 cinescopios del fabricante  $B$ ?
- b) Calcule la probabilidad a) si los tamaños de muestra se consideran iguales, 20-20 y 26-26.
- c) El fabricante  $B$  afirma que en al menos 90% de los casos el promedio muestral de sus cinescopios duran más de un mes que el del fabricante  $B$ . ¿Cuál debe ser el tamaño mínimo de la muestra que debe considerar el fabricante  $B$  para corroborar su afirmación de manera estadística?
- 2.29** Se sabe que el medicamento estándar usado para tratar cierta enfermedad ha resultado efectivo en un lapso de tres días en 75% de los casos en que se empleó. Al evaluar la efectividad de un nuevo medicamento para tratar la misma enfermedad, se les administró a 10 pacientes que la padecían. ¿Cuál es la probabilidad de observar máximo a tres pacientes que se recuperen de la muestra? Considere que el nuevo medicamento es al menos tan efectivo como el estándar.
- 2.30** Un estudio llevado a cabo en un distrito de la ciudad dio como resultado que 70% de los trabajadores habían cambiado de empleo por lo menos una vez en su vida. Si se selecciona una muestra aleatoria de 20 trabajadores de este distrito:
- a) ¿Cuál es la probabilidad de que más de 14 han cambiado de empleo por lo menos una vez en su vida?
- b) ¿Cuál es la varianza para muestras aleatorias de tamaño 200 para los trabajadores que hayan cambiado de empleo al menos una vez en su vida?
- 2.31** En un proceso de llenado de botellas de soda, 10% no se llenan por completo. Si para este proceso se selecciona al azar una muestra de 30 botellas, responda:
- a) ¿Cuál es la probabilidad de que más de cuatro botellas de la muestra no estén completamente llenas?
- b) ¿Cuál es la probabilidad de que entre cinco y 25 botellas de la muestra no estén llenas en su totalidad? ¿Qué puede decir con respecto a la respuesta del inciso a)?
- 2.32** Cierta tipo de pistones se fabrica con un diámetro de 10 cm y una desviación estándar de 1 mm, se toman muestras de tamaño  $n$ , las cuales se mandan a una revisión total, cuando su diámetro promedio es menor a 9.9 cm.
- a) ¿Qué tamaño mínimo de la muestra debe elegirse para que los pistones tengan que mandarse a revisión total máximo en 1% de los casos?
- b) ¿Cuál es la probabilidad de que el diámetro promedio de 100 pistones sea mayor a 10.09?
- 2.33** Suponga que el peso de los paquetes de café de cierta marca tiene una media de 1.05 kg, y desviación estándar de 0.05 kg, si en una caja se colocan 50. Si se consideran faltantes de peso a las cajas en las cuales el peso total de los paquetes es inferior a 51.80 kg. ¿En cuántas de las 200 cajas del camión que las transporta habrá faltantes de peso?
- 2.34** Un vendedor de automóviles sospecha que su margen de beneficios promedio por auto vendido está por debajo del promedio nacional de \$2700. Se admite que el margen de beneficios por auto vendido tiene una desviación estándar de \$600. Una muestra aleatoria de 50 autos vendidos da un margen de beneficios  $\bar{X}$ .
- a) Calcule la probabilidad de que  $\bar{X}$  sea superior a \$2850.
- b) Calcule la probabilidad de que  $\bar{X}$  difiera de la media poblacional en menos de \$250.
- 2.35** Suponga que se estudia el rendimiento de las empresas WM y EK, para esto se toman dos muestras aleatorias independientes de tamaño 88 para cada una. Los rendimientos tienen distribución normal con desviación estándar de 0.0009 y 0.0014, respectivamente.
- a) Calcule la probabilidad de que la diferencia muestral de los rendimientos difiera de la verdadera diferencia en más de 0.0002.
- b) ¿Cuál es el tamaño de muestra mínimo que se requiere para que la probabilidad calculada en a) sea al menos de 0.90?
- 2.36** ¿Cuántas personas se tendrán que entrevistar por parte de un periódico para estimar la proporción de personas en contra de la pena de muerte, con un error máximo de 5% y probabilidad mayor a 95%? Suponga que la verdadera proporción que está a favor de la pena de muerte es de 52%.
- 2.37** Se piensa que dos medicamentos son igual de efectivos para reducir el nivel de ansiedad en ciertas personas perturbadas en sus emociones. La proporción de personas en que los medicamentos resultan ser efectivas es de 70%. En una muestra de 100 personas con este padecimiento se les administró el medicamento  $A$ , mientras que a otra muestra independiente de 150 personas se les aplicó el medicamento  $B$ . Calcule la probabilidad de que las proporciones muestrales se diferencien máximo en 8%.

## Ejercicios complementarios con grado de dificultad dos

- 2.38** El peso de ciertos paquetes de azúcar es una variable aleatoria normal con una media de 48 g y una desviación estándar de 5 g. ¿Cuál es el tamaño mínimo de la muestra para cubrir la necesidad de 1 kg con una probabilidad mayor a 0.90? *Sugerencia:* Obtenga una ecuación de segundo grado para  $n$ .
- 2.39** Para determinar la calidad de diferentes tipos de secadoras de cabello que se venden en el mercado, una empresa ha realizado un estudio intensivo de este tipo de producto, en el cual se reportó que la duración de los secadores de una marca determinada tiene una distribución normal con una media de 1200 horas y una desviación estándar de 100 horas. Si un comprador afirma que puede probar de manera estadística que menos de 20% de las secadoras del fabricante duran en promedio menos de 1180 horas, ¿cuál es el tamaño mínimo de la muestra que debe el comprador elegir para que pueda corroborar su afirmación?

- 2.40** Una planta industrial fabrica bombillas de luz cuya duración tiene una distribución normal con una media de 780 horas y una desviación estándar de 50 horas.
- Ante reclamaciones de los compradores el fabricante de bombillas indica que hará un estudio de calidad si la probabilidad de que una muestra aleatoria de 25 bombillas tenga una duración promedio entre 750 y 800 es menor a 0.95. ¿Tendrá que hacer el estudio?
  - Si el grupo de compradores afirma que puede probar de manera estadística que menos de 1% de las bombillas duran en promedio más 790 horas, cuál es el tamaño mínimo de la muestra que deben elegir los compradores para que puedan corroborar su afirmación.
- 2.41** Una máquina envasa sal en bolsas con  $X$  kg de peso, las cuales se introducen después en cajas que contienen 20 bolsas cada una. Si  $X$  es una variable aleatoria distribuida normalmente con media de 2 kg, y desviación estándar de 0.20 kg.
- ¿Qué porcentaje de cajas tendrá menos de 39 kg?
  - ¿Cuál es el tamaño mínimo de la muestra para cubrir la necesidad de 40 kg con una probabilidad mayor a 0.90? *Sugerencia:* Obtenga una ecuación de segundo grado para  $n$ .
- 2.42** Para determinar la calidad de diferentes tipos de secadoras de cabello que se venden en el mercado, una empresa ha realizado un estudio intensivo de este tipo de producto, en el cual se reportó que la duración de los secadores de una marca determinada tiene una media de 1 200 horas y una desviación estándar de 200 horas.
- Se elige una muestra aleatoria de tamaño 80, ¿cuál es la probabilidad de que el tiempo promedio de duración de la muestra esté entre 1 150 y 1 220 horas?
  - ¿Cuál es el tamaño mínimo de la muestra que debe seleccionarse para que con una probabilidad máxima de 0.10 la media muestral sea menor a 1 100 horas?
- 2.43** Los refrigeradores de un fabricante  $A$  tienen una vida media de 9.5 años con una desviación estándar de dos años; en tanto que la vida media de los refrigeradores de un fabricante  $B$  es de nueve años con una desviación estándar de tres años.
- ¿Cuál es la probabilidad que una muestra de 50 refrigeradores del fabricante  $A$  tenga una vida promedio que sea al menos 1.5 años mayor que la vida promedio de una muestra de 60 refrigeradores del fabricante  $B$ ?
  - El fabricante  $A$  afirma que en al menos 90% de los casos sus refrigeradores duran en promedio muestral por lo menos tres meses más que los del fabricante  $B$ . ¿Cuál debe ser el tamaño mínimo de la muestra que debe considerar el fabricante  $A$  para corroborar su afirmación de manera estadística?
  - El fabricante  $B$  afirma que en a lo más 25% de los casos el promedio muestral de sus refrigeradores duran a lo más dos meses menos que los del fabricante  $A$ . ¿Cuál debe ser el tamaño mínimo de la muestra que debe considerar el fabricante  $B$  para corroborar estadísticamente su afirmación?
- 2.44** Una compañía de auditores se encuentra interesada en estimar la proporción de cuentas para las que existe una discrepancia entre los balances contables reportados entre los clientes y los bancos. ¿Cuántas cuentas deberán seleccionarse de manera que haya una probabilidad de 90% de que la proporción de la muestra se diferencie máximo 0.02 unidades de la proporción real que es de 35%?
- 2.45** Se cree que 16% de los hogares del sur de la Ciudad de México tienen ingresos totales que se clasifican en nivel económico alto. En el norte de la ciudad se cree que este porcentaje es de 11%. Si estas cifras son exactas:
- Calcule la probabilidad de que la proporción muestral de los hogares del sur sea mayor a los del norte máximo en 4%, si se seleccionaron muestras respectivas de 500 y 625 hogares.
  - Encuentre el tamaño mínimo de muestra que se tiene que considerar para que la probabilidad en el inciso  $a$ ) sea menor a 0.20.
  - Resuelva los incisos  $a$ ) y  $b$ ) para una ventaja menor a 2% y explique las diferencias.

## Proyectos de la unidad 2

- Pruebe que los únicos casos que se pueden presentar en el cálculo del tamaño mínimo de muestra son del teorema 2.5.
- Por medio de un paquete genere 1 000 muestras de tamaño 5 de una distribución uniforme 0 y 1. En cada una calcule su promedio y trace el histograma de clases de frecuencia de los 1 000 promedios. Después repita lo anterior para muestras de tamaño 10, 20, 30, 40 y compruebe de manera gráfica que se cumple el TCL.
- Compruebe que si en la práctica anterior genera 5 000 números aleatorios uniformes (1 000 muestras de tamaño 5), y después traza su histograma de clases de frecuencia no obtendrá el mismo resultado que en la práctica anterior. Repita esto para los 10 000, 20 000, 30 000 y 40 000 números aleatorios uniformes de los otros tamaños de muestra y explique qué sucede.
- Por medio de un paquete genere 1 000 muestras de tamaño 5 de una distribución binomial con parámetros 5 y 0.1. En cada una calcule su promedio y trace el histograma de clases de frecuencia de los 1 000 promedios. Después repita lo anterior para muestras de tamaño 10, 20, 30, 40 y compruebe de manera gráfica que se cumple el TCL.
- Repita la práctica anterior con  $p = 0.25$  y  $p = 0.5$ . ¿Qué se puede concluir?



# Estimación puntual y por intervalos de confianza

UNIDAD  
**3**



## Competencia específica a desarrollar

- Interpretar, analizar e integrar datos para determinar los intervalos de confianza que permitan estimar los parámetros poblacionales.

## ¿Qué sabes?

- ¿Qué son los estimadores puntuales?
- ¿Qué es una prueba de hipótesis?
- ¿Cuál es la utilidad de los intervalos de confianza?
- ¿Cómo se determinan los estimadores insesgados?
- ¿Por qué es útil un coeficiente de confianza?

## Introducción

En las unidades 1 y 2 se inicia el estudio de la inferencia estadística, así como las bases para la distribución muestral con la que es posible obtener estimaciones con respecto a los parámetros de interés. En la presente unidad revisamos algunos conceptos básicos sobre el estudio clásico de la estimación de parámetros, los cuales se pueden clasificar en tres áreas fundamentales:

- **Estimadores puntuales.** En esta parte se revisan los conceptos fundamentales sobre los estimadores.
- **Intervalos de confianza.** Aquí se presentan los intervalos en los que se infiere la localización del parámetro de interés con cierta probabilidad.
- **Pruebas de hipótesis.** En esta sección se estudia la comprobación de los supuestos sobre un parámetro por medio de los llamados contrastes de hipótesis (véase unidad 4).

Se puede decir que los **estimadores puntuales** son la base teórica para el desarrollo de la inferencia estadística, tema que tratamos en ésta y la siguiente unidad. Aunque en cuestiones prácticas no es muy apropiado utilizarlos para realizar estimaciones de parámetros.

En la unidad iniciamos el estudio de los estimadores de los parámetros con el conjunto de valores posibles que puede tomar uno, al que llamaremos espacio paramétrico, y seguiremos con las definiciones de estimador y estimador puntual.

Definidos los estimadores puntuales y enumeradas algunas propiedades básicas que deben cumplir para que sean buenos, como: estimadores insesgados, de varianza mínima, error cuadrado medio de un estimador; asimismo, estudiamos algunas propiedades asintóticas de los estimadores.

Después de iniciar el estudio de los estimadores puntuales es lógico suponer que la inferencia llevada a cabo mediante un valor puntual no es la más adecuada, puesto que puede variar mucho de realización en realización. Por tal razón, se deduce que es preferible indicar un intervalo de valores en el que se estime, con cierto *grado de confianza*, la localización del parámetro en estudio.

Lo anterior da origen al estudio de los intervalos de confianza, para esto suponemos que se tiene una población de la que desconocemos su parámetro,  $\theta$ , y que bajo ciertas condiciones (analizadas en las siguientes secciones), encontramos que  $\theta \in (\hat{\theta}_i, \hat{\theta}_s)$ , donde los puntos extremos  $\hat{\theta}_i$  y  $\hat{\theta}_s$  dependen del valor del estadístico  $\hat{\Theta}$  para la realización particular de una muestra aleatoria y se conocen como **extremo inferior** y **extremo superior**. Debido a que los extremos,  $\hat{\theta}_i$  y  $\hat{\theta}_s$ , del intervalo dependen de la realización de la muestra, representan valores particulares de las variables aleatorias  $\hat{\Theta}_i$  y  $\hat{\Theta}_s$ , que son función del estadístico  $\hat{\Theta}$  correspondiente a la muestra aleatoria y la distribución muestral.

Con base en las variables aleatorias anteriores es posible calcular la probabilidad de que el parámetro  $\theta$  se encuentre en el intervalo establecido. Es decir, si simbolizamos por  $1 - \alpha$ , con  $\alpha \in (0, 1)$  la probabilidad mencionada tenemos:

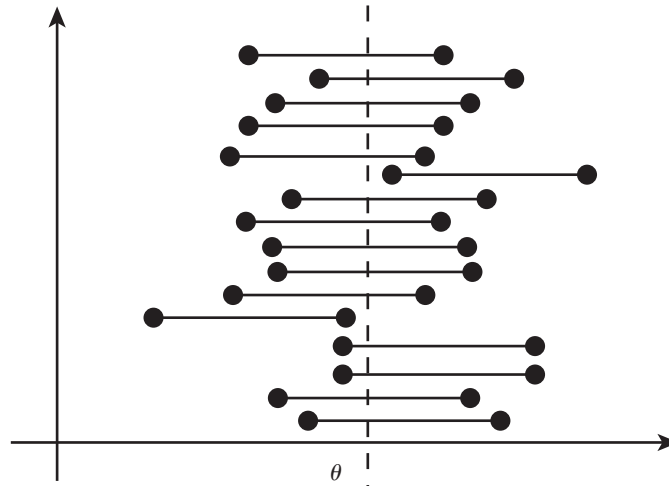
$$P(\hat{\Theta}_i < \theta < \hat{\Theta}_s) = 1 - \alpha$$

Al intervalo en el que está localizado el parámetro  $\theta$ , correspondiente a una realización de la muestra  $\hat{\theta}_i < \theta < \hat{\theta}_s$ , se le llama **intervalo de confianza**, mientras que a la fracción  $1 - \alpha$  se le da el nombre de **coeficiente** o **grado de confianza** y los extremos  $\hat{\theta}_i$  y  $\hat{\theta}_s$  se denominan **límites de confianza inferior** y **superior**. Ahora bien, ¿cómo debemos interpretar un intervalo de confianza? Suponga que tenemos una población con parámetro  $\theta$  y elegimos una muestra aleatoria  $X_1, X_2, \dots, X_n$ , de la cual definimos la estadística  $\hat{\Theta}$ , correspondiente al parámetro  $\theta$ . Por otro lado, definimos como funciones de  $\hat{\Theta}$  a las variables aleatorias  $\hat{\Theta}_i$  y  $\hat{\Theta}_s$ , de manera que para cada realización  $x_1, x_2, \dots, x_n$  de la muestra aleatoria  $X_1, X_2, \dots, X_n$ , obtenemos un intervalo de confianza  $(\hat{\theta}_i, \hat{\theta}_s)$  para el parámetro, donde  $\hat{\theta}_i$  y  $\hat{\theta}_s$  son los valores de las variables  $\hat{\Theta}_i$  y  $\hat{\Theta}_s$ . Entonces:

$$P(\hat{\Theta}_i < \theta < \hat{\Theta}_s) = 1 - \alpha$$

indica que  $(1 - \alpha)$  100% de los intervalos encontrados,  $(\hat{\theta}_i, \hat{\theta}_s)$  para cada realización  $x_1, x_2, \dots, x_n$  de la muestra aleatoria sí contienen al parámetro.

Lo anterior se puede representar como se observa en la figura 3.1.



**Figura 3.1** Cierta porcentaje de los intervalos contienen al parámetro.

En la figura 3.1 los segmentos de recta representan la longitud de los intervalos para cada realización de la muestra, mientras que la recta vertical punteada sirve como referencia para mostrar en qué intervalos sí se localiza el parámetro  $\theta$ .

En este momento cabe aclarar un error de interpretación que se comete con frecuencia en los intervalos de confianza. Sea el intervalo de confianza  $(\hat{\theta}_i, \hat{\theta}_s)$  obtenido de la realización  $x_1, x_2, \dots, x_n$  de una muestra aleatoria para el parámetro  $\theta$ ; en muchas ocasiones se dice que:

$$P(\hat{\theta}_i < \theta < \hat{\theta}_s) = 1 - \alpha$$

¡Esto es incorrecto!, ya que en  $P(\hat{\theta}_i < \theta < \hat{\theta}_s)$  no tenemos variables aleatorias; por consiguiente, no existen probabilidades que calcular (en este caso la probabilidad es cero o uno).

Por ejemplo, si a partir de la realización de una muestra aleatoria de 20 focos se encuentra que la duración promedio en horas es  $\bar{x} = 750$  y con base en este valor, y alguna regla que veremos más adelante, *estimamos* que el parámetro  $\mu$  puede encontrarse, con una probabilidad  $1 - \alpha$  (establecida de antemano) en el intervalo (740 760). Lo anterior no debe interpretarse como:

$$P(740 < \mu < 760) = 1 - \alpha$$

Puesto que en la expresión anterior no existen variables aleatorias, no podemos calcular probabilidades.

Hasta el momento se expusieron algunas ideas generales sobre lo que tratamos respecto a los intervalos de confianza. En las siguientes secciones veremos la metodología que nos ayudará a la obtención de los intervalos de confianza para los parámetros de una población con distribución normal. Es decir, determinar los intervalos de confianza de los parámetros en poblaciones normales como son:

- Media y diferencia de medias de poblaciones aproximadamente normales.
- Varianza y razón entre varianzas de poblaciones aproximadamente normales.

Además estudiamos los intervalos de confianza para poblaciones con distribución de Bernoulli para:

- Proporciones y diferencia de proporciones de poblaciones con distribución de Bernoulli.

### 3.1 Conceptos básicos sobre estimadores puntuales

Los principios básicos de la inferencia estadística consisten en crear métodos para realizar conclusiones o inferencias acerca de la población. En la actualidad, estos métodos se dividen en dos grupos: **clásicos** y **bayesianos**.

En los primeros, la inferencia se realiza por medio de los resultados de un muestreo aleatorio; en los segundos, se lleva a efecto con base en el conocimiento previo sobre la distribución de los parámetros desconocidos (en los métodos bayesianos los parámetros son considerados variables aleatorias). Desde hace algunas décadas los métodos bayesianos tienen gran auge en diferentes esferas de la estadística, pero su estudio queda fuera de los objetivos del texto.

En estos momentos pueden surgir las preguntas: ¿cómo saber qué parámetro estimar?, ¿cómo determinar un valor para el estimador del parámetro?

Las respuestas se pueden dar basándonos en los fundamentos de la inferencia estadística, los cuales tratamos en las secciones siguientes.

## Espacio paramétrico

Iniciamos el estudio de la inferencia estadística determinando el conjunto de valores posibles de un parámetro. Sea  $X_1, \dots, X_n$  una muestra aleatoria con función de densidad  $f(x; \theta)$ , donde la forma de la función es conocida pero el parámetro  $\theta$  desconocido, solo sabemos que pertenece al espacio paramétrico denotado por  $\Omega$ .

### Ejemplos 3.1 Espacio paramétrico

1. Las variables aleatorias tienen función de densidad exponencial con parámetro  $\beta$ ,

$$f(x; \beta) = \begin{cases} \frac{1}{\beta} e^{-\frac{x}{\beta}}, & x \geq 0, \beta > 0 \\ 0, & \text{para cualquier otro valor de } x \end{cases}$$

En este caso, el parámetro es  $\theta = \beta$  y el conjunto de valores del parámetro es  $\Omega = (0, \infty)$ .

2. Las variables tienen función de densidad normal con parámetros  $\mu$  y  $\sigma^2$ ,

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \text{ con } \mu \in \mathbb{R} \text{ y } \sigma > 0$$

En esta situación el espacio paramétrico es  $\Omega = \mathbb{R} \times \mathbb{R}^+$ .

3. Las variables tienen función de densidad uniforme con parámetros  $a$  y  $b$ ,

$$f(x; a, b) = \begin{cases} \frac{1}{b-a}, & a < b \text{ con } a, b \in \mathbb{R} \\ 0 & \text{para cualquier otro valor de } x \end{cases}$$

En esta situación el espacio paramétrico es  $\Omega = \mathbb{R} \times \mathbb{R}^+$ .

4. Las variables tienen función de densidad de Poisson con parámetro  $\mu$ ,

$$f(x; \mu) = \begin{cases} \frac{e^{-\mu} \mu^x}{x!}, & x = 0, 1, 2, \dots \text{ con } \mu \in (0, \infty) \\ 0, & \text{para cualquier otro valor de } x \end{cases}$$

En esta situación el espacio paramétrico es  $\Omega = (0, \infty)$ .

5. Las variables tienen función de densidad de Bernoulli, con parámetro  $p$ ,

$$f(x; p) = \begin{cases} p, & x = 1 \\ 0 & x = 0 \end{cases}, \text{ con } p \in [0, 1]$$

En esta situación el espacio paramétrico es  $\Omega = [0, 1]$ .

De los ejemplos anteriores, podemos notar que en la inferencia estadística al hablar del parámetro hacemos referencia a una familia de densidades. Es decir, debemos estudiar la forma de obtener información sobre el parámetro  $\theta$  de la función de densidad fijada. Para facilitar este estudio, en el texto lo restringimos a funciones de densidad con uno o dos parámetros.

### Ejercicios 3.1

1. Encuentre el espacio paramétrico para una muestra aleatoria con densidad normal con varianza conocida e igual a 4.
2. Encuentre el espacio paramétrico para una muestra aleatoria con densidad uniforme entre  $(-\theta, \theta)$ .
3. Encuentre el espacio paramétrico para una muestra aleatoria con densidad Weibull.
4. Encuentre el espacio paramétrico para una muestra aleatoria con densidad *jicuada*.
5. Encuentre el espacio paramétrico para una muestra aleatoria con densidad geométrica.

En un estudio formal de los estimadores puntuales se deben analizar las familias de densidades de los parámetros bajo conceptos un poco más complicados de los que se requiere en este libro; por ello, solo veremos los conceptos necesarios para su desarrollo; estamos conscientes de que en ocasiones los temas tratados no se podrán ver con la profundidad que algunos lectores quisieran.

### Valores de los estimadores puntuales

Suponga que se quiere realizar una inferencia con respecto a la calificación media de todos los alumnos que cursan la materia de cálculo, para lo cual se analiza la realización de una muestra aleatoria de 10 de éstos con calificaciones:

8, 4, 9, 9, 6, 8, 2, 7, 3 y 6.

Con los datos anteriores calculamos un valor del estadístico  $\bar{x}$ :

$$\bar{x} = \frac{1}{10}(8 + 4 + 9 + 9 + 6 + 8 + 2 + 7 + 3 + 6) = 6.2$$

Con base en el valor calculado del estadístico  $\bar{x}$  es posible realizar una inferencia del parámetro  $\mu$  para las calificaciones de la materia de cálculo. Es decir, se puede hacer una *estimación puntual* del parámetro media con respecto a las calificaciones de esta materia. En este caso, el parámetro  $\mu$  se estima con respecto al estimador puntual  $\bar{x} = 6.2$ . En general, se define como sigue:

Sea una población con parámetro  $\theta$  y  $X_1, X_2, \dots, X_n$  una muestra aleatoria de la población, con  $\hat{\Theta} = U(X_1, \dots, X_n)$  el estadístico correspondiente de  $\theta$ , en la parte de estimación a  $\hat{\Theta}$  se le llama **estimador** de  $\theta$ , mientras que al valor  $\hat{\theta}$  de  $\hat{\Theta}$  obtenido de una realización de la muestra aleatoria se le llama **estimador puntual** de  $\theta$ .

Con esto se puede concluir que si  $\theta$  es el parámetro con espacio paramétrico  $\Omega$ , cualquier estimador puntual de  $\theta$  debe estar contenido en  $\Omega$ . Es decir, si  $U(X_1, \dots, X_n)$  es el estimador del parámetro su distribución tendrá como dominio a  $\Omega$ .

Podemos observar que el estimador  $\hat{\Theta}$  correspondiente al parámetro  $\theta$  es un estadístico; por tanto, la respuesta a la pregunta sobre qué estimador utilizar la podemos contestar de manera parcial. Sea una población descrita por la función de densidad  $f(x; \theta)$ , suponga que deseamos obtener un estimador de  $\theta$ . Para iniciar su búsqueda comencamos por revisar las relaciones que tiene el parámetro y los valores que conocimos en la estadística descriptiva, como son los centrales (la media, moda, mediana, cuantiles), de desviación (varianza, rangos, desviación estándar), coeficiente de variación, etcétera.

El problema de la búsqueda de estimadores es el tema central de la unidad, por lo que no debemos preocuparnos demasiado en este momento por tener uno bueno, ya que faltan por revisar todas las propiedades que deben cumplir para que sean buenos estimadores. Además, en la práctica, en la mayoría de situaciones los que buscamos estarán relacionados con la media y varianza, ya que si no se dice otra cosa  $\mu$  lo estimaremos con  $\bar{x}$  y  $\sigma^2$  con  $S_{n-1}^2$ .

### Ejemplos 3.2 Estimador puntual

1. Sea una población con función de densidad exponencial con parámetro  $\beta$ :

$$f(x; \beta) = \begin{cases} \frac{1}{\beta} e^{-\frac{x}{\beta}}, & x \geq 0, \beta > 0 \\ 0, & \text{para cualquier otro valor de } x \end{cases}$$

Por otro lado, sea  $X_1, X_2, \dots, X_6$  una muestra aleatoria para estimar a  $\beta$ , de la cual se elige una realización dada por, 3.2, 4.5, 2.3, 1.9, 3.5 y 2.8, utilice los valores de la realización para estimar al parámetro  $\theta = \beta$ .

#### Solución

De la distribución exponencial se sabe que el parámetro  $\beta = E(X)$ ; por otro lado, el valor esperado se puede estimar con la media aritmética  $\bar{x}$ , de manera que:

$$\hat{\beta} = \frac{1}{6}(3.2 + 4.5 + 2.3 + 1.9 + 3.5 + 2.8) = 3.033$$

Con base en la realización tenemos un estimador puntual de  $\beta$ .

Recuerde que el espacio paramétrico para la distribución exponencial es  $\Omega = (0, \infty)$  y  $3.033 \in \Omega$ .

2. Sea una población con función de densidad uniforme con parámetros  $a$  y  $b$ ,

$$f(x; a, b) = \begin{cases} \frac{1}{b-a}, & a < b \text{ con } a, b \in \mathbb{R} \\ 0 & \text{para cualquier otro valor de } x \end{cases}$$

Por otro lado, sea  $X_1, X_2, \dots, X_{10}$  una muestra aleatoria para estimar los parámetros  $a$  y  $b$ , de la cual se elige una realización dada por, 6.2, 3.5, 5.3, 8.9, 3.5, 4.8, 5.2, 6.7, 3.0 y 8.1, utilice los valores de la realización para estimar los parámetros  $a$  y  $b$ .

#### Solución

Se sabe de la distribución uniforme que los parámetros  $a$  y  $b$  son los valores mínimo y máximo de los valores de la variable aleatoria. Por consiguiente:

$$\hat{a} = \min \{6.2, 3.5, 5.3, 8.9, 3.5, 4.8, 5.2, 6.7, 3.0, 8.1\} = 3.0$$

$$\hat{b} = \max \{6.2, 3.5, 5.3, 8.9, 3.5, 4.8, 5.2, 6.7, 3.0, 8.1\} = 8.9$$

Con base en la realización tenemos un estimador puntual de  $a$  y  $b$ .

Recuerde que el espacio paramétrico para la uniforme es  $\Omega = \mathbb{R} \times \mathbb{R}$  y  $(3.0, 8.9) \in \Omega$ .

## Ejercicios 3.2

1. Sea una población con función de densidad uniforme distribuida entre  $(-2, b)$ :

$$f(x; a, b) = \begin{cases} \frac{1}{b+2}, & -2 < b \\ 0, & \text{para cualquier otro valor de } x \end{cases}$$

Por otro lado, sea  $X_1, X_2, \dots, X_6$  una muestra aleatoria para estimar el parámetro  $b$ , de la cual se elige una realización dada por: 4.1, 3.5, -1.2, 2.5, 8.2 y 5.2, utilice los valores de la realización para estimar al parámetro  $b$ .

2. Sea una población con función de densidad normal con parámetros  $\mu$  y  $\sigma^2$ :

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \text{ con } \mu \in \mathbb{R} \text{ y } \sigma \in (0, \infty)$$

Por otro lado, sea  $X_1, X_2, \dots, X_{12}$  una muestra aleatoria para estimar los parámetros  $\mu$  y  $\sigma^2$ , de la cual se elige una realización dada por: 124, 122, 143, 135, 128, 135, 140, 122, 127, 138, 131 y 141, utilice los valores de la realización para estimar los parámetros  $\mu$  y  $\sigma^2$ .

3. Sea una población con función de densidad de Bernoulli con parámetro  $p$ :

$$f(x; p) = \begin{cases} 0, & \text{en caso de fracaso} \\ 1, & \text{en caso de éxito} \end{cases}$$

Por otro lado, sea  $X_1, X_2, \dots, X_{10}$  una muestra aleatoria para estimar al parámetro  $p$ , de la que se elige una realización dada por 0, 1, 1, 0, 0, 1, 0, 1, 0 y 0, utilice los valores de la realización para estimar al parámetro  $p$ .

## Estimadores insesgados

En los últimos ejemplos de la subsección anterior sobre los estimadores puntuales apreciamos que éstos dependen del valor de la realización, por tanto, no se puede esperar que el valor puntual estime con certeza al parámetro, pues también depende de la estadística utilizada. Por ejemplo, si toda la población estudiantil tiene una calificación promedio  $\mu = 6.5$  en álgebra y se considera una muestra aleatoria de tres estudiantes  $X_1, X_2, X_3$  con una realización 3, 6 y 6 para una estimación del parámetro, resulta:

$$\bar{x} = (3 + 6 + 6)/3 = 5$$

Es decir, el estadístico media difiere del parámetro media en 1.5 unidades; mientras que el estadístico mediana,  $\tilde{x} = 6$  ( $3 < 6 < 6$ ) difiere del parámetro en solo 0.5 unidades. De este modo, con la realización anterior el estadístico mediana estima mejor al parámetro. Pero, ¿qué pasará si en una segunda realización de la muestra aleatoria resultan 4, 4 y 10 las calificaciones para la estimación del parámetro? Se tendrá:

$$\bar{x} = (4 + 4 + 10)/3 = 6$$

Luego, para esta realización el estadístico media difiere del parámetro en 0.5 unidades, mientras que el estadístico mediana,  $\tilde{x} = 4$ , en 2.5 unidades.

Es decir, con la realización anterior el estadístico media estima mejor al parámetro. Por tanto, en forma natural surgen las preguntas: ¿Habrà algún estimador puntual que sea mejor a los demás? ¿Cómo encontrar un estimador mejor a los otros? ¿Qué propiedades deben cumplir los mejores estimadores?

Se formularon varias preguntas referentes a los estimadores puntuales, a los cuales, en este momento solo podemos dar respuesta parcial. De manera que los estimadores que se prefieran deben cumplir ciertas propiedades, por el momento podemos restringirnos a:

- Estimadores cuyos valores siempre pertenecen al espacio paramétrico.
- Otra propiedad deseable del estimador  $\hat{\Theta} = U(X_1, \dots, X_n)$  se presenta cuando el valor esperado de la distribución del estimador  $\hat{\Theta}$  es igual al parámetro  $\theta$ .

El estimador  $\hat{\Theta} = U(X_1, \dots, X_n)$  de una función  $g(\theta)$  del parámetro  $\theta$  se llama **estimador insesgado** de  $g(\theta)$  si  $E_{\theta}(\hat{\Theta}) = g(\theta)$ , en caso contrario ( $E_{\theta}(\hat{\Theta}) \neq g(\theta)$ ), se conoce como **estimador sesgado** de  $g(\theta)$ .

Para probar si un estimador es insesgado utilizamos la linealidad del operador valor esperado:

$$E(a_1X_1 + a_2X_2 + \dots + a_nX_n) = a_1E(X_1) + a_2E(X_2) + \dots + a_nE(X_n)$$

### Ejemplos 3.3 Estimadores insesgados

1. Sea  $X_1, X_2, \dots, X_5$  una muestra aleatoria, pruebe cuáles de los siguientes estimadores de  $\mu$  son insesgados:

$$T_1 = \frac{X_1 + X_2 + X_5}{3}, \quad T_2 = \frac{X_1 + X_2 + \dots + X_5}{10} \quad \text{y} \quad T_3 = \frac{X_1 + X_2 + X_3 - X_4 + X_5}{3}$$

#### Solución

Para verificar qué estimadores son insesgados se empleará la definición y la propiedad de linealidad del valor esperado.

- Para el estadístico  $T_1$ :

$$\begin{aligned} E(T_1) &= E\left[\frac{X_1 + X_2 + X_5}{3}\right] = \frac{1}{3}E[X_1 + X_2 + X_5] = \frac{1}{3}[E(X_1) + E(X_2) + E(X_5)] \\ &= \frac{1}{3}(3\mu) = \mu \end{aligned}$$

Esto muestra que  $T_1$  es un estimador insesgado de  $\mu$

- Para el estadístico  $T_2$ :

$$\begin{aligned} E(T_2) &= E\left[\frac{X_1 + X_2 + X_3 + X_4 + X_5}{10}\right] = \frac{1}{10}E[X_1 + X_2 + X_3 + X_4 + X_5] \\ &= \frac{1}{10}[E(X_1) + E(X_2) + E(X_3) + E(X_4) + E(X_5)] = \frac{1}{10}(5\mu) = \frac{1}{2}\mu \neq \mu \end{aligned}$$

Esto muestra que  $T_2$  es un estimador sesgado.

- Para el estadístico  $T_3$ :

$$\begin{aligned} E(T_3) &= E\left[\frac{X_1 + X_2 + X_3 - X_4 + X_5}{3}\right] = \frac{1}{3}E[X_1 + X_2 + X_3 - X_4 + X_5] \\ &= \frac{1}{3}[E(X_1) + E(X_2) + E(X_3) - E(X_4) + E(X_5)] = \frac{1}{3}(\mu + \mu + \mu - \mu + \mu) = \frac{1}{3}(3\mu) = \mu \end{aligned}$$

Por tanto,  $T_3$  también es un estimador insesgado de la media.

2. Pruebe que en general, si  $X_1, X_2, \dots, X_n$  es la muestra aleatoria de una población con media  $\mu$ , entonces  $\bar{X}$  es un estimador insesgado de  $\mu$ .

#### Solución

La demostración se concluye, puesto que se demostró en la unidad 2 que  $E(\bar{X}) = \mu$ .

3. Pruebe que en general, si  $X_1, X_2, \dots, X_n$  es la muestra aleatoria de una población con media  $\mu$  y varianza  $\sigma^2$  (finita), entonces  $S_{n-1}^2$  es un estimador insesgado y  $S_n^2$  es un estimador sesgado de  $\sigma^2$ .

#### Solución

Debido a que  $(n-1)S_{n-1}^2 = \sum_{i=1}^n (X_i - \bar{X})^2$  se tiene:



$$\begin{aligned}
 (n-1)E\{S_{n-1}^2\} &= E\left\{\sum_{i=1}^n (X_i - \bar{X})^2\right\} = E\left\{\sum_{i=1}^n (X_i^2 - 2\bar{X}X_i + \bar{X}^2)\right\} \\
 &= E\left\{\sum_{i=1}^n (X_i^2) - n\bar{X}^2\right\} \\
 &= \sum_{i=1}^n E(X_i^2) - nE(\bar{X}^2)
 \end{aligned}$$

Por otro lado, como las variables de una muestra aleatoria tienen la misma distribución que la población, resulta:

$$E(X_i^2) = V(X_i) + E^2(X_i) = \sigma^2 + \mu^2, \text{ para } i = 1, 2, \dots, n$$

De manera similar, al utilizar el resultado mostrado en la unidad 2:

$$E(\bar{X}^2) = V(\bar{X}) + E^2(\bar{X}) = \frac{\sigma^2}{n} + \mu^2$$

Al continuar con el ejemplo y utilizar los dos resultados anteriores, tenemos:

$$\begin{aligned}
 (n-1)E\{S_{n-1}^2\} &= \sum_{i=1}^n E(X_i^2) - nE(\bar{X}^2) = \sum_{i=1}^n (\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right) \\
 &= n(\sigma^2 + \mu^2) - \sigma^2 - n\mu^2 \\
 &= (n-1)\sigma^2
 \end{aligned}$$

Por último, si se divide entre  $n-1$  se obtiene el resultado esperado.

Para la varianza  $S_n^2$  resulta que:

$$S_n^2 = \frac{n-1}{n} S_{n-1}^2, \text{ luego } E(S_n^2) = \frac{n-1}{n} E(S_{n-1}^2) = \frac{n-1}{n} \sigma^2 \neq \sigma^2$$

Por tanto, se concluye que  $S_n^2$  es un estimador sesgado de  $\sigma^2$ .

De los ejemplos 3.3 y los ejercicios 3.3 que se presentan a continuación se obtienen los estimadores insesgados más comunes que serán empleados con mayor regularidad en la parte metodológica del texto (véase tabla 3.1).

**Tabla 3.1**

	Parámetro	Estadístico insesgado
Media	$\mu$	$\bar{X}$
Diferencia de medias	$\mu_Y - \mu_X$	$\bar{Y} - \bar{X}$
Variancia	$\sigma^2$	$S_{n-1}^2$
Proporciones	$p$	$\bar{X}$
Diferencia de proporciones	$p_X - p_Y$	$\bar{X} - \bar{Y}$

## Ejercicios 3.3

1. Sean  $X_1, X_2, X_3$  y  $X_4$  una muestra aleatoria seleccionada de una población con media  $\mu$  y desviación estándar  $\sigma$ . Considere los siguientes estimadores de  $\mu$ :

$$T_1 = \frac{X_1 + X_2 + X_3 + X_4}{6}, T_2 = \frac{X_1 + X_2 + X_3 + 2X_4}{5}, T_3 = \frac{X_1 + 2X_2 + 3X_3 + 4X_4}{10}$$

Determine cuáles son insesgados.

2. Sean  $X_1, X_2, X_3$  y  $X_4$  variables de una muestra aleatoria seleccionada de una población con media  $\mu$ . De los siguientes estimadores indique cuál es un estimador insesgado de la media  $\mu$ .

a)  $\hat{\theta}_1 = \frac{X_1 + 2X_2 + 3X_3 + 4X_4}{4}$

b)  $\hat{\theta}_2 = \frac{X_1 + 2X_2 + X_3 + X_4}{5}$

c)  $\hat{\theta}_3 = X_1 + X_2 + X_3 - 2X_4$

d)  $\hat{\theta}_4 = \frac{X_1 + 2X_2 - 3X_3 + 4X_4}{4}$

3. Sea  $X_1, X_2$  una muestra aleatoria de tamaño 2 de una distribución normal con  $\mu = \theta$  y  $\sigma = 1$ . Considere los siguientes tres estimadores de  $\theta$ .

$$T_1 = \frac{2}{3}X_1 + \frac{1}{3}X_2, T_2 = \frac{1}{4}X_1 + \frac{3}{4}X_2, T_3 = \frac{1}{2}X_1 + \frac{1}{2}X_2$$

Pruebe que  $T_i$  es un estimador insesgado de  $\mu$  para  $i = 1, 2, 3$ .

4. Sea  $X_1, X_2, X_3, \dots, X_n$  una muestra aleatoria seleccionada de una población con media  $\mu$  y desviación estándar

$\sigma$ . Pruebe si  $T = \frac{s_{n-1}^2 + s_n^2}{2}$  es un estimador insesgado de  $\sigma^2$ .

5. Sean  $X_1, X_2$  y  $X_3$  y  $Y_1, Y_2$  y  $Y_3$  muestras aleatorias independientes de dos poblaciones con medias  $\mu_1$  y  $\mu_2$  y varianzas  $\sigma_1^2$  y  $\sigma_2^2$ , respectivamente, demuestre que  $\bar{X} - \bar{Y}$  es un estimador insesgado de  $\mu_1 - \mu_2$ .

## Estimadores insesgados de distribuciones específicas

En los ejemplos anteriores, para comprobar si un estimador era o no insesgado no requerimos del conocimiento de la distribución de la variable debido a que solo era necesario recurrir a la linealidad del operador del valor esperado, pero sí debemos conocer la distribución de la variable para calcular su valor esperado, por lo que recomendamos repasar este tema.

## Ejemplo 3.4 Estimadores insesgados

Sea  $X_1, X_2, \dots, X_n$  una muestra aleatoria de distribuciones exponenciales,

$$f(x; \beta) = \begin{cases} \frac{1}{\beta} e^{-\frac{x}{\beta}}, & x > 0, \quad \beta > 0 \\ 0 & \text{en otro caso} \end{cases}$$

Pruebe que el estadístico  $T = \bar{X}$ , es un estimador insesgado de  $\beta$ .

**Solución**

Para probar que el estadístico  $T = \bar{X}$  es un estimador insesgado de  $\beta$  necesitamos calcular su valor esperado. Vimos en la unidad 2 que  $E(\bar{X}) = \mu = E(X)$ , por tratarse de una distribución exponencial  $E(X) = \beta$  y con esto concluimos la comprobación.

Sea  $T$  un estimador del parámetro  $g(\theta)$ , se llama **sesgo** a la función que representa la diferencia entre el valor esperado de un estimador y el parámetro:

$$S(\theta) = E(T) - g(\theta)$$

A continuación, se muestra un ejemplo para ilustrar este concepto.

### Ejemplo 3.5 Ilustración del sesgo del estimador

Sea  $X_1, X_2, \dots, X_n$  una muestra aleatoria de densidades uniformes entre  $(\theta - 2, \theta)$ , encuentre el sesgo del estimador  $T = \bar{X}$  de  $\theta$ .

#### Solución

Para encontrar el sesgo tenemos en cuenta que  $E(\bar{X}) = \mu = [\theta + (\theta - 2)]/2$  es el valor esperado de la distribución uniforme. Luego:

$$S(\theta) = E(T) - \theta = E(\bar{X}) - \theta = \mu - \theta = \frac{\theta + (\theta - 2)}{2} - \theta = \theta - 1 - \theta = -1$$

El sesgo vale  $-1$ , entonces  $T = \bar{X} + 1$  es un estimador insesgado de  $\theta$ .

Ahora bien, ¿se cumple la existencia y unicidad de los estimadores insesgados?

En el siguiente ejemplo se muestra que el estimador insesgado de un parámetro no siempre existe.

### Ejemplo 3.6 Estimador insesgado de un parámetro

Sea  $X$  una sola observación de una distribución Bernoulli con parámetro  $p \in (0, 1)$  y  $X = \{0, 1\}$ , demuestre que el parámetro  $g(p) = \log(p)$  no tiene estimador insesgado.

#### Solución

La prueba se hará por reducción al absurdo; es decir, se supondrá que existe un estimador insesgado para  $g(p) = \log(p)$  y se llegará a un absurdo (algo que no puede ser).

Sea  $T$  un estimador insesgado de  $g(p) = \log(p)$ , entonces debe cumplirse:

$$g(p) = \log(p) = E(T) = \sum_{x=0}^1 T(x)f(x; p) = T(0)f(0; p) + T(1)f(1; p) = T(0) + [T(1) - T(0)] p$$

Es decir,  $\log(p) = T(0) + [T(1) - T(0)] p$ . ¡El logaritmo es una función lineal!, lo que no puede ser.

Por consiguiente, la suposición de que  $T$  es un estimador insesgado de  $g(p) = \log(p)$  es falsa, luego dicho parámetro no tiene estimadores insesgados.

Por otro lado, tenemos la interrogante sobre la unicidad del estimador insesgado. Es decir, en el caso de encontrar un estimador insesgado de  $T$  para un parámetro  $\theta$  surge la pregunta, ¿si existe un estimador insesgado, será único o existen más?

Con respecto a la unicidad, la respuesta no resulta tan sencilla. Por ahora veremos un resultado que concluye: "Si un parámetro tiene dos estimadores insesgados, entonces tiene una infinidad de estimadores insesgados".

### ◆ Proposición 3.1

Sean  $T_1$  y  $T_2$  dos estimadores insesgados de  $\theta$ , entonces  $\alpha T_1 + (1 - \alpha)T_2$  es otro estimador insesgado de  $\theta$ , para cualquier  $\alpha \in \mathbb{R}$ .

De lo anterior concluimos que un parámetro puede **no tener** estimadores insesgados, **tener uno** o tener **una infinidad** de estimadores insesgados.

### Demostración

Se calcula el valor esperado de la estadística  $\alpha T_1 + (1 - \alpha)T_2$ :

$$E\{\alpha T_1 + (1 - \alpha)T_2\} = \alpha E\{T_1\} + (1 - \alpha)E\{T_2\} = \alpha\theta + (1 - \alpha)\theta = \theta$$

Lo que implica que  $\alpha T_1 + (1 - \alpha)T_2$  es un estimador insesgado  $\theta$ .

Se había visto que una de las propiedades de interés en los estimadores puntuales es que sean insesgados, pero el estimador puede serlo y no contener toda la información muestral; por ejemplo, si tenemos  $X_1, X_2, \dots, X_n$  la muestra aleatoria de una población con parámetro  $\mu$ , entonces un estimador de éste puede ser  $(X_1 + X_2)/2$ , que es insesgado, pero no contiene toda la información de la muestra. Por esta razón, el hecho de que el estimador sea insesgado no es suficiente, falta probar que tenga toda la información muestral.

El estudio de las propiedades de los estimadores puntuales es muy extenso y queda fuera de los objetivos prácticos del texto, pero podemos numerar las propiedades deseables que deberían cumplir los estimadores puntuales. Para su análisis se recomienda consultar libros especializados en estos temas.

1. Estadísticos suficientes, criterio de factorización de Neyman-Fisher.
2. Estimadores que cumplan la propiedad de invarianza.
3. Estimadores insesgados con menor varianza.
4. Estimador más eficiente.
5. Estimadores con el menor error cuadrado medio.
6. Estimadores consistentes en probabilidad y error cuadrado medio.
7. Mejores estimadores asintóticamente normales.

### Ejercicios 3.4

1. Sea  $X_1, X_2, \dots, X_{10}$  una muestra aleatoria de densidades uniformes entre  $(\theta, 2\theta)$ ,  $\theta > 0$ , pruebe que  $T = \frac{2}{3}\bar{X}$  es un estimador insesgado de  $\theta$ .
2. Sea  $X_1, X_2, \dots, X_n$  una muestra aleatoria de densidades uniformes entre  $(\theta, \theta + 4)$ , pruebe que  $T = \bar{X} - 2$  es un estimador insesgado de  $\theta$ .
3. Sea  $X_1, X_2, \dots, X_n$  una muestra aleatoria de densidades uniformes entre  $(0, \theta)$ , pruebe que  $T = \left(\frac{n+1}{n}\right)Y_n$  es un estimador insesgado de  $\theta$ .
4. Sea una muestra aleatoria de una sola variable de población con densidad:

$$f(x; \theta) = \begin{cases} \frac{2x}{\theta^2}, & 0 < x < \theta, \quad \theta > 0 \\ 0 & \text{en otro caso} \end{cases}$$

pruebe que  $T = \frac{3}{2}X$  es un estimador insesgado  $\theta$ .

5. Sea  $X$  una sola observación con distribución  $N(1, \sigma^2)$ , pruebe que  $X^2 - 1$  es un estimador insesgado de  $\sigma^2$ .
6. Sea  $X_1, X_2, \dots, X_n$  una muestra aleatoria con función de densidad Pareto:

$$f(x; \theta) = \begin{cases} \frac{\theta}{x^2}, & x > \theta, \quad \theta > 0 \\ 0 & \text{en otro lugar} \end{cases}$$

pruebe que  $T = \frac{n-1}{n}Y_1$  es un estimador insesgado de  $\theta$ .

## 3.2 Conceptos básicos de los intervalos de confianza

En esta sección estableceremos los conceptos básicos necesarios para la comprensión de las fórmulas que utilizaremos más adelante y en general para el estudio de diferentes situaciones en la estimación por intervalos.

En la introducción, iniciamos los preparativos de un intervalo de confianza, en el que concluimos que la estimación por intervalos resulta cuando las variables aleatorias de los extremos cubren al parámetro (recuerde que este último es constante).

Un **intervalo aleatorio** es aquel en el que al menos uno de sus dos extremos es una variable aleatoria.

Veamos un ejemplo de intervalo aleatorio.

### Ejemplo 3.7 Intervalo aleatorio

Suponga que se tiene una variable aleatoria  $X$  con distribución binomial con parámetros  $n = 20$  y  $p = 0.3$ . ¿Cuál es la probabilidad de que el intervalo aleatorio  $(X, 2.5X)$  contenga al número 5 (el 5 se puede considerar como el valor real de un parámetro en estudio)?

#### Solución

Tenemos que calcular la probabilidad  $P(X \leq 5 \leq 2.5X)$ , si consideramos que  $X$  es una variable aleatoria con distribución binomial,  $n = 20$  y  $p = 0.3$ . Para esto utilizaremos las tablas de la distribución binomial acumulada:

$$P(X \leq 5 \leq 2.5X) = P(X \leq 5, X \geq 2) = P(2 \leq X \leq 5) = F(5) - F(1) = 0.4164 - 0.0076 = 0.4088$$

donde  $F(5)$  es la distribución acumulada de la binomial hasta el valor 5, de forma similar  $F(1)$ . Del resultado anterior, se dice que la confianza de que  $5 \in (X, 2.5X)$  es solo de 40.88%, esto es un resultado pobre.

Una forma de aumentar la probabilidad en el problema anterior es aumentar la longitud del intervalo aleatorio, pero éste puede dejar de ser atractivo cuando es muy grande.

Con base en lo anterior, concluimos que en la estimación por intervalos se busca un intervalo aleatorio que **no tenga una longitud grande**, pero que la probabilidad de que el parámetro esté en el intervalo sí lo sea.

Sea  $X_1, X_2, \dots, X_n$  una muestra aleatoria de una población con parámetro  $\theta$  y función de densidad  $f(x_1, \dots, x_n; \theta)$ , además  $(\hat{\Theta}_i(X_1, \dots, X_n), \hat{\Theta}_s(X_1, \dots, X_n))$  un intervalo aleatorio con estadísticos  $\hat{\Theta}_i$  y  $\hat{\Theta}_s$  para  $\theta$ .

La **amplitud esperada** de un intervalo aleatorio  $(\hat{\Theta}_i, \hat{\Theta}_s)$  se expresa como:

$$AE = E\{\hat{\Theta}_s - \hat{\Theta}_i\}, P(\hat{\Theta}_i < \hat{\Theta}_s) = 1$$

Ahora, podemos definir tanto el intervalo de confianza como el coeficiente de confianza.

Se dice que tenemos un **intervalo de confianza** de  $(1 - \alpha)100\%$  para  $\theta$  cuando:

$$P(\hat{\Theta}_i < \theta < \hat{\Theta}_s) = 1 - \alpha$$

En el caso de una realización  $x_1, \dots, x_n$  de la muestra aleatoria  $X_1, \dots, X_n$ , se dice que el intervalo de números reales  $(\hat{\theta}_i(x_1, \dots, x_n), \hat{\theta}_s(x_1, \dots, x_n))$  es un **intervalo de confianza con coeficiente de confianza** o nivel de confianza  $1 - \alpha$ . Este intervalo debe *tener una amplitud mínima*. Es decir, para una buena estimación de parámetros, primero se calcula su amplitud esperada:

$$\text{Calcular la } AE = E\{\hat{\Theta}_s(X_1, \dots, X_n) - \hat{\Theta}_i(X_1, \dots, X_n)\}.$$

Luego, se minimiza la amplitud esperada que se encontró.

En general, el problema de los intervalos de confianza es amplio y requiere de técnicas del cálculo diferencial para encontrar el deseado que sea de amplitud mínima. Trabajar de esta forma requiere de una mayor preparación de los estimadores, por estas razones en las siguientes secciones tratamos la parte metodológica de los intervalos de confianza para poblaciones normales y de Bernoulli, al formular los resultados, pero sin llegar a realizar la demostración de que los construidos son los mejores, es decir, función de un estadístico y con la amplitud mínima.

### 3.3 Intervalos de confianza para los parámetros de una población normal

Una de las principales distribuciones en el comportamiento de procesos de producción, control de calidad, modelos de inventarios, etc., es la distribución normal, que tiene dos parámetros, la media ( $\mu$ ) y la varianza ( $\sigma^2$ ). Por esta razón, su desarrollo metodológico inicia con el parámetro media, analizando tres casos; después, estudiamos los intervalos de confianza para la varianza. Las fórmulas que utilizamos aquí para estos intervalos de la media cumplen con las condiciones de un buen estimador por intervalos de confianza; es decir, su amplitud es mínima, mientras que en el caso de la varianza tienen amplitudes casi mínimas, pero que suelen utilizarse por su sencillez.

Cabe aclarar que la aplicación de los intervalos de confianza para los parámetros media y varianza que se analizarán se restringen a **poblaciones normales o aproximadamente normales**. Es decir, las fórmulas que se establezcan en esta sección no aplican a poblaciones que no entren en esta categoría.

#### Intervalos de confianza para la media de poblaciones normales o aproximadamente normales cuando se conoce $\sigma$

##### Teorema 3.1

Sea  $x_1, x_2, \dots, x_n$  una realización tomada de una muestra aleatoria de variables con distribución normal con parámetros  $\mu$  y  $\sigma_0^2$  conocida, entonces un intervalo de  $(1 - \alpha)$  100% de confianza para el parámetro  $\mu$  está dado por:

$$\bar{x} - Z_{\frac{\alpha}{2}} \left( \frac{\sigma_0}{\sqrt{n}} \right) < \mu < \bar{x} + Z_{\frac{\alpha}{2}} \left( \frac{\sigma_0}{\sqrt{n}} \right) \text{ o } \bar{x} + F_Z^{-1}(\alpha/2) \left( \frac{\sigma_0}{\sqrt{n}} \right) < \mu < \bar{x} + F_Z^{-1}(1 - \alpha/2) \left( \frac{\sigma_0}{\sqrt{n}} \right)$$

en donde,  $Z_{\alpha/2}$  es el valor de la distribución normal estándar con un área derecha igual a  $\alpha/2$ . De la misma forma,  $F_Z^{-1}(\gamma)$  es el cuantil  $\gamma$  de la normal estándar para  $\gamma \in (0, 1)$ .

##### Ejemplo 3.8 Aplicación de los intervalos de confianza para el parámetro media

Una máquina de refrescos está ajustada de manera que la cantidad de líquido despachada tiene una distribución aproximadamente normal con una desviación estándar igual a 0.15 dl. Encuentre un intervalo de confianza de 95% para la verdadera media de los refrescos que sirve la máquina, si una muestra aleatoria de 36 tiene un contenido promedio de 2.25 dl.

**Solución**

Los datos proporcionados por el problema son:  $n = 36$ ,  $\sigma_0 = 0.15$  y  $\bar{x} = 2.25$  dl. El intervalo de confianza para el parámetro  $\mu$  se obtiene del teorema 3.1. Así, al calcular el valor de  $Z_{\alpha/2}$  con  $1 - \alpha = 0.95$ . De las tablas porcentuales para  $\alpha/2 = 0.025$ , la distribución normal estándar resulta  $Z_{0.025} = 1.960 = F_Z^{-1}(0.975)$ . Entonces:

$$2.25 - 1.96 \left( \frac{0.15}{\sqrt{36}} \right) < \mu < 2.25 + 1.96 \left( \frac{0.15}{\sqrt{36}} \right) \Rightarrow 2.201 < \mu < 2.299$$

Por tanto, se puede asegurar que el parámetro media del líquido despachado por la máquina de refrescos se encuentra entre 2.201 y 2.299 dl en 95% de los casos.

## Intervalos de confianza para medias de poblaciones normales o aproximadamente normales cuando se desconoce $\sigma$

**Teorema 3.2**

Sea  $x_1, x_2, \dots, x_n$  la realización tomada de una muestra aleatoria de variables con distribución normal con parámetros  $\mu$  y  $\sigma^2$  desconocida, entonces un intervalo de  $(1 - \alpha)100\%$  de confianza para el parámetro  $\mu$  está dado por:

$$\bar{x} - t_{\frac{\alpha}{2}} \left( \frac{s_{n-1}}{\sqrt{n}} \right) < \mu < \bar{x} + t_{\frac{\alpha}{2}} \left( \frac{s_{n-1}}{\sqrt{n}} \right) \text{ o}$$

$$\bar{x} + F_{t_{n-1}}^{-1}(\alpha/2) \left( \frac{s_{n-1}}{\sqrt{n}} \right) < \mu < \bar{x} + F_{t_{n-1}}^{-1}(1 - \alpha/2) \left( \frac{s_{n-1}}{\sqrt{n}} \right)$$

donde,  $t_{\alpha/2}$  es el valor de la distribución t-Student con  $\nu = n - 1$  grados de libertad, con un área derecha igual  $\alpha/2$ ,  $s_{n-1}$  es el valor de la varianza muestral calculada con la realización  $x_1, x_2, \dots, x_n$ ,  $F_{t_{n-1}}^{-1}(\gamma)$  es el cuantil  $\gamma$  de la t-Student con  $n - 1$  grados de libertad para  $\gamma \in (0, 1)$ .

1. Debido a que las tablas de la distribución t-Student en general están diseñadas para  $n \leq 30$ , en ocasiones el teorema anterior se restringe a muestras no mayores a 30, pero si las tablas tienen valores para grados de libertad mayores, no es necesaria la restricción.
2. En ocasiones suele simplificarse la notación de los intervalos de confianza con:

$$\bar{x} \pm Z_{\frac{\alpha}{2}} \left( \frac{\sigma_0}{\sqrt{n}} \right) \text{ teorema 3.1 y}$$

$$\bar{x} \pm t_{\frac{\alpha}{2}} \left( \frac{s_{n-1}}{\sqrt{n}} \right) \text{ teorema 3.2.}$$

**Ejemplo 3.9** Aplicación de los intervalos de confianza para los parámetros media y varianza

Un fabricante de máquinas despachadoras de refrescos asegura que sus productos sirven en promedio 240 ml en 99.9% de los casos. Un comprador decide verificar una de las máquinas, para esto toma una muestra aleatoria de 20 refrescos, de la que obtiene las siguientes medidas:

243	250	240	248	245	250	238	246	252	247
246	240	250	249	248	240	245	247	238	248

Si se supone normalidad en los datos y una confianza de 99.9%, determine si es válida la afirmación del fabricante.

**Solución**

Para encontrar el intervalo de confianza se necesita calcular la media y varianza insesgadas de los datos:

$$\bar{x} = 245.5 \text{ y } s_{n-1}^2 = 18.37, \text{ es decir, } s_{n-1} = 4.29$$

Como la muestra es de tamaño 20 y se desconoce la varianza poblacional tenemos que emplear el teorema 3.2. Para esto se calcula el valor  $t_{\alpha/2}$ , con  $\nu = 20 - 1 = 19$  grados de libertad y  $1 - \alpha = 0.999$ , de donde  $\alpha = 0.001$ , es decir  $\alpha/2 = 0.0005$ . Por tanto, de las *tablas porcentuales* para la distribución t-Student,  $t_{0.0005}(19) = 3.883$  o en caso de la acumulada  $F_{t_{19}}^{-1}(0.9995) = 3.883$ . Por último:

$$245.5 - 3.883 \left( \frac{4.29}{\sqrt{20}} \right) < \mu < 245.5 + 3.883 \left( \frac{4.29}{\sqrt{20}} \right) \Rightarrow 241.77 < \mu < 249.22$$

Es decir, el parámetro media del líquido despachado por la máquina de refrescos se encuentra entre 241.77 y 249.22 ml en 99.9 de los casos. Por tanto, la afirmación del fabricante no será válida con una confianza de 99.9%, pues el valor 240 ml quedó fuera del intervalo de confianza.

Como se puede apreciar, el uso del teorema 3.2 está limitado a las tablas de la distribución t-Student. Además, en general estas tablas están elaboradas para valores  $n \leq 30$ ; por consiguiente, surge la pregunta, ¿qué hacer cuando se desconoce  $\sigma$  y el tamaño de muestra es mayor a 30?

Cuando la muestra es grande, se puede utilizar el resultado que dice: La distribución t-Student se aproxima a la distribución normal cuando los grados de libertad son grandes. Es decir, cuando el tamaño de la muestra sea grande.

Luego, en caso de no contar con una tabla de la distribución t-Student para  $n > 30$  se puede usar la siguiente fórmula o una interpolación para los grados de libertad.

**Teorema 3.3**

Sea  $x_1, x_2, \dots, x_n$  una realización tomada de una muestra aleatoria de variables con distribución normal con parámetros  $\mu$  y  $\sigma^2$  desconocidos, entonces un intervalo de  $(1 - \alpha)$  100% de confianza para el parámetro  $\mu$  cuando  $n > 30$  está dado por:

$$\bar{x} - Z_{\frac{\alpha}{2}} \left( \frac{s_{n-1}}{\sqrt{n}} \right) < \mu < \bar{x} + Z_{\frac{\alpha}{2}} \left( \frac{s_{n-1}}{\sqrt{n}} \right) \text{ o } \bar{x} + F_Z^{-1}(\alpha/2) \left( \frac{s_{n-1}}{\sqrt{n}} \right) < \mu < \bar{x} + F_Z^{-1}(1 - \alpha/2) \left( \frac{s_{n-1}}{\sqrt{n}} \right)$$

donde  $Z_{\alpha/2}$  es el valor de la distribución normal estándar, a la derecha de ésta se tiene un área igual a  $\alpha/2$ ,  $F_Z^{-1}(\gamma)$  es el cuantil  $\gamma$  de la normal estándar para  $\gamma \in (0, 1)$ .

En el siguiente ejemplo se muestra una aplicación del teorema 3.3.

**Ejemplo 3.10 Intervalo de confianza**

Sea una máquina de refrescos como en el ejemplo 3.9 con  $\sigma$  desconocida. Para estimar la cantidad promedio de líquido despachada se sirven 50 vasos, de lo que se obtiene un promedio de 240 ml, con una desviación estándar de 20. Encuentre un intervalo de confianza de 99% para la cantidad promedio de líquido despachado por la máquina.



**Solución**

A partir de los datos  $\bar{x} = 240$  y  $s_{n-1} = 20$  ml, el intervalo de confianza del parámetro media se calcula al sustituir estos valores en la fórmula del teorema 3.3. Primero se busca el valor de  $Z_{\alpha/2}$ , con  $1 - \alpha = 0.99$ , de las *tablas porcentuales* para la distribución normal estándar  $Z_{\alpha/2} = 2.576$ . Por tanto:

$$240 - 2.576 \left( \frac{20}{\sqrt{50}} \right) < \mu < 240 + 2.576 \left( \frac{20}{\sqrt{50}} \right) \Rightarrow 232.71 < \mu < 247.29$$

Es decir, el parámetro media del líquido despachado por la máquina de refrescos se encuentra entre 232.71 y 247.29 ml en 99% de los casos.

**Ejemplos variados para la estimación de la media**

Se han visto los diferentes casos para llevar a cabo una estimación de la media por intervalos. Falta analizar ejemplos en donde se realicen ciertas modificaciones a las expresiones que intervienen en cada uno de los tres casos para la media.

Sea  $\theta$  un parámetro y  $\hat{\theta}$  su estimador, el **error en la estimación** se refiere a la separación que puede existir entre  $\theta$  y  $\hat{\theta}$ , sin importar que  $\hat{\theta}$  sea mayor o menor al parámetro. Por tanto, si denotamos por  $\varepsilon$  al error, tenemos que  $\varepsilon = |\theta - \hat{\theta}|$ .

De la definición anterior y las fórmulas para la estimación por intervalos es posible obtener el error por estimación en cada caso. Por ejemplo, cuando se conoce  $\sigma$ :

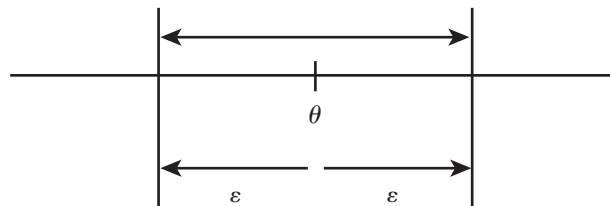
$$\bar{x} - Z_{\frac{\alpha}{2}} \left( \frac{\sigma_0}{\sqrt{n}} \right) < \mu < \bar{x} + Z_{\frac{\alpha}{2}} \left( \frac{\sigma_0}{\sqrt{n}} \right)$$

Al restar la media muestral resulta  $-Z_{\alpha/2} \sigma_0 / \sqrt{n} < \mu - \bar{x} < +Z_{\alpha/2} \sigma_0 / \sqrt{n}$ . Por definición de valor absoluto en una desigualdad, tenemos:

$$\varepsilon = |\mu - \bar{x}| < Z_{\frac{\alpha}{2}} \left( \frac{\sigma_0}{\sqrt{n}} \right)$$

Se puede representar con la figura 3.2.

Intervalo donde se puede localizar el estimador de  $\theta$



Longitud igual al tamaño del error a la izquierda y derecha del parámetro.

**Figura 3.2**

Ahora bien, ¿cómo encontrar un tamaño de muestra adecuado al error?

Conocidos el tamaño del error de estimación, la varianza poblacional y el grado de confianza o el valor de  $Z_{\alpha/2}$  se puede calcular el tamaño mínimo de muestra que satisfaga el error de estimación,

$$n \geq \left( \frac{Z_{\alpha/2}}{\varepsilon} \right)^2 \sigma_0^2 \text{ o } n = \left\lceil \left( \frac{Z_{\alpha/2}}{\varepsilon} \right)^2 \sigma_0^2 \right\rceil + 1, \text{ donde } [q] \text{ parte entera de } q.$$

**Caso I.** Se conoce  $\sigma_0$ ,  $|\mu - \bar{x}| < Z_{\frac{\alpha}{2}} \left( \frac{\sigma_0}{\sqrt{n}} \right)$ , luego  $\varepsilon = Z_{\frac{\alpha}{2}} \left( \frac{\sigma_0}{\sqrt{n}} \right)$  o  $n \geq \left( \frac{Z_{\alpha/2}}{\varepsilon} \right)^2 \sigma_0^2$

**Caso II.** Se desconoce grandes  $\sigma$ ,  $|\mu - \bar{x}| < t_{\frac{\alpha}{2}} \left( \frac{s_{n-1}}{\sqrt{n}} \right)$ , luego  $\varepsilon = t_{\frac{\alpha}{2}} \left( \frac{s_{n-1}}{\sqrt{n}} \right)$

**Caso III.** Se desconoce  $\sigma$ , muestras  $|\mu - \bar{x}| < Z_{\frac{\alpha}{2}} \left( \frac{s_{n-1}}{\sqrt{n}} \right)$ , luego  $n \geq \left( \frac{Z_{\alpha/2}}{\varepsilon} \right)^2 s_{n-1}^2$

Los errores por estimación,  $\varepsilon$ , suelen utilizarse para calcular el tamaño mínimo de la muestra necesario para que la media muestral se desvíe de la poblacional en un valor menor o igual a  $\varepsilon$ . Estas fórmulas coinciden con las obtenidas en la unidad 2 con  $\varepsilon = |\mu - \bar{x}|$ .

En caso de querer calcular el tamaño de la muestra por medio del error por estimación, tanto en el caso II como en el III se tiene el problema de conocer el valor de  $s_{n-1}$ . Para esto requerimos el tamaño de la muestra. En el caso III, el problema se puede solucionar al tomar una estimación de  $s_{n-1}$ , con alguna otra realización de un tamaño determinado, pero el caso II no se puede utilizar ya que requiere también del tamaño de la muestra para calcular el valor  $t_{\alpha/2}$ .

### Ejemplo 3.11 Tamaño mínimo de la muestra

Una máquina de refresco está ajustada de manera que la cantidad de líquido despachada tiene una distribución aproximadamente normal con una desviación estándar igual a 0.15 dl. Encuentre el tamaño mínimo de la muestra que se necesita elegir para que la media muestral se desvíe de la media poblacional en menos de 0.035 dl con una confianza de 95%.

#### Solución

De los datos del enunciado se obtiene que  $\sigma_0 = 0.15$  y  $\varepsilon = |\mu - \bar{x}| < 0.035$  dl. Con base en la fórmula del tamaño del error para el caso I (se conoce  $\sigma_0 = 0.15$ ):

$$|\mu - \bar{x}| < Z_{\frac{\alpha}{2}} \left( \frac{\sigma_0}{\sqrt{n}} \right), \text{ entonces } \varepsilon = Z_{\frac{\alpha}{2}} \left( \frac{\sigma_0}{\sqrt{n}} \right)$$

Falta calcular el valor de  $Z_{\alpha/2}$  con  $1 - \alpha = 0.95$ . De las *tablas porcentuales* para la distribución normal estándar se tiene que  $Z_{\alpha/2} = 1.960$ . Por tanto, al despejar el tamaño de la muestra,

$$n = \left\lceil \left( \frac{Z_{\alpha/2}}{\varepsilon} \right)^2 \sigma_0^2 \right\rceil + 1 = \left\lceil \left( \frac{1.96}{0.035} \right)^2 (0.15)^2 \right\rceil = 70.56$$

Es decir, el tamaño mínimo de muestra que satisface las condiciones del problema es 70.

En este sentido, ¿cómo calcular el tamaño de muestra cuando se da la longitud del intervalo?

De la definición del error por estimación se aprecia que la longitud del intervalo de confianza es igual a:

$$\text{longitud del intervalo} = 2\varepsilon$$

Esto se verifica con facilidad, pues la longitud de un intervalo es igual al límite superior menos el inferior, de manera que a partir del intervalo:

$$\bar{x} - Z_{\frac{\alpha}{2}} \left( \frac{\sigma_0}{\sqrt{n}} \right) < \mu < \bar{x} + Z_{\frac{\alpha}{2}} \left( \frac{\sigma_0}{\sqrt{n}} \right)$$

se restan los límites, de lo que se obtiene:

$$\bar{x} + Z_{\frac{\alpha}{2}} \left( \frac{\sigma_0}{\sqrt{n}} \right) - \left( \bar{x} - Z_{\frac{\alpha}{2}} \left( \frac{\sigma_0}{\sqrt{n}} \right) \right) = 2Z_{\frac{\alpha}{2}} \left( \frac{\sigma_0}{\sqrt{n}} \right) = 2\varepsilon$$

### Ejemplo 3.12 Tamaño de la muestra

Una máquina de refresco está ajustada de manera que la cantidad de líquido despachado se distribuye de manera aproximadamente normal con una desviación estándar igual a 0.15 dl. ¿Qué tan grande debe ser la muestra requerida, si se desea tener una confianza de 98% que la estimación media estará dentro de un intervalo de longitud 0.10 dl?

#### Solución

De los datos del enunciado se obtiene que  $\sigma_0 = 0.15$  y  $2\varepsilon = 0.10$  dl, donde  $\varepsilon = 0.05$ . Luego, al utilizar la fórmula anterior para el tamaño de muestra con  $Z_{\alpha/2}$ , para  $1 - \alpha = 0.98$ . De las *tablas porcentuales* para la distribución normal estándar se tiene  $Z_{\alpha/2} = 2.3263$ . Por tanto:

$$n = \left[ \left( \frac{Z_{\alpha/2}}{\varepsilon} \right)^2 \sigma_0^2 \right] + 1 = \left[ \left( \frac{2.3263}{0.05} \right)^2 (0.15)^2 \right] + 1 = [48.71] + 1 = 49$$

Es decir, el tamaño mínimo de muestra que satisface las condiciones del problema es 49.

En este sentido, ¿cómo determinar el grado de confianza cuando se tienen los límites del intervalo?

En ocasiones, en la práctica el investigador determina los límites que quisiera tener para acotar al parámetro, pero no sabe cómo encontrar el grado de confianza apropiado que satisfaga sus necesidades y reportarlo en su investigación. En estos casos se usa el hecho de que  $\bar{x}$  siempre es el valor medio del intervalo de confianza para el parámetro.

### Teorema 3.4

Sea  $x_1, x_2, \dots, x_n$  una realización de la muestra aleatoria  $X_1, X_2, \dots, X_n$  de una población normal con parámetros  $\mu$  y  $\sigma^2$  (conocida) y  $a < \mu < b$  un intervalo del  $(1 - \alpha)100\%$  de confianza para el parámetro  $\mu$ , entonces:

$$1 - \alpha = D \left( \left( \frac{\bar{x} - a}{\sigma_0} \right) \sqrt{n} \right) = D \left( \left( \frac{b - \bar{x}}{\sigma_0} \right) \sqrt{n} \right) = 1 - 2\Phi \left( \left( \frac{\bar{x} - b}{\sigma_0} \right) \sqrt{n} \right) = 2\Phi \left( \left( \frac{\bar{x} - a}{\sigma_0} \right) \sqrt{n} \right) - 1$$

donde  $\Phi$  representa la distribución acumulada de la normal estándar y  $D$  la distribución acumulada simétrica de la normal estándar; ambas se encuentran en las tablas de la normal.

### Ejemplo 3.13 Grado de confianza cuando se tienen los límites del intervalo

Una máquina de refrescos está ajustada de manera que la cantidad de líquido despachado tiene una distribución aproximadamente normal con una desviación estándar de 15 ml. Se elige una muestra de tamaño 50, de la que resulta que la cantidad de líquido promedio es de 245 ml. Si el administrador de la empresa quiere

entregar un reporte de una estimación por intervalos para el líquido promedio despachado, en la que el límite superior del intervalo sea de 250 ml, ¿cuál será el límite inferior del reporte y con qué grado de confianza justificaría su reporte?

### Solución

De los datos del enunciado, tenemos  $\sigma_0 = 15$ ,  $n = 50$ ,  $\bar{x} = 245$  y el límite superior del intervalo  $b = 250$ . Para determinar el valor del límite inferior se puede utilizar  $\bar{x} - a = b - \bar{x} = 250 - 245 = 5$  y  $a = 245 - 5 = 240$ .

Para el grado de confianza se usa alguna de las cuatro igualdades del teorema 3.4. Por ejemplo:

$$1 - \alpha = D\left(\frac{\bar{x} - a}{\sigma_0} \sqrt{n}\right) = D\left(\frac{245 - 240}{15} \sqrt{50}\right) = D(2.36) = 0.9817$$

Es decir, el administrador puede reportar con un grado de confianza de 98.17% que la máquina despacha entre 240 y 250 ml de refresco.

En caso de que se desconozca  $\sigma$ , en el teorema 3.4 se cambia  $\Phi$  por la acumulada de t-Student.

Ahora surge la pregunta, ¿cómo calcular un intervalo de confianza cuando se tiene una tabla de frecuencias?

Se hace de la misma forma que en los casos anteriores, se emplean las fórmulas correspondientes de la unidad 1 para la media y varianza muestrales.

### Ejemplo 3.14 Cálculo de un intervalo de confianza cuando hay una tabla de frecuencias

Una máquina de refrescos está ajustada de manera que la cantidad de líquido despachado tiene una distribución aproximadamente normal, con una desviación estándar de 15 ml. Se elige una muestra de tamaño 60 y un trabajador registra el líquido despachado por clases de frecuencia, que da como resultado la tabla 3.2.

Usando un coeficiente de confianza de 99%, estime el promedio real  $\mu$ .

Tabla 3.2

Intervalos de clase	Frecuencias ( $n_i$ )
[239, 241)	4
[241, 243)	10
[243, 245)	20
[245, 247)	11
[247, 249)	12
[249, 251]	3

### Solución

De los datos del enunciado tenemos  $\sigma_0 = 15$  y  $n = 60$ . Por tanto, la fórmula para el intervalo de confianza está dada por:

$$\bar{x}_f - Z_{\frac{\alpha}{2}} \left( \frac{\sigma_0}{\sqrt{n}} \right) < \mu < \bar{x}_f + Z_{\frac{\alpha}{2}} \left( \frac{\sigma_0}{\sqrt{n}} \right)$$

Al calcular la media por clases de frecuencias  $\bar{x}_f = \frac{1}{n} \sum_{i=1}^m x_i^m n_i = \frac{1}{60} \sum_{i=1}^6 x_i^m n_i$  faltan las marcas de clase de la tabla anterior 240, 242, 244, 246, 248 y 250. De esta manera:

$$\bar{x}_f = \frac{1}{60} (240 \times 4 + 242 \times 10 + 244 \times 20 + 246 \times 11 + 248 \times 12 + 250 \times 3) = 244.8667$$

Para el valor de  $Z_{\alpha/2}$ , tenemos  $1 - \alpha = 0.99$ , luego  $\alpha/2 = 0.005$ . De las tablas porcentuales para la distribución normal estándar  $Z_{\alpha/2} = 2.5758$ . Al final, el intervalo queda:

$$244.8667 - 2.5758 \left( \frac{15}{\sqrt{60}} \right) < \mu < 244.8667 + 2.5758 \left( \frac{15}{\sqrt{60}} \right) \Rightarrow 239.8787 < \mu < 249.8547$$

Es decir, con un grado de confianza de 99%, la máquina despacha entre 239.88 y 249.85 ml de refresco.

## Intervalos de confianza para la varianza de poblaciones normales

La distribución normal tiene dos parámetros de los cuales se estudió la media, ahora se analizará la varianza y quedarán resueltos ambos parámetros de una población normal. La fórmula que utilizaremos para los intervalos de confianza de la varianza no cumple con una amplitud mínima, pero da una buena aproximación. La construcción de los intervalos de confianza para la varianza con amplitud mínima resulta bastante compleja y depende de cada realización, por estas razones en la práctica no se usa.

Para estudiar el parámetro  $\sigma^2$  recurriremos al hecho de que  $\frac{(n-1)s_{n-1}^2}{\sigma^2}$  tiene una distribución  $\chi_{n-1}^2$  con  $n - 1$  grados de libertad, de manera que resulta el teorema 3.5.

### Teorema 3.5

Si  $x_1, x_2, \dots, x_n$  es la realización de una muestra aleatoria de tamaño  $n$ , tomada de una población normal con parámetros  $\mu$  y  $\sigma^2$ , y  $s_{n-1}^2$  es el valor de la varianza muestral, entonces un intervalo de confianza de  $(1 - \alpha)100\%$  para  $\sigma^2$  está dado por:

$$\frac{(n-1)s_{n-1}^2}{\chi_{\alpha/2}^2} < \sigma^2 < \frac{(n-1)s_{n-1}^2}{\chi_{1-\alpha/2}^2},$$

donde  $\chi_{\alpha/2, n-1}^2$  y  $\chi_{1-\alpha/2, n-1}^2$  son valores de la distribución ji cuadrada  $\chi^2$  con  $\nu = n - 1$  grado de libertad, con áreas a la derecha de  $\alpha/2$  y  $1 - \alpha/2$ , respectivamente.

1. El intervalo de confianza encontrado no siempre es el de amplitud esperada mínima, pero da una buena aproximación.

Debido a que las tablas de la distribución ji cuadrada suelen ser para  $n \leq 30$ ; en ocasiones, el teorema anterior se restringe para este tipo de muestras, pero en caso de tener tablas para valores de grados de libertad mayores no es necesaria la restricción.

En el siguiente ejemplo se realizan los cálculos para ilustrar el teorema 3.5.

### Ejemplo 3.15 Valor de la varianza muestral

Un antropólogo midió el ancho (en centímetros) de una muestra aleatoria de nueve cráneos de los miembros de cierta tribu cuyos resultados fueron 13.3, 14.2, 13.5, 16.7, 11.1, 13.1, 13.0, 12.2, 13.0. Estime un intervalo de confianza de 95% para la varianza poblacional de dicha tribu.

**Solución**

Primero se calcula la variancia insesgada de los datos  $s_{n-1}^2 = 2.33$ . Por otro lado, el coeficiente de confianza está dado por  $1 - \alpha = 0.95$ , de donde  $\alpha = 0.05$ , es decir  $\alpha/2 = 0.025$  y  $1 - \alpha/2 = 0.975$ . Al buscar en las tablas de la distribución *ji* cuadrada con  $\nu = 9 - 1 = 8$  grados de libertad, resulta:

$$\chi_{\alpha/2, n-1}^2 = \chi_{0.025, 8}^2 = 17.5345 \text{ y } \chi_{1-\alpha/2, n-1}^2 = \chi_{0.975, 8}^2 = 2.1797$$

Por último, se aplica la fórmula del teorema 3.5:

$$\frac{(9-1)2.33}{17.5345} < \sigma^2 < \frac{(9-1)2.33}{2.1797} \Rightarrow 1.06 < \sigma^2 < 8.55$$

Es decir, con una probabilidad de 0.95 se asegura que el parámetro  $\sigma^2$  del ancho de los cráneos de la tribu se encuentra entre 1.06 y 8.55 cm.

**Ejemplos variados para varianzas**

A diferencia de los casos para medias y diferencia de medias, donde se usaban las distribuciones *Z* y *t-Student* que son simétricas, la distribución para la varianza *ji* cuadrada no lo es. La simetría facilitó el estudio sobre el tamaño mínimo de la muestra y el error por estimación, ya que las medias muestrales eran el centro del intervalo de confianza.

Con base en lo anterior, se puede utilizar el resultado del teorema 3.5 para situaciones en que se quiera fijar uno de los extremos del intervalo y se desee conocer el grado de confianza con el que se obtiene; en esta situación, como consecuencia directa, es posible encontrar el otro extremo del intervalo de confianza.

**Teorema 3.6**

Si  $x_1, x_2, \dots, x_n$  es una realización de una muestra aleatoria de tamaño  $n$ , tomada de una población normal con parámetros  $\mu$  y  $\sigma^2$ , y  $a < \sigma^2 < b$  un intervalo de  $(1 - \alpha)100\%$  de confianza para el parámetro  $\sigma^2$ , entonces:

$$1 - \alpha = 1 - 2\chi_{n-1} \left( \frac{(n-1)s_{n-1}^2}{a} \right) = 2\chi_{n-1} \left( \frac{(n-1)s_{n-1}^2}{b} \right) - 1$$

donde  $\chi_{n-1}(q)$  representa probabilidades de la distribución  $\chi^2$  con  $\nu = n - 1$  grados de libertad, a la derecha del valor  $q$ . Es decir,  $\chi_{n-1}(q) = 1 - F_{\chi_{n-1}^2}(q)$ ,  $F$  distribución acumulada.

En el ejemplo 3.16 se observa una aplicación del teorema 3.6.

**Ejemplo 3.16 Intervalo de confianza**

Con el fin de estimar la variancia de una población normal se tomó una muestra aleatoria de tamaño 22, con  $s^2 = 0.3486$ . Suponga que por ciertas circunstancias de la investigación se desea que el límite inferior del intervalo de confianza sea 0.2241.

- ¿Qué grado de confianza se tiene que anotar en el reporte de la investigación?
- ¿Cuál es su límite superior?

**Solución**

- Tenemos los datos  $s^2 = 0.3486$ ,  $n = 22$  y  $\alpha = 0.2241$ , de la fórmula del teorema 3.6:

$$1 - \alpha = 1 - 2\chi_{n-1} \left( \frac{(n-1)s_{n-1}^2}{\alpha} \right) = 1 - 2\chi_{n-1} \left( \frac{(22-1)0.3486}{0.2241} \right) = 1 - 2\chi_{n-1}(32.667) = 1 - 2(0.05) = 0.90$$

Luego, el coeficiente de confianza utilizado  $1 - \alpha = 0.90$ .

b) Para el límite superior, primero obtenemos de las tablas porcentuales,  $\chi_{1-\alpha/2, 21}^2 = 11.591$ , (parte derecha de 0.95), luego el límite superior  $(n-1)s_{n-1}^2 / \chi_{1-\alpha/2, 21}^2 = 21(0.3486) / 11.5913 = 0.6316$ .

El valor  $\chi_{n-1}(32.667) = 0.05$  se buscó en las tablas porcentuales  $\chi^2$  en el renglón de  $\nu = n - 1 = 21$  g.l., hasta encontrar 32.667. En caso de no estar el valor se puede hacer una interpolación.

### Ejemplo 3.17 Intervalo de confianza

Se registraron las siguientes mediciones de las horas de secado de una marca de pintura látex.

3.4	2.5	4.8	2.9	3.6
2.6	3.3	5.6	3.7	2.8
4.4	4.0	5.2	3.0	4.8

Suponga que las mediciones representan una muestra aleatoria. Calcule el intervalo de confianza para la desviación estándar con 90% de confianza.

#### Solución

De la fórmula del teorema 3.5:

$$\frac{(n-1)s_{n-1}^2}{\chi_{\alpha/2, n-1}^2} < \sigma^2 < \frac{(n-1)s_{n-1}^2}{\chi_{1-\alpha/2, n-1}^2}$$

Luego, se extrae raíz cuadrada para obtener la desviación estándar:

$$s_{n-1} \sqrt{\frac{n-1}{\chi_{\alpha/2, n-1}^2}} < \sigma < s_{n-1} \sqrt{\frac{n-1}{\chi_{1-\alpha/2, n-1}^2}}$$

De los datos del problema se calcula la desviación estándar, resulta  $s_{n-1} = 0.9867$ .

Para los valores de la *ji* cuadrada se tiene el grado de confianza  $1 - \alpha = 0.90$ , de donde  $\alpha = 0.10$ . Luego,  $\alpha/2 = 0.05$  y los grados de libertad  $\nu = n - 1 = 15 - 1 = 14$ . Ahora, de las tablas porcentuales de la distribución *ji* cuadrada tendremos:

$$\chi_{\alpha/2, n-1}^2 = \chi_{0.05, 14}^2 = 23.6848 \text{ y } \chi_{1-\alpha/2, n-1}^2 = \chi_{0.95, 14}^2 = 6.5706$$

Por último, el intervalo de confianza para la desviación estándar:

$$0.9867 \sqrt{\frac{14}{23.6848}} < \sigma < 0.9867 \sqrt{\frac{14}{6.5706}} \Rightarrow 0.7586 < \sigma < 1.4403$$

## Ejercicios 3.5

- Una máquina produce piezas metálicas de forma cilíndrica. Se toma una muestra de piezas cuyos diámetros son 10, 12, 11, 11.5, 9, 9.8, 10.4, 9.8, 10 y 9.8 ml. Suponga que los diámetros tienen una distribución aproximadamente normal. Con 99% de confianza:
  - Construya un intervalo de confianza para el diámetro promedio de todas las piezas de esta máquina, suponga que  $\sigma = 1$ .
  - Determine el tamaño mínimo de la muestra que debe elegirse para que el error de los diámetros sea menor a un cuarto de milímetro.
  - Si el límite inferior del intervalo de confianza es 9.75 ml, ¿cuál es el límite superior y el nivel de confianza?
  - Construya un intervalo de confianza para el diámetro promedio de todas las piezas de esta máquina, si no se conoce  $\sigma$ .

- Del ejercicio anterior con 99% de confianza:

- Construya un intervalo de confianza para la desviación estándar y decida si fue válida la suposición de que  $\sigma = 1$ .
- Si el límite inferior del intervalo de confianza para  $\sigma$  es 0.699 mm, ¿cuál es el límite superior y el nivel de confianza?

- La cámara de comercio de una ciudad asegura que según sus estudios económicos, la cantidad promedio de dinero que gasta al día la gente que asiste a convenciones, que incluye comidas, alojamiento y entretenimiento, es menor a \$950. Para probar esta afirmación un supervisor de la cámara seleccionó 16 personas que asisten a convenciones y les preguntó qué cantidad de dinero gastaban por día, de lo que obtuvo la siguiente información (en pesos):

940	875	863	948	942	989	835	874	868	852	958	884	1034	1046	955	963
-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	------	------	-----	-----

Suponga que la cantidad de dinero gastada en un día se distribuye de manera normal. La gerente decide que si el límite superior del intervalo de confianza para la media es menor a \$960, entonces será válida su afirmación. Con 95% de confianza:

- Construya un intervalo de confianza para la media de todos los gastos diarios, suponga que  $\sigma = \$45$ , ¿es válida la afirmación?
  - Determine el tamaño mínimo de la muestra que debe elegirse para que la estimación media esté dentro de un intervalo de longitud \$40.
  - Si el límite inferior del intervalo de confianza es \$900, ¿cuál es el límite superior y el nivel de confianza?
  - Construya un intervalo de confianza para la media de todos los gastos diarios, si no se conoce  $\sigma$ , ¿es válida la afirmación?
- Del ejercicio anterior con 95% de confianza:
    - Construya un intervalo de confianza para la desviación estándar y decida si fue válida la suposición de que  $\sigma = \$45$ .
    - Si el límite superior del intervalo de confianza para  $\sigma$  es 70 pesos, ¿cuál es el límite inferior y el nivel de confianza?
  - Un geólogo que pretendía estudiar el movimiento de los cambios relativos en la corteza terrestre en un sitio particular, en un intento por determinar el ángulo medio de las fracturas eligió  $n = 51$  fracturas y encontró que la media era  $39.8^\circ$ , si la desviación estándar muestral es  $17.2^\circ$ . Si supone normalidad en la población, con 90% de confianza:
    - Construya un intervalo de confianza para la media de ángulo de fractura en la corteza terrestre, suponga que  $\sigma = 19^\circ$ .
    - Determine el tamaño mínimo de la muestra que debe elegirse para que el error de la estimación media de las fracturas sea menor a  $3^\circ$ .
    - Si el límite inferior del intervalo de confianza es  $35.2246^\circ$ , ¿cuál es el límite superior y el nivel de confianza?
    - Construya un intervalo de confianza para la media de ángulo de fractura en la corteza terrestre, si no se conoce  $\sigma$ .
  - Del ejercicio anterior con 90% de confianza:



- a) Construya un intervalo de confianza para la desviación estándar y decida si fue válida la suposición de que  $\sigma = 19^\circ$ .
- b) Si el límite superior del intervalo de confianza para  $\sigma$  es  $21^\circ$ , ¿cuál es el límite inferior y el nivel de confianza?
7. El espesor de las paredes de 25 botellas de vidrio de 2 l fue medido por un supervisor de control de calidad. La media muestral fue de 4.02 mm, y la desviación estándar muestral de 0.5 mm. Suponga normalidad en la distribución del espesor de las paredes de las botellas de vidrio de 2 l. Con 96% de confianza:
- a) Construya un intervalo de confianza para la media del espesor de las paredes de las botellas de vidrio, suponga que  $\sigma = 0.4$  mm.
- b) Determine el tamaño mínimo de la muestra que debe elegirse para que el error de la estimación media del espesor de las paredes sea menor a 0.20 mm.
- c) Si el límite inferior del intervalo de confianza es 3 mm, ¿cuál es el límite superior y el nivel de confianza?
- d) Construya un intervalo de confianza para la media del espesor de las paredes de las botellas de vidrio, si no se conoce  $\sigma$ .
8. Del ejercicio anterior con 96% de confianza:
- a) Construya un intervalo de confianza para la desviación estándar y decida si fue válida la suposición del inciso a) de que  $\sigma = 0.4$  mm.
- b) Si el límite inferior del intervalo de confianza para  $\sigma$  es 0.4 mm, ¿cuál es el límite superior y el nivel de confianza?
9. Los clientes que llegan a una gasolinera se han quejado ante la Profeco de que reciben menos gasolina que la marcada en el medidor. La Profeco manda un supervisor a verificar la bomba. En la bomba de la estación de gasolina el supervisor realiza 10 mediciones en garrafones de 20 l: 20.5, 19.99, 20.0, 20.3, 19.90, 20.05, 19.79, 19.85, 19.95 y 20.15. Si supone normalidad en las mediciones de 20 l. Con 95% de confianza:
- a) Construya un intervalo de confianza para la media del contenido de gasolina de los garrafones, suponga que  $\sigma = 0.14$  l.
- b) Determine el tamaño mínimo de la muestra que debe elegirse para que la estimación media esté dentro de un intervalo de longitud 0.09 l.
- c) Si el límite superior del intervalo de confianza es 20.3 l, ¿cuál es el límite inferior y el nivel de confianza?
- d) Construya un intervalo de confianza para la media del contenido de gasolina de los garrafones, si no se conoce  $\sigma$ .
10. Del ejercicio anterior con 95% de confianza:
- a) Construya un intervalo de confianza para la desviación estándar y decida si fue válida la suposición de que  $\sigma = 0.14$ .
- b) Si el límite superior del intervalo de confianza para  $\sigma$  es 0.3 l, ¿cuál es el límite inferior y el nivel de confianza?
11. El IPC de la empresa WM se muestra en la tabla 3.3 y se supone que tiene una distribución normal durante el año. Con 99% de confianza:
- a) Construya un intervalo de confianza para el IPC medio, suponga que  $\sigma = 1.1$ .
- b) Determine el tamaño mínimo de la muestra que debe elegirse para que la estimación media esté dentro de un intervalo de longitud 0.10.
- c) Si el límite superior del intervalo de confianza es 36.5, ¿cuál es el límite inferior y el nivel de confianza?
- d) Construya un intervalo de confianza para el IPC medio, si no se conoce  $\sigma$ .

Tabla 3.3

Fecha	WM
09/06/2013	37.10
09/03/2013	36.99
09/02/2013	37.83
09/01/2013	36.36
08/31/2013	36.17
08/30/2013	35.98

Fecha	WM
08/27/2013	35.87
08/26/2013	35.68
08/25/2013	35.92
08/24/2013	35.91
08/23/2013	35.29
08/20/2013	34.86
08/19/2013	34.83

12. Del ejercicio anterior con 99% de confianza:
- Construya un intervalo de confianza para la varianza y decida si fue válida la suposición de que  $\sigma = 1.1$ .
  - Si el límite inferior del intervalo de confianza para  $\sigma$  es 0.6527, ¿cuál es el límite superior y el nivel de confianza?
13. Mientras efectúan una tarea determinada en condiciones simuladas de ausencia de gravedad, el ritmo cardiaco de 31 astronautas en adiestramiento se incrementa en un promedio de 26.4 pulsaciones por minuto con una desviación estándar de 4.28 pulsaciones por minuto. Suponga que en estas condiciones el ritmo cardiaco de los astronautas tiene una distribución normal. Con 95% de confianza:
- Construya un intervalo de confianza para el ritmo cardiaco medio de los astronautas en adiestramiento, suponga que  $\sigma = 3$ .
  - Determine el tamaño mínimo de la muestra que debe elegirse para que el error de la estimación media del ritmo cardiaco sea menor a 1.5 pulsaciones por minuto.
  - Si el límite inferior del intervalo de confianza es 25, ¿cuál es el límite superior y el nivel de confianza?
  - Construya un intervalo de confianza para el ritmo cardiaco medio de los astronautas en adiestramiento, si no se conoce  $\sigma$ .
14. Del ejercicio anterior con 95% de confianza:
- Construya un intervalo de confianza para la varianza y decida si fue válida la suposición de que  $\sigma = 1$ .
  - Si el límite superior del intervalo de confianza para  $\sigma$  es 6, ¿cuál es el límite inferior y el nivel de confianza?
15. Los resultados de una investigación se muestran en la distribución de frecuencias de la tabla 3.4. Si supone que se obtuvieron de una población aproximadamente normal, con 98% de confianza:
- Construya un intervalo de confianza para la media de los resultados de la investigación, suponga que  $\sigma = 2$ .
  - Determine el tamaño mínimo de la muestra que debe elegirse para que el error de la estimación media sea menor a 0.5.
  - Si el límite inferior del intervalo de confianza es 4.5, ¿cuál es el límite superior y el nivel de confianza?
  - Construya un intervalo de confianza para la media de los resultados de la investigación, si no se conoce  $\sigma$ .

Tabla 3.4

Intervalos de clase	Frecuencias
[0.15, 1.55]	2
(1.55, 2.95]	6
(2.95, 4.35]	11
(4.35, 5.75]	17
(5.75, 7.15]	10
(7.15, 8.55]	3
(8.55, 9.95]	2

16. Del ejercicio anterior con 98% de confianza:
- Construya un intervalo de confianza para la varianza y decida si fue válida la suposición de que  $\sigma = 2$ .
  - Si el límite superior del intervalo de confianza para  $\sigma$  es 2.3, ¿cuál es el límite inferior y el nivel de confianza?

### 3.4 Intervalos de confianza para comparar dos poblaciones normales

En la sección anterior estudiamos los parámetros de una distribución normal, la media y la varianza. Ahora el estudio de dichos parámetros se puede extender a los casos en que requerimos comparar dos poblaciones normales. Por ejemplo, para contrastar los rendimientos medios de dos líneas o procesos de producción pensaríamos en tomar una muestra de cada población y estudiar su diferencia de medias.

Al comparar dos poblaciones existen más de dos casos para inferir sobre la diferencia de medias, veremos que existen en total cinco. Cabe mencionar que las fórmulas que utilizaremos para los intervalos de confianza de la diferencia de medias cumplen con las condiciones de un buen estimador por intervalos de confianza; es decir, su amplitud es mínima.

Antes de construir los intervalos de confianza para la diferencia de medias, hay que hacer énfasis en que ésta tiene como objetivo primordial la comparación de medias entre dos poblaciones. Luego, es necesario establecer los diferentes tipos de comparación más comunes.

## Resultados posibles de las comparaciones entre dos medias

Suponga que buscamos un intervalo con un coeficiente de confianza del  $1 - \alpha$  para  $\mu_1 - \mu_2$  y resulta  $(a, b)$ , entonces podemos tener alguno de los tres casos siguientes:

1. Si  $a$  y  $b$  son ambos positivos, entonces al  $1 - \alpha$  de confianza se puede suponer  $\mu_1 > \mu_2$ .  
 Por ejemplo,  $\mu_1 - \mu_2 \in (2, 8)$ , entonces se puede suponer  $\mu_1 > \mu_2$  al nivel mínimo  $1 - \alpha$ .
  - a) Si se agrega que  $\mu_1$  aventaja a  $\mu_2$  en al menos (o más de)  $k$ , la aseveración  $\mu_1 \geq \mu_2 + k$ . El problema se resuelve de la misma forma. Si  $\mu_1 - \mu_2$ , se obtiene  $\mu_1 - \mu_2 \in (a, b)$ , entonces la aseveración  $\mu_1 \geq \mu_2 + k$  se cumplirá al  $1 - \alpha$  que  $k < b$ . Por ejemplo, si la aseveración es  $\mu_1 \geq \mu_2 + 4$  y resultó  $\mu_1 - \mu_2 \in (2, 8)$ , entonces al nivel de confianza mínimo de  $1 - \alpha$  se puede decir que la aseveración  $\mu_1 \geq \mu_2 + 4$  es válida,  $4 < 8$ . Si la aseveración es  $\mu_1 \geq \mu_2 + 10$  y resultó  $\mu_1 - \mu_2 \in (2, 8)$ , entonces al nivel de confianza  $1 - \alpha$  no se puede decir que la aseveración  $\mu_1 \geq \mu_2 + 10$  es válida, ya que  $10 > 8$ . Para que se cumpla tenemos que aumentar el nivel de significancia  $1 - \alpha$ .
  - b) Si además se agrega que  $\mu_1$  aventaja a  $\mu_2$  en máximo (o menos de)  $k$ , la aseveración  $\mu_2 < \mu_1 \leq \mu_2 + k$ . El problema se resuelve de la misma forma. Si  $\mu_1 - \mu_2$ , obtiene  $\mu_1 - \mu_2 \in (a, b)$ , entonces la aseveración  $\mu_2 < \mu_1 \leq \mu_2 + k$  se cumplirá si  $(0, k) \subset (a, b)$ . Por ejemplo, si la aseveración es  $\mu_2 < \mu_1 \leq \mu_2 + 4$  y resultó  $\mu_1 - \mu_2 \in (2, 8)$ , entonces al nivel de confianza de  $1 - \alpha$  no es posible decir que la aseveración  $\mu_2 < \mu_1 \leq \mu_2 + 4$  es válida, porque no hay seguridad para situaciones entre  $(0, 2)$ , hay que aumentar  $1 - \alpha$ .
  - c) Si además se agrega que  $\mu_1$  aventaja a  $\mu_2$  en  $k$ , la aseveración  $\mu_1 = \mu_2 + k$ , el problema se resuelve de la misma forma. Si  $\mu_1 - \mu_2$ , y se obtiene  $\mu_1 - \mu_2 \in (a, b)$ , entonces la aseveración  $\mu_1 = \mu_2 + k$  se cumplirá si  $k \in (a, b)$ . Por ejemplo, si la aseveración es  $\mu_1 = \mu_2 + 4$  y resultó  $\mu_1 - \mu_2 \in (2, 8)$ , entonces al nivel de confianza mínimo de  $1 - \alpha$  se puede decir que la aseveración  $\mu_1 = \mu_2 + 4$  es válida. Ahora si la aseveración es  $\mu_1 = \mu_2 + 1$  y resultó  $\mu_1 - \mu_2 \in (2, 8)$ , entonces al nivel de confianza de  $1 - \alpha$  no es posible decir que la aseveración  $\mu_1 = \mu_2 + 1$  es válida  $1 \notin (2, 8)$ .
2. Si  $a$  y  $b$  son negativos, entonces al  $1 - \alpha$  de confianza se puede suponer  $\mu_1 < \mu_2$ . Por ejemplo,  $\mu_1 - \mu_2 \in (-8, -2)$ , entonces se puede suponer  $\mu_1 < \mu_2$  al nivel de confianza  $1 - \alpha$ .  
 Los otros casos se analizan de la misma forma que en 1), o solo invirtiendo la diferencia de medias. En el ejemplo anterior  $\mu_1 - \mu_2 \in (2, 8)$  y estamos en el inciso a).
3. Si  $a$  es negativo y  $b$  positivo, al  $1 - \alpha$  de confianza se puede suponer  $\mu_1 = \mu_2$ . Por ejemplo,  $\mu_1 - \mu_2 \in (2, 8)$ , entonces se puede suponer  $\mu_1 = \mu_2$  al nivel mínimo  $1 - \alpha$ . Las demás situaciones se concluyen de la misma forma que a).

## Intervalos de confianza para la diferencia de medias, poblaciones aproximadamente normales cuando se conocen $\sigma_1$ y $\sigma_2$

De la misma forma que en el caso de una población normal, el problema de la comparación de medias inicia para el caso cuando se conocen las varianzas poblacionales, como se muestra en el teorema 3.7.

### Teorema 3.7

Si  $\bar{x}_1$  y  $\bar{x}_2$  son las medias aritméticas de las realizaciones de dos muestras aleatorias independientes de tamaños  $n_1$  y  $n_2$ , respectivamente, de poblaciones con distribuciones aproximadamente normales de las que se conoce  $\sigma_1^2$  y  $\sigma_2^2$ , entonces un intervalo de confianza de  $(1 - \alpha)100\%$  para  $\mu_1 - \mu_2$  está dado por:

$$(\bar{x}_1 - \bar{x}_2) - Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

donde  $Z_{\alpha/2}$  es el valor de la distribución normal estándar a la derecha del cual se tiene un área de  $\alpha/2$ . Si se

simplifica  $(\bar{x}_1 - \bar{x}_2) \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

### Ejemplo 3.18 Intervalos de confianza para la diferencia de medias, poblaciones aproximadamente normales cuando se conocen $\sigma_1$ y $\sigma_2$

Se comparan dos tipos de rosca de tornillo para ver su resistencia a la tensión. Para esto se prueban 12 piezas de cada tipo de cuerda bajo condiciones similares, obteniéndose los siguientes resultados en kilogramos (véase la tabla 3.5).

Tabla 3.5

Tipo de rosca	1	2	3	4	5	6	7	8	9	10	11	12
I	68	70	72	69	71	72	70	69	75	69	70	71
II	75	73	73	68	68	67	69	75	74	68	73	74

Si  $\mu_1$  y  $\mu_2$  son las resistencias promedio a la tensión y  $\sigma_1^2 = 5$  y  $\sigma_2^2 = 10$ , las varianzas de las resistencias de los tornillos tipo I y tipo II, respectivamente:

- Calcule un intervalo de 90% de confianza para  $\mu_1 - \mu_2$ , suponga normalidad en la tensión de los tornillos y que las muestras son independientes.
- ¿A 90% de confianza se puede decir que el tipo de rosca II tiene una mayor resistencia a la tensión que el tipo I?

#### Solución

Al realizar los cálculos se obtiene  $\bar{x}_1 = 70.50$  y  $\bar{x}_2 = 71.42$ , las muestras son de tamaño  $n_1 = n_2 = 12$ . Falta encontrar el valor para  $Z_{\alpha/2}$  con una confianza de 90%. De las tablas porcentuales para la distribución normal  $Z_{\alpha/2} = 1.645$ . Se conocen las varianzas poblacionales, esto indica que se puede utilizar la fórmula del teorema 3.7,

$$(70.50 - 71.42) - 1.645 \sqrt{\frac{5}{12} + \frac{10}{12}} < \mu_1 - \mu_2 < (70.50 - 71.42) + 1.645 \sqrt{\frac{5}{12} + \frac{10}{12}}$$

$$-2.76 < \mu_1 - \mu_2 < 0.92$$

- Es decir, la diferencia de la resistencia promedio a la tensión en la fabricación de los tornillos tipo I y II se encuentra en el intervalo  $(-2.76, 0.92)$  con una probabilidad de 0.90.
- El intervalo contiene tanto valores negativos como positivos; luego, con 90% de confianza no se puede asegurar que el tipo de rosca II tenga una mayor resistencia a la tensión que el tipo I, ya que con el resultado obtenido se concluye que ambas resistencias medias de los tornillos son iguales con 90% de confianza.

## Intervalos de confianza para la diferencia de medias de poblaciones normales cuando se desconocen $\sigma_1$ y $\sigma_2$ , pero se sabe que $\sigma_1^2 = \sigma_2^2$

El segundo caso que debemos revisar en la comparación de medias, de forma similar a una población, es cuando se desconocen las varianzas poblacionales. Pero a diferencia de una población en la diferencia de medias pueden existir dos situaciones.

**Teorema 3.8**

Si  $\bar{x}_1$  y  $\bar{x}_2$  son las medias aritméticas de las realizaciones de dos muestras aleatorias independientes de tamaños  $n_1$  y  $n_2$  respectivamente, de poblaciones aproximadamente normales de las cuales se desconoce  $\sigma_1^2$  y  $\sigma_2^2$ , pero se sabe que  $\sigma_1^2 = \sigma_2^2$ , entonces un intervalo de confianza de  $(1 - \alpha)100\%$  para  $\mu_1 - \mu_2$  está dado por:

$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2}(s_p) \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2}(s_p) \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

donde  $t_{\alpha/2}$  es el valor de la distribución t-Student con  $\nu = n_1 + n_2 - 2$  grados de libertad a la derecha del cual se

tiene un área de  $\alpha/2$ ;  $s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$  es la estimación común de la desviación estándar poblacional

y  $s_1^2, s_2^2$  son las varianzas insesgadas muestrales, 1 y 2, respectivamente. Al simplificar  $(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}(s_p) \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ .

Vemos un ejemplo en el que se muestre el uso del teorema 3.8.

**Ejemplo 3.19 Varianzas insesgadas**

Las pruebas de tracción en 10 puntos de soldadura en un dispositivo semiconductor produjeron los siguientes resultados en libras requeridas para romper la soldadura:

15.8, 12.7, 13.2, 16.9, 10.6, 18.8, 11.1, 14.3, 17.0, 12.5

Otro conjunto de ocho puntos fue probado después de recibir el dispositivo para determinar si la resistencia a la tracción se incrementa con el recubrimiento y se obtienen los siguientes resultados:

24.9, 23.6, 19.8, 22.1, 20.4, 21.6, 21.8, 22.5

Si se supone normalidad en las pruebas de tracción:

- Calcule un intervalo de confianza de 90% para  $\mu_1 - \mu_2$ , si considera  $\sigma_1^2 = \sigma_2^2$ , pero desconocidas.
- A 90% de confianza, ¿será posible decir que el recubrimiento utilizado aumenta la resistencia a la tracción en el dispositivo semiconductor empleado?, ¿en al menos seis?, y ¿en al menos 10?

**Solución**

Primero, se calculan las medias y varianzas insesgadas de los datos y del conjunto 1 se obtiene:  $\bar{x}_1 = 14.29$  y  $s_1^2 = 7.50$ ,  $n_1 = 10$  y del conjunto 2:  $\bar{x}_2 = 22.09$  y  $s_2^2 = 2.68$ ,  $n_2 = 8$ . Luego:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(10 - 1)7.50 + (8 - 1)2.68}{10 + 8 - 2}} = 2.32$$

Ahora, falta encontrar en las tablas porcentuales de la distribución t-Student el valor de  $t_{\alpha/2}$  con una confianza de 90%, es decir,  $\alpha = 0.10$  esto es  $\alpha/2 = 0.05$  y  $\nu = n_1 + n_2 = 16$  grados de libertad. Al buscar en las tablas de la t-Student con 16 grados de libertad,  $t_{0.05}(16) = 1.746$ .

- Al aplicar la fórmula del teorema 3.8, tenemos:

$$14.29 - 22.09 - (1.746)2.32 \sqrt{\frac{1}{10} + \frac{1}{8}} < \mu_1 - \mu_2 < 14.29 - 22.09 + (1.746)2.32 \sqrt{\frac{1}{10} + \frac{1}{8}} - 9.72 < \mu_1 - \mu_2 < -5.88$$

- Como se puede notar, el intervalo de confianza de la diferencia  $\mu_1 - \mu_2$  es siempre negativo. Por tanto, se puede decir con 90% de confianza que el recubrimiento utilizado aumenta la resistencia a la tracción en el dispositivo semiconductor utilizado.

Para los otros dos casos tenemos:

$$\mu_2 > \mu_1 + 6 \text{ o } \mu_1 - \mu_2 < -6 \text{ y } -6 \in (-9.72, -5.88).$$

Entonces, con 90% de confianza es válido suponer  $\mu_1 > \mu_2 + 6$ .

Por último, para  $\mu_2 > \mu_1 + 11$  o  $\mu_1 - \mu_2 < -11$ , tenemos  $-11 \notin (-9.72, -5.88)$ , entonces con 90% de confianza no es válido suponer  $\mu_2 > \mu_1 + 11$ .

## Intervalos de confianza para la diferencia de medias de poblaciones normales cuando se desconocen $\sigma_1$ y $\sigma_2$ , pero se sabe $\sigma_1^2 \neq \sigma_2^2$

El otro caso para comparar las medias de dos poblaciones es cuando las varianzas son desconocidas, pero diferentes.

### Teorema 3.9

Si  $\bar{x}_1$  y  $\bar{x}_2$  son las medias aritméticas de las realizaciones de dos muestras aleatorias independientes de tamaños  $n_1$  y  $n_2$ , de poblaciones aproximadamente normales de las que se desconoce  $\sigma_1^2$  y  $\sigma_2^2$ , pero se sabe  $\sigma_1^2 \neq \sigma_2^2$ , entonces un intervalo de confianza de  $(1 - \alpha)100\%$  para  $\mu_1 - \mu_2$  está dado por:

$$(\bar{x}_1 - \bar{x}_2) - t_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

donde:  $t_{\alpha/2}$  es el valor de la distribución t-Student con:

$$\nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\left(\frac{s_1^2}{n_1}\right)^2 \left(\frac{1}{n_1 - 1}\right) + \left(\frac{s_2^2}{n_2}\right)^2 \left(\frac{1}{n_2 - 1}\right)} \text{ grados de libertad,}$$

y a la derecha del cual se tiene un área de  $\alpha/2$ ;  $s_1^2$ ,  $s_2^2$  son las variancias insesgadas muestrales 1 y 2, al simplificar

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Se puede apreciar que los grados de libertad están en función de  $s_1^2$  y  $s_2^2$ ; por tanto, en realidad, representan una estimación de los grados de libertad. Por consiguiente, los grados de libertad calculados con esta fórmula en general darán una cantidad no entera, y debemos redondear al entero más próximo (¡no el siguiente!). Es decir,  $\nu = 14.3 \approx 14$  o en caso de que  $\nu = 14.7 \approx 15$ , si  $\nu = 14.5 \approx 15$ .

### Ejemplo 3.20 Intervalos de confianza

Resuelva el ejemplo anterior si considera que  $\sigma_1^2 \neq \sigma_2^2$  y ambas desconocidas. Al suponer normalidad, calcule un intervalo de confianza con 90% para  $\mu_1 - \mu_2$ . Determine qué tipo de semiconductor, sin recubrimiento 1 o con recubrimiento 2, tiene mayor resistencia a la tracción.

#### Solución

Las medias y variancias muestrales se calcularon antes y resultaron:  $\bar{x}_1 = 14.29$  y  $s_1^2 = 7.50$ ,  $n_1 = 10$ , y  $\bar{x}_2 = 22.09$  y  $s_2^2 = 2.68$ ,  $n_2 = 8$ . Con estos valores se calculan los grados de libertad:

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2 \left(\frac{1}{n_1-1}\right) + \left(\frac{s_2^2}{n_2}\right)^2 \left(\frac{1}{n_2-1}\right)} = \frac{\left(\frac{7.50}{10} + \frac{2.68}{8}\right)^2}{\left(\frac{7.50}{10}\right)^2 \left(\frac{1}{10-1}\right) + \left(\frac{2.68}{8}\right)^2 \left(\frac{1}{8-1}\right)} = 14.99 \approx 15$$

Falta encontrar de las tablas porcentuales de la distribución t-Student el valor de  $t_{\alpha/2}$  con una confianza de 90% ( $\alpha = 0.10$ , entonces  $\alpha/2 = 0.05$ ) y  $\nu = 15$  grados de libertad. Al buscar en las tablas de la distribución t-Student se obtiene  $t_{0.05} = 1.753$ .

a) De la fórmula del teorema 3.9 se tendrá:

$$14.29 - 22.09 - 1.753\sqrt{\frac{7.50}{10} + \frac{2.68}{8}} < \mu_1 - \mu_2 < 14.29 - 22.09 + 1.753\sqrt{\frac{7.50}{10} + \frac{2.68}{8}}$$

$$-9.63 < \mu_1 - \mu_2 < -5.97$$

b) Como se puede observar del intervalo de confianza calculado, la diferencia  $\mu_1 - \mu_2$  es siempre negativa. Por tanto, se tiene que con 90% de confianza, la resistencia a la tracción con recubrimiento es mayor que éste.

De forma similar se concluye en los otros dos casos:

c)  $\mu_2 > \mu_1 + 6$  y  $-6 \in (-9.63, -5.97)$ , con 90% de confianza es válido suponer  $\mu_2 > \mu_1 + 6$ .

d)  $\mu_2 > \mu_1 + 11$ , pero  $-11 \notin (-9.63, -5.97)$ , luego con 90% de confianza no es válido suponer  $\mu_2 > \mu_1 + 11$ , se tendría que aumentar el nivel de confianza.

## Intervalos de confianza para la diferencia de medias de poblaciones aproximadamente normales, se desconocen $\sigma_1$ y $\sigma_2$ muestras grandes

En los dos teoremas 3.8 y 3.9 la distribución que se tiene es la t-Student. Como se mencionó el uso de las tablas de esta distribución está limitado al tamaño de la muestra, de manera que en caso de no conocer los valores de  $t$  para tamaños de muestra grandes se buscan alternativas para los casos cuando los grados de libertad lo sean. Una de éstas se refiere a la aproximación de la distribución t-Student con la normal estándar. Así, las fórmulas de los teoremas 3.8 y 3.9 se pueden utilizar, al cambiar la distribución t-Student por la Z. Pero en general, para tamaños de muestras grandes se utiliza el siguiente resultado.

### Teorema 3.10

Si  $\bar{x}_1$  y  $\bar{x}_2$  son las medias aritméticas de las realizaciones de dos muestras aleatorias independientes de tamaños  $n_1 < 30$  y  $n_2 > 30$ , respectivamente, de poblaciones con distribuciones aproximadamente normales de las cuales se desconocen  $\sigma_1^2$  y  $\sigma_2^2$ , entonces un intervalo de confianza de  $(1 - \alpha)100\%$  para  $\mu_1 - \mu_2$  está dado por:

$$(\bar{x}_1 - \bar{x}_2) - Z_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + Z_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

donde  $Z_{\alpha/2}$  es el valor de la distribución normal estándar a la derecha del cual se tiene un área de  $\alpha/2$ ,  $s_1^2$  y  $s_2^2$  son las variancias insesgadas muestrales 1 y 2.



### Ejemplo 3.21 Intervalos de confianza para la diferencia de medias de poblaciones aproximadamente normales

Para comparar dos tipos de rosca de tornillos en su resistencia a la tensión, se prueban 40 piezas de cada tipo de cuerda bajo condiciones similares; se obtienen los siguientes resultados en kilogramos.

Del tipo I:  $\bar{x}_1 = 72.5$  y  $s_1 = 2.45$ ,  $n_1 = 40$  y del tipo II:  $\bar{x}_2 = 69.8$  y  $s_2 = 1.75$ ,  $n_2 = 40$ .

Si  $\mu_1$  y  $\mu_2$  son las resistencias promedio a la tensión de los tornillos tipo I y tipo II, respectivamente.

- Calcule un intervalo de 95% de confianza para  $\mu_1 - \mu_2$ , suponga normalidad en la tensión de los tornillos y que las muestras son independientes.
- ¿Con 95% de confianza se puede decir que uno de los dos tipos de rosca tiene una mayor resistencia a la tensión que el otro?

#### Solución

Debido a que los valores muestrales para la media y las desviaciones estándar son conocidos y aunado a esto las muestras son grandes ( $n_1 = n_2 = 40 > 30$ ), falta encontrar el valor para  $Z_{\alpha/2}$  con una confianza de 95%. Por tablas porcentuales  $Z_{0.025} = 1.960$ .

a) Luego

$$(72.5 - 69.8) - 1.96\sqrt{\frac{2.45^2}{40} + \frac{1.75^2}{40}} < \mu_1 - \mu_2 < (72.5 - 69.8) + 1.96\sqrt{\frac{2.45^2}{40} + \frac{1.75^2}{40}}$$

$$1.78 < \mu_1 - \mu_2 < 3.62$$

- b) Puesto que el intervalo de confianza para la diferencia de las medias poblacionales siempre es positivo, con una confianza de 95% se puede decir que la resistencia a la tensión de los tornillos tipo I es mayor a la del tipo II:

$$\mu_1 - \mu_2 \in (1.78, 3.62) \text{ indica que } \mu_1 - \mu_2 > 0, \text{ o sea } \mu_1 - \mu_2$$

### Ejemplo 3.22 Intervalos de confianza para la diferencia de medias de poblaciones aproximadamente normales

Resuelva el problema anterior, con los teoremas 3.8 y 3.9, y verifique que, en efecto, la aproximación encontrada en el ejemplo anterior es muy buena.

*Sugerencia.* El valor de la t-Student para los grados de libertad resultantes se puede calcular con la ayuda de Excel o interpolando el valor.

#### Solución

- a) Con los resultados del teorema 3.8, los grados de libertad son:  $n_1 + n_2 - 2 = 40 + 40 - 2 = 78$ , falta encontrar el valor para  $t_{\alpha/2}$  con 78 grados de libertad. El coeficiente de confianza es de 95%, por tanto  $\alpha/2 = 0.025$ , entonces  $t_{0.025}(78) = 1.9908 \approx 1.991$ . Ahora se calculará el valor de  $s_p$ :

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{39 \times 2.45^2 + 39 \times 1.75^2}{40 + 40 - 2}} = 2.129$$

Por último:

$$(72.5 - 69.8) - 1.991(2.129)\sqrt{\frac{1}{40} + \frac{1}{40}} < \mu_1 - \mu_2 < (72.5 - 69.8) + 1.991(2.129)\sqrt{\frac{1}{40} + \frac{1}{40}}$$

$$1.752 < \mu_1 - \mu_2 < 3.648$$



El intervalo de confianza con la aproximación fue  $\mu_1 - \mu_2 \in (1.78, 3.62)$ , el cual difiere un poco del caso cuando se desconocen  $\sigma_1^2$  y  $\sigma_2^2$ , pero se sabe que son iguales.

b) Con el resultado del teorema 3.9 y los grados de libertad, tenemos:

$$v = \frac{\left[ \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right]^2}{\left( \frac{s_1^2}{n_1} \right)^2 \left( \frac{1}{n_1 - 1} \right) + \left( \frac{s_2^2}{n_2} \right)^2 \left( \frac{1}{n_2 - 1} \right)} = \frac{\left[ \frac{2.45^2}{40} + \frac{1.75^2}{40} \right]^2}{\left( \frac{2.45^2}{40} \right)^2 \left( \frac{1}{40 - 1} \right) + \left( \frac{1.75^2}{40} \right)^2 \left( \frac{1}{40 - 1} \right)} = 70.576 \approx 71$$

Solo falta encontrar el valor para  $t_{\alpha/2}$  con 71 grados de libertad. El coeficiente de confianza es de 95%; por tanto,  $\alpha/2 = 0.025$ , entonces  $t_{0.025}(71) = 1.9936$ . Ahora:

$$(72.5 - 69.8) - 1.9936 \sqrt{\frac{2.45^2}{40} + \frac{1.75^2}{40}} < \mu_1 - \mu_2 < (72.5 - 69.8) + 1.9936 \sqrt{\frac{2.45^2}{40} + \frac{1.75^2}{40}}$$

$$1.751 < \mu_1 - \mu_2 < 3.649$$

El intervalo de la aproximación fue  $\mu_1 - \mu_2 \in (1.78, 3.62)$  el cual difiere ligeramente del caso en que se desconozcan  $\sigma_1^2$  y  $\sigma_2^2$  pero se sabe que son diferentes.

## Intervalos de confianza para la diferencia de medias de observaciones pareadas con diferencias normales

En los cuatro casos revisados sobre intervalos de confianza para diferencia de medias siempre se consideraron muestras independientes, pero en la práctica existe una infinidad de situaciones en las que son dependientes. En el caso de diferencia de medias se puede introducir la dependencia cuando se tienen observaciones pareadas.

Es necesario aclarar qué entendemos por observaciones pareadas, para lo cual se verá la siguiente terminología. En una situación experimental a los entes de estudio se les llama **unidades experimentales**, las cuales pueden ser personas, animales, áreas de cultivo, etcétera.

A las variaciones que existen en las diferentes observaciones de una misma unidad experimental se les llama **variaciones dentro de una misma unidad**. Cuando se trata de las variaciones de las observaciones de diferentes unidades experimentales se les llama **variaciones entre unidades**.

Con base en lo anterior, **un primer caso de observaciones pareadas** se tiene cuando a cada unidad experimental o de estudio se le toman dos observaciones. Veamos los siguientes ejemplos.

### Ejemplos 3.23 Observaciones pareadas

1. Suponga que se quiere analizar si un medicamento es o no apropiado para mejorar cierta enfermedad. Para este efecto se toman mediciones en cada paciente antes y después de haber aplicado el medicamento. En esta situación, sí se trata de observaciones pareadas ya que a cada unidad experimental se le tomaron dos mediciones para después realizar una comparación de resultados.
2. Suponga que para medir la efectividad de una dieta se eligen 15 personas a las que se pesan antes y después de haber utilizado la dieta y se comparan los resultados. En esta situación, sí se trata de observaciones pareadas porque las parejas de observaciones se tomaron de la misma unidad experimental.
3. Suponga que en un proceso químico se comparan dos marcas de catalizadores para verificar su efecto en el resultado de la reacción del proceso. Se prepara una muestra de 12 procesos con el uso del catalizador marca  $L$  y otra muestra de 12 procesos con el uso del catalizador de la marca  $M$ .

En esta situación, no se trata de observaciones pareadas, ya que son diferentes las unidades experimentales a las que se les aplica un tipo de catalizador. Este ejemplo aclara que el tema de observaciones pareadas no se refiere solo a parejas de observaciones, sino que deben ser observaciones obtenidas de la misma unidad experimental.

En las observaciones pareadas se mencionó la restricción de que éstas sean obtenidas de la misma unidad experimental, esto se debe a que en la teoría se requiere que las variaciones dentro de una misma unidad experimental sean pequeñas. Es decir, que las unidades experimentales seleccionadas sean relativamente homogéneas (dentro de las unidades). Por consiguiente, para que se trate de observaciones pareadas con diferentes unidades experimentales es necesario que en el apareamiento cada par tenga características muy similares.

### Ejemplos 3.24 Observaciones pareadas

1. Suponga que se quiere analizar si el proceso de enseñanza aprendizaje de las matemáticas mejora al utilizar paquetería. De una población de estudiantes que están en el mismo curso y tienen la misma edad e igual promedio académico se eligen dos muestras de tamaño 20. A cada una se le enseña el mismo tema, pero una con paquetería y la otra sin ésta, y se les aplica un examen para medir los rendimientos.

En esta situación, sí es posible considerar que se trata de observaciones pareadas, porque no importa que cada par se refiera a diferentes unidades experimentales; las muestras se obtuvieron de estudiantes que tienen las mismas características y el resultado es el promedio.

2. Suponga que se quiere analizar si el proceso de enseñanza aprendizaje de las matemáticas mejora al utilizar paquetería. De una población de estudiantes se eligen dos muestras de tamaño 20. A cada una se le enseña el mismo tema, pero una sin paquetería y la otra con el uso de ésta y se les aplica un examen para medir los rendimientos.

En esta situación, no se trata de observaciones pareadas, porque no se especifica que los estudiantes tengan las mismas características.

Con esta breve introducción a las observaciones pareadas se espera tener una visión más clara de la aplicación y el alcance de los experimentos cuando se lleva a cabo un apareamiento de observaciones.

Si se representa por  $X$  y  $Y$  a las variables aleatorias correspondientes de las parejas de observaciones,  $(x_1, y_1)$ ,  $(x_2, y_2)$ , ...,  $(x_n, y_n)$ , con  $(X_1, Y_1)$ ,  $(X_2, Y_2)$ , ...,  $(X_n, Y_n)$  con medias y varianzas respectivas;  $\mu_X$ ,  $\mu_Y$  y  $\sigma_X^2$ ,  $\sigma_Y^2$ . Por otro lado, sea  $D$  la variable aleatoria de la diferencia entre las variables  $X$  y  $Y$ , de manera que:

$$D_i = X_i - Y_i, i = 1, 2, \dots, n,$$

representa la variable aleatoria resultante de la diferencia entre las variables  $X_i$  y  $Y_i$ . Suponga que las  $D_i$  tienen distribución normal con media  $\mu_D$  y varianza  $\sigma_D^2$  y que son independientes (es decir, las variables aleatorias entre parejas diferentes son independientes, pero las variables dentro del mismo par son dependientes). De esta forma, se tiene:

$$\mu_D = E(X - Y) = \mu_X - \mu_Y$$

Mientras que la varianza:

$$\sigma_D^2 = V(X - Y) = \sigma_X^2 + \sigma_Y^2 - \text{cov}(X, Y)$$

es estimada por una realización de las parejas  $(X_1, Y_1)$ ,  $(X_2, Y_2)$ , ...,  $(X_n, Y_n)$  dada por  $(x_1, y_1)$ ,  $(x_2, y_2)$ , ...,  $(x_n, y_n)$ , de donde se calcula  $d_i = x_i - y_i$ . Luego:

$$\bar{x}_d = \frac{1}{n} \sum_{i=1}^n d_i = \frac{1}{n} \sum_{i=1}^n (x_i - y_i) \text{ estimará a } \mu_D \text{ y}$$

$$s_d^2 = \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{x}_d)^2 = \frac{1}{n-1} \sum_{i=1}^n [(x_i - y_i) - \bar{x}_d]^2 \text{ estimará a } \sigma_D^2$$

De manera que surge el teorema 3.11.

### Teorema 3.11

Si  $\bar{x}_d$  y  $s_d$  son la media y la desviación estándar muestrales de la diferencia de  $n$  pares de realizaciones de muestras aleatorias pareadas, tomadas de mediciones normalmente distribuidas de las cuales se desconoce  $\sigma_X^2$  y  $\sigma_Y^2$ , entonces un intervalo de confianza de  $(1 - \alpha)100\%$  para  $\mu_d = \mu_X - \mu_Y$  está dado por:

$$\bar{x}_d - t_{\alpha/2} \left( \frac{s_d}{\sqrt{n}} \right) < \mu_d < \bar{x}_d + t_{\alpha/2} \left( \frac{s_d}{\sqrt{n}} \right)$$

donde  $t_{\alpha/2}$  es el valor de la distribución t-Student con  $\nu = n - 1$  grados de libertad a la derecha del cual se tiene un área de  $\alpha/2$ . Se simplifica  $\bar{x}_d \pm t_{\alpha/2} \frac{s_d}{\sqrt{n}}$ .

A continuación, analizamos un ejemplo que ilustra el resultado del teorema 3.11.

### Ejemplo 3.25 Muestras pareadas

Para probar si un tratamiento reductor de peso es eficiente se toma una muestra de 10 personas, las cuales serán sujetas al tratamiento. Se anotan sus pesos en kilogramos antes y después del tratamiento, cuyos resultados se aprecian en la tabla 3.6.

Tabla 3.6

Persona	1	2	3	4	5	6	7	8	9	10
Antes	81	75	74	69	71	72	70	84	75	79
Después	75	72	71	67	68	67	68	75	72	73

Se considera que el tratamiento de peso es eficaz si reduce el peso de la persona en 5 kg.

- Calcule un intervalo de 95% de confianza para  $\mu_1 - \mu_2$ , suponga normalidad en las diferencias de los pesos.
- ¿Con 95% de confianza será posible decir que el tratamiento es eficaz?

#### Solución

En este caso se trata de muestras pareadas porque las parejas de observaciones se toman de la misma persona (unidad experimental), de manera que al calcular las diferencias se tienen los resultados de la tabla 3.7.

Tabla 3.7

Persona	1	2	3	4	5	6	7	8	9	10
Antes	81	75	74	69	71	72	70	84	75	79
Después	75	72	71	67	68	67	68	75	72	73
Diferencia	6	3	3	2	3	5	2	9	3	6

Al realizar los cálculos con las diferencias,  $\bar{x}_d = 4.2$ ,  $s_d = 2.251$ ,  $n = 10$  y  $1 - \alpha = 0.95$ . Luego,  $\alpha/2 = 0.025$ . Así, de las tablas porcentuales para la distribución t-Student,  $t_{0.025}(9) = 2.262$ ,

a) Por la fórmula del teorema 3.11, se tiene:

$$\bar{x}_d - t_{\frac{\alpha}{2}} \left( \frac{s_d}{\sqrt{n}} \right) < \mu_d < \bar{x}_d + t_{\frac{\alpha}{2}} \left( \frac{s_d}{\sqrt{n}} \right)$$

$$4.2 - 2.262 \left( \frac{2.251}{\sqrt{10}} \right) < \mu_d < 4.2 + 2.262 \left( \frac{2.251}{\sqrt{10}} \right)$$

$$2.590 < \mu_d < 5.810$$

Es decir, la diferencia de los pesos antes y después del tratamiento se encuentra en el intervalo (2.590, 5.810), con una confianza de 95 por ciento.

b) Debido a que el intervalo contiene el valor de 5 kg se dice que con 95% de confianza el tratamiento para rebajar de peso es eficaz.

## Ejemplos variados para la estimación de diferencia de medias

De forma similar como se trabajó en los intervalos de confianza para la media se puede hacer en la diferencia de medias, para calcular no solo los intervalos de confianza, sino los errores de estimación ( $\varepsilon = |(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)|$ ) y tamaños de muestra, los cuales se pueden obtener de forma similar en el teorema 2.6 en la unidad 2. Para los tamaños de muestra en diferencia de medias, se reducen a los casos en que éstos son iguales.

### Ejemplo 3.26 Estimación de diferencia de medias

Se comparan dos tipos de rosca de tornillo para ver su resistencia a la tensión, para lo cual se prueban  $n$  piezas de cada tipo de cuerda bajo condiciones similares. Si  $\mu_1$  y  $\mu_2$  son resistencias promedio a la tensión de los tornillos tipo I y tipo II, respectivamente. Sean  $\sigma_1^2 = 5$  y  $\sigma_2^2 = 10$  varianzas a la tensión de los tornillos tipo I y tipo II, respectivamente, ¿cuál debe ser el tamaño mínimo de las muestras para que la diferencia de las medias muestrales sea de un kilogramo con respecto a las medias poblacionales con una confianza de 95%? Suponga normalidad en la tensión de los tornillos y que las muestras son independientes.

#### Solución

De los datos del enunciado, se obtiene que  $\sigma_1^2 = 5$  y  $\sigma_2^2 = 10$  y  $\varepsilon = |(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)| = 1$  kg. Luego, de la fórmula del tamaño del error para el caso I (se conocen las varianzas poblacionales).

$$|(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)| < Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{n}} \text{ así, } \varepsilon = Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{n}}$$

Falta calcular el valor de  $Z_{\alpha/2}$ , con  $1 - \alpha = 0.95$ . De las *tablas porcentuales* para la distribución normal estándar se tiene que  $Z_{\alpha/2} = 1.960$ . Por tanto, si se despeja el tamaño de la muestra tenemos:

$$n = \left\lceil \left[ \left( \frac{Z_{\alpha/2}}{\varepsilon} \right)^2 (\sigma_1^2 + \sigma_2^2) \right] + 1 \right\rceil = \left\lceil \left[ \left( \frac{1.96}{1} \right)^2 (5 + 10) \right] + 1 \right\rceil = \lceil [57.624] + 1 \rceil = 58$$

Es decir, el tamaño mínimo de las muestras para que el error de estimación valga 1 kg es  $n \geq 58$ .

**Ejemplo 3.27 Estimación de diferencia de medias**

En un proceso químico se comparan dos catalizadores de las marcas  $L$  y  $M$  para verificar su efecto en el resultado de la reacción del proceso. Según estudios, los catalizadores  $L$  y  $M$  tienen idénticas características. Se preparó una muestra de 10 procesos utilizando el catalizador marca  $L$  y otra muestra de 10 procesos utilizando el catalizador de la marca  $M$ , en la tabla 3.8 se muestran los datos con los rendimientos.

**Tabla 3.8**

$L$	0.99	0.90	0.32	0.70	0.43	0.67	0.65	0.61	0.44	0.92
$M$	0.95	0.40	0.60	0.62	0.44	0.62	0.42	0.72	0.26	0.86

- a) Calcule un intervalo de confianza para la diferencia de los rendimientos a 99%.  
 b) ¿Con 99% de confianza se puede decir que los dos tipos de catalizadores tienen el mismo efecto en la reacción del proceso?

Suponga que las diferencias están distribuidas normalmente.

**Solución**

Primero se determina que se trata de observaciones pareadas porque los tipos de catalizadores  $L$  y  $M$  tienen características idénticas. Se calculan las diferencias de los datos muestrales.

**Tabla 3.9**

$L$	0.99	0.90	0.32	0.70	0.43	0.67	0.65	0.61	0.44	0.92
$M$	0.95	0.40	0.60	0.62	0.44	0.62	0.42	0.72	0.26	0.86
$L - M$	0.04	0.50	-0.28	0.08	-0.01	0.05	0.23	-0.11	0.18	0.06

Con las diferencias anteriores se calcula su valor medio y desviación estándar:

$$\bar{x}_d = 0.074 \text{ y } s_d = 0.207$$

El tamaño de la muestra es 10, por consiguiente, los grados de libertad  $\nu = 10 - 1 = 9$ . De las tablas porcentuales correspondientes a la distribución t-Student con 99% de confianza,  $\alpha = 0.01$  y  $\alpha/2 = 0.005$ , se tiene que  $t_{0.005} = 3.25$ .

Por último, el intervalo de confianza resulta:

$$0.074 - 3.250 \left( \frac{0.207}{\sqrt{10}} \right) < \mu_d < 0.074 + 3.250 \left( \frac{0.207}{\sqrt{10}} \right)$$

$$-0.139 < \mu_d < 0.287$$

Debido a que el intervalo de confianza contiene al cero, se puede decir que con 99% de confianza los dos tipos de catalizadores tienen el mismo efecto de la reacción del proceso.

**Ejemplo 3.28 Coeficiente de confianza**

En el ejemplo anterior, ¿qué se puede suponer respecto a los catalizadores con 70% de confianza?

**Solución**

De los cálculos anteriores, se tiene  $\nu = 10 - 1 = 9$ ,  $\bar{x}_d = 0.074$  y  $s_d = 0.207$ .

Ahora, de las tablas porcentuales de la distribución t-Student con 70% de confianza,  $\alpha = 0.30$  y  $\alpha/2 = 0.15$ , resulta  $t_{0.15} = 1.0997$ . Por último, el intervalo de confianza está dado por:

$$0.074 - 1.0997 \left( \frac{0.207}{\sqrt{10}} \right) < \mu_d < 0.074 + 1.0997 \left( \frac{0.207}{\sqrt{10}} \right)$$

$$0.002 < \mu_d < 0.146$$

Con base en  $\mu_d = \mu_L - \mu_M$  y que el intervalo de confianza no contiene valores negativos, se puede concluir con 70% de confianza que la reacción del catalizador  $L$  es mayor a la del  $M$ .

Ahora bien, ¿cómo determinar el coeficiente de confianza cuando se tienen los límites del intervalo?

En estos casos  $\bar{x}_1 - \bar{x}_2$  siempre es el valor medio del intervalo de confianza para el parámetro  $\mu_1 - \mu_2$  y se puede formular un teorema similar al 3.4.

### Teorema 3.12

Si  $\bar{x}_1$  y  $\bar{x}_2$  son las medias aritméticas y  $s_1^2$  y  $s_2^2$  varianzas insesgadas de las realizaciones de dos muestras aleatorias independientes de tamaños  $n_1$  y  $n_2$ , de poblaciones con distribuciones aproximadamente normales y  $a < \mu_1 - \mu_2 < b$  un intervalo de  $(1 - \alpha)100\%$  de confianza para la diferencia de parámetros  $\mu_1 - \mu_2$ , entonces:

- $1 - \alpha = 1 - 2\Phi \left( \frac{\bar{x}_1 - \bar{x}_2 - b}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \right) = 2\Phi \left( \frac{\bar{x}_1 - \bar{x}_2 - a}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \right) - 1$ , se conocen  $\sigma_1^2$  y  $\sigma_2^2$ .

- $1 - \alpha = 1 - 2F_{t_\nu} \left( \frac{\bar{x}_1 - \bar{x}_2 - b}{s_p \sqrt{1/n_1 + 1/n_2}} \right) = 2F_{t_\nu} \left( \frac{\bar{x}_1 - \bar{x}_2 - a}{s_p \sqrt{1/n_1 + 1/n_2}} \right) - 1$ , se desconocen  $\sigma_1^2$  y  $\sigma_2^2$ ,

pero se sabe que  $\sigma_1^2 = \sigma_2^2$  con  $s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$  y  $\nu = n_1 + n_2 - 2$  grados de libertad.

- $1 - \alpha = 1 - 2F_{t_\nu} \left( \frac{\bar{x}_1 - \bar{x}_2 - b}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \right) = 2F_{t_\nu} \left( \frac{\bar{x}_1 - \bar{x}_2 - a}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \right) - 1$  se desconocen  $\sigma_1^2$  y  $\sigma_2^2$ ,

pero se sabe que  $\sigma_1^2 \neq \sigma_2^2$  con  $\nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\left(\frac{s_1^2}{n_1}\right)^2 \left(\frac{1}{n_1 - 1}\right) + \left(\frac{s_2^2}{n_2}\right)^2 \left(\frac{1}{n_2 - 1}\right)}$  grados de libertad.

- $1 - \alpha = 1 - 2F_{t_\nu} \left( \frac{\bar{x}_d - b}{s_d} \sqrt{n} \right) = 2F_{t_\nu} \left( \frac{\bar{x}_d - a}{s_d} \sqrt{n} \right) - 1$  muestras pareadas  $\nu = n - 1$ .

Donde  $\Phi$  representa la distribución acumulada de la normal estándar y  $F_{t_\nu}$  la distribución acumulada de la t-Student con  $\nu$  grados de libertad.

### Ejemplo 3.29 Coeficiente de confianza del intervalo para la diferencia de medias

Las pruebas de tracción en 10 puntos de soldadura en un dispositivo semiconductor produjeron los siguientes resultados en libras requeridas para romper la soldadura:

15.8, 12.7, 13.2, 16.9, 10.6, 18.8, 11.1, 14.3, 17.0, 12.5

Otro conjunto de ocho puntos fue probado después de recibir el dispositivo para determinar si la resistencia a la tracción se incrementa con el recubrimiento y se obtuvieron los siguientes resultados:

24.9, 23.6, 19.8, 22.1, 20.4, 21.6, 21.8, 22.5

Suponga normalidad en las pruebas de tracción y  $\sigma_1^2 = \sigma_2^2$ , pero desconocidas. ¿Cuál tendría que ser el coeficiente de confianza del intervalo para la diferencia de medias, si se quiere que la media de la resistencia a la tracción con recubrimiento exceda entre 5.8 y 9.8 lb la resistencia a la tracción sin recubrimiento del dispositivo semiconductor?

### Solución

Este problema trata del caso en que se desconocen las varianzas poblacionales, pero se sabe que son iguales.

Primero se calculan las medias y variancias muestrales, y se obtiene:

Del conjunto uno sin recubrimiento,  $\bar{x}_1 = 14.29$  y  $s_1^2 = 7.50$ ,  $n_1 = 10$ .

Del conjunto dos con recubrimiento,  $\bar{x}_2 = 22.09$  y  $s_2^2 = 2.68$ ,  $n_2 = 8$ .

Así, con los valores anteriores se calcula la estimación de la desviación estándar:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(10 - 1)7.50 + (8 - 1)2.68}{10 + 8 - 2}} = 2.32$$

Los grados de libertad son  $\nu = n_1 + n_2 - 2 = 16$ , luego  $a = 5.8$  y  $b = 9.8$ , utilizando  $b = 9.8$

$$1 - \alpha = 1 - 2F_{t_\nu} \left( \frac{\bar{x}_2 - \bar{x}_1 - b}{s_p \sqrt{1/n_2 + 1/n_1}} \right) = 1 - 2F_{t_\nu} \left( \frac{22.09 - 14.29 - 9.8}{2.32 \sqrt{1/8 + 1/10}} \right) = 1 - 2F_{t_\nu} (-1.8174)$$

El valor se busca en las tablas de la distribución t-Student, pero se tiene un problema debido a que en éstas se da el valor para un área derecha, luego:

$$\frac{\alpha}{2} = P(T_{\alpha/2}(16) \geq 1.8174)$$

De manera que el valor de 1.8174 se busca en el renglón correspondiente a 16 grados de libertad, es obvio que es muy improbable que justo este valor se encuentre en las tablas. Por tanto, se toman los dos valores más próximos.

Para 1.805 corresponde 0.045 y para 1.869 corresponde 0.040.

Se puede tomar el valor más cercano a 1.8174, entonces  $\alpha/2 = 0.045$ . Es decir, el coeficiente de confianza utilizado se considera,  $1 - \alpha = 1 - 2(0.045) = 0.91$  o 91%.

Otra manera más formal se lleva a cabo mediante una interpolación para el valor de 1.8174 y se obtiene  $\alpha/2 = 0.04403$ , en estas condiciones el nivel de confianza es:

$$1 - \alpha = 1 - 2(0.04403) = 0.91194 \text{ o } 91.19\%$$

La fórmula para la interpolación lineal está dada por:

$$y^* = y_0 + \frac{y_1 - y_0}{x_1 - x_0} (x^* - x_0)$$

Al aplicarla, el valor interpolado es:

$$y^* = 0.045 + \frac{0.040 - 0.045}{1.869 - 1.805} (1.8174 - 1.805) = 0.04403$$

## Ejemplo 3.30 Diferencia de medias poblacionales

De dos poblaciones normales con la misma varianza se tomaron muestras independientes y se obtuvo la siguiente información:

$$n_1 = 8; \sum_{i=1}^{n_1} x_i = 768; \sum_{i=1}^{n_1} x_i^2 = 74120 \text{ y } n_2 = 10, \sum_{i=1}^{n_2} y_i = 970, \sum_{i=1}^{n_2} y_i^2 = 94360$$

Obtenga un intervalo de confianza de 90% para la diferencia de las medias poblacionales.

**Solución**

Primero, se calculan las medias y variancias de los datos, lo cual da como resultado:

$$\bar{x} = \frac{1}{n_1} \sum_{i=1}^{n_1} x_i = \frac{1}{8}(768) = 96 \text{ y } \bar{y} = \frac{1}{n_2} \sum_{i=1}^{n_2} y_i = \frac{1}{10}(970) = 97$$

$$s_X^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} x_i^2 - \frac{n_1}{n_1 - 1} \bar{x}^2 = \frac{1}{7}(74120) - \frac{8}{7}(96)^2 = 56$$

$$s_Y^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} y_i^2 - \frac{n_2}{n_2 - 1} \bar{y}^2 = \frac{1}{9}(94360) - \frac{10}{9}(97)^2 = 30$$

Se trata del caso en el que se desconocen las variancias poblacionales, pero se sabe que son iguales (en el enunciado dice: con la misma varianza). Por consiguiente, se tiene que calcular el valor de:

$$s_p = \sqrt{\frac{(n_1 - 1)s_X^2 + (n_2 - 1)s_Y^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(8 - 1)56 + (10 - 1)30}{8 + 10 - 2}} = 6.4323$$

Aún falta encontrar con las tablas porcentuales de la distribución t-Student al valor de  $t_{\alpha/2}$  con una confianza de 90% ( $\alpha = 0.10$ , es decir,  $\alpha/2 = 0.05$ ) y  $\nu = 16$  grados de libertad. Al buscar en las tablas correspondientes se obtiene  $t_{0.05}(16) = 1.746$ . De la fórmula del teorema 3.8 se tiene:

$$96 - 97 - (1.746)(6.4323)\sqrt{\frac{1}{8} + \frac{1}{10}} < \mu_1 - \mu_2 < 96 - 97 + (1.746)(6.4323)\sqrt{\frac{1}{8} + \frac{1}{10}}$$

$$-6.327 < \mu_1 - \mu_2 < 4.327$$

Por último, se dice que la diferencia de medias poblacionales  $\mu_2 - \mu_1$  está entre  $-6.327$  y  $4.327$  con una confianza de 90%.

## Intervalos de confianza para la razón entre varianzas de poblaciones normales

Cuando se tienen dos poblaciones normales con parámetros  $(\mu_1, \sigma_1^2)$  y  $(\mu_2, \sigma_2^2)$ , respectivamente, se vio que podemos comparar sus rendimientos promedio con base en la diferencia de medias. Ahora se estudiará la forma de hacer inferencia con respecto a las varianzas, para ver cuál de éstas es más grande. Es decir, determinaremos la población más heterogénea, para esto recurriremos a un resultado de distribuciones muestrales que establece:

$$\frac{S_{n_1-1}^2 \sigma_2^2}{S_{n_2-1}^2 \sigma_1^2},$$



que tiene una distribución  $F$  con  $\nu_1 = n_1 - 1$  y  $\nu_2 = n_2 - 1$  grados de libertad en el numerador y denominador, de manera respectiva. La fórmula que aquí utilizaremos para los intervalos de confianza de la razón entre varianzas no cumple con una amplitud mínima, sin embargo, da una muy buena aproximación. La construcción de los intervalos de confianza para la razón entre varianzas con amplitud mínima resulta bastante compleja y depende de cada realización. Por estas razones no se usa.

### Teorema 3.13

Si  $s_1^2$  y  $s_2^2$  son las varianzas insesgadas de las realizaciones de dos muestras aleatorias independientes de tamaños  $n_1$  y  $n_2$  de poblaciones normales con parámetros  $(\mu_1, \sigma_1^2)$   $(\mu_2, \sigma_2^2)$ , respectivamente, entonces un intervalo de confianza de  $(1 - \alpha)$  100% para la razón de las varianzas  $\sigma_1^2/\sigma_2^2$  está dado por:

$$\left(\frac{s_1^2}{s_2^2}\right) \frac{1}{f_{\alpha/2}(\nu_1, \nu_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \left(\frac{s_1^2}{s_2^2}\right) f_{\alpha/2}(\nu_2, \nu_1)$$

Donde  $f_{\alpha/2}(\nu_1, \nu_2)$  es el valor de la distribución  $F$  (ver tablas correspondientes), con  $\nu_1 = n_1 - 1$  grados de libertad del numerador y  $\nu_2 = n_2 - 1$  grados de libertad del denominador, con un área a la derecha  $\alpha/2$ , de manera similar  $f_{\alpha/2}(\nu_2, \nu_1)$ .

Revisemos un ejemplo para ver la aplicación del teorema 3.13.

### Ejemplo 3.31 Intervalo de confianza para la razón entre varianzas

En el ejemplo 3.29 sobre las pruebas para la resistencia a la tracción de soldadura se hizo la suposición de que  $\sigma_1^2 = \sigma_2^2$ . Encuentre un intervalo de confianza para la razón de varianzas y determine si fue válida la suposición con una confianza de 90%.

#### Solución

Los datos para la prueba 1 fueron: 15.8, 12.7, 13.2, 16.9, 10.6, 18.8, 11.1, 14.3, 17.0, 12.5.

Para la prueba 2: 24.9, 23.6, 19.8, 22.1, 20.4, 21.6, 21.8, 22.5.

Al calcular las varianzas muestrales se obtuvo  $s_1^2 = 7.50$ ,  $n_1 = 10$  y  $s_2^2 = 2.68$ ,  $n_2 = 8$ .

Falta encontrar de las tablas porcentuales de la distribución  $F$  a los valores de  $f_{\alpha/2}(\nu_1, \nu_2)$  y  $f_{\alpha/2}(\nu_2, \nu_1)$  con una confianza de 90% ( $\alpha = 0.10$ , es decir,  $\alpha/2 = 0.05$ ) y  $\nu_1 = n_1 - 1 = 10 - 1 = 9$  y  $\nu_2 = n_2 - 1 = 8 - 1 = 7$  grados de libertad. Al buscar  $F$  en las tablas de la distribución obtenemos:

$$f_{\alpha/2}(\nu_1, \nu_2) = f_{0.05}(9, 7) = 3.677 \quad \text{y} \quad f_{\alpha/2}(\nu_2, \nu_1) = f_{0.05}(7, 9) = 3.293$$

Por tanto, el intervalo de confianza resulta:

$$\left(\frac{7.50}{2.68}\right) \frac{1}{3.677} < \frac{\sigma_1^2}{\sigma_2^2} < \left(\frac{7.50}{2.68}\right) 3.293$$

$$0.76 < \frac{\sigma_1^2}{\sigma_2^2} < 9.22$$

En esta sección se llevará a cabo una comparación entre dos varianzas, por esta razón es conveniente tener en cuenta los casos que pueden ocurrir. Suponga que se busca un intervalo con un coeficiente de confianza del  $1 - \alpha$  para  $\sigma_1^2/\sigma_2^2$ , y  $(a, b)$ , entonces se tiene lo siguiente:

- Si  $a$  y  $b$  ambos están entre 0 y 1, significa que al  $1 - \alpha$  de confianza se puede suponer que  $\sigma_1^2 < \sigma_2^2$ .
- Si  $a$  y  $b$  ambos son mayores a 1, significa que al  $1 - \alpha$  de confianza se puede suponer que  $\sigma_1^2 > \sigma_2^2$ .

Si  $a$  está entre 0 y 1 y  $b$  es mayor a 1, significa que al  $1 - \alpha$  de confianza se puede suponer que  $\sigma_1^2 = \sigma_2^2$ .

Por último, del intervalo de confianza para la razón entre variancias se puede observar que el valor 1 está contenido en el intervalo. Por tanto, con 90% de confianza se justifica la suposición de que  $\sigma_1^2 = \sigma_2^2$ , ya que  $\sigma_1^2/\sigma_2^2 = 1 \in (0.76, 9.22)$  y si multiplicamos por  $\sigma_2^2$  ambos miembros de la igualdad se obtiene  $\sigma_1^2 = \sigma_2^2$ .

### Ejemplo 3.32 Comparación entre dos variancias

Una empresa requiere de un cierto producto fabricado por dos distribuidores diferentes  $A$  y  $B$ . Para estimar la diferencia en la duración del producto entre estos dos distribuidores se lleva a cabo un experimento con doce artículos de cada distribuidor. Se obtuvo que el promedio y desviación estándar en  $A$  son 70.5 y 1.88, respectivamente, mientras que para la muestra de  $B$  se obtuvo el promedio 73.4 y desviación estándar 3.12. Determine un intervalo de confianza de 95% para  $\sigma_A^2/\sigma_B^2$  e indique al 95% de confianza qué distribuidor fabrica un producto más homogéneo.

#### Solución

Los datos son  $s_A = 1.88$ ,  $s_B = 3.12$  y  $n_A = n_B = 12$ . Falta encontrar de las tablas porcentuales de la distribución  $F$  a los valores de  $f_{\alpha/2}(v_A, v_B)$  y  $f_{\alpha/2}(v_B, v_A)$  con una confianza de 95% ( $\alpha = 0.05$ , es decir,  $\alpha/2 = 0.025$ ) y  $v_A = n_A - 1 = 12 - 1 = 11$  y  $v_B = n_B - 1 = 12 - 1 = 11$  grados de libertad del numerador y denominador. Al buscar en las tablas de la distribución  $F$  se obtiene:

$$f_{\alpha/2}(v_A, v_B) = f_{0.025}(11, 11) = 3.474$$

El intervalo de confianza resulta:

$$\left(\frac{1.88}{3.12}\right)^2 \frac{1}{3.474} < \frac{\sigma_A^2}{\sigma_B^2} < \left(\frac{1.88}{3.12}\right)^2 3.474 \Rightarrow 0.1045 < \frac{\sigma_A^2}{\sigma_B^2} < 1.2614$$

Por último, del intervalo de confianza para la razón entre variancias se puede observar que el valor 1 está contenido en el intervalo. Por tanto, con 95% de confianza no se puede decir qué distribuidor tiene su fabricación más homogénea.

### Ejemplo 3.33 Comparación entre dos variancias

Resuelva el problema anterior con 80% de confianza.

#### Solución

Con respecto al ejemplo anterior solo cambia el nivel de confianza con 80% ( $\alpha = 0.20$ , es decir,  $\alpha/2 = 0.10$ ). Al buscar en las tablas de la distribución  $F$  se obtiene:

$$f_{\alpha/2}(v_A, v_B) = f_{0.10}(11, 11) = 2.227$$

El intervalo de confianza resulta:

$$\left(\frac{1.88}{3.12}\right)^2 \frac{1}{2.227} < \frac{\sigma_A^2}{\sigma_B^2} < \left(\frac{1.88}{3.12}\right)^2 2.227 \Rightarrow 0.163 < \frac{\sigma_A^2}{\sigma_B^2} < 0.809$$

Por último, del intervalo de confianza para la razón entre variancias se puede notar que la razón está entre (0, 1), luego el denominador tiene que ser mayor que el numerador. Por tanto, se puede concluir que el distribuidor  $A$  tiene un producto más homogéneo que el fabricante  $B$  y esto se puede asegurar con una confianza de 80%.

De forma similar que con la media y la diferencia de medias, también podemos establecer un resultado para calcular el coeficiente de confianza para la razón entre variancias, el problema está en que se requiere de un paquete estadístico para calcular los valores.

**Teorema 3.14**

Si  $s_1^2$  y  $s_2^2$  son las variancias insesgadas de las realizaciones de dos muestras aleatorias independientes de tamaños  $n_1$  y  $n_2$  de poblaciones normales con parámetros  $(\mu_1, \sigma_1^2)$  y  $(\mu_2, \sigma_2^2)$ , y  $a < \sigma_1^2/\sigma_2^2 < b$  un intervalo de  $(1 - \alpha)100\%$  de confianza  $\sigma_1^2/\sigma_2^2$ , entonces:

$$1 - \alpha = 1 - 2f_{\nu_1, \nu_2}^p \left( \frac{s_1^2}{s_2^2} \frac{1}{a} \right) = 2f_{\nu_2, \nu_1}^p \left( \frac{s_2^2}{s_1^2} b \right) - 1$$

Donde  $f_{\nu_1, \nu_2}^p(q)$  representa la probabilidad de la distribución  $F$ , con  $\nu_1 = n_1 - 1$  g.l. en el numerador y  $\nu_2 = n_2 - 1$  g.l. en el denominador, a la derecha del valor  $q$ .

Para utilizar esta fórmula es muy probable que deba hacer uso de Excel, porque en las tablas solo se tienen los valores de áreas derechas 0.005, 0.010, 0.025, 0.05, 0.10 y 0.20.

**Ejercicios 3.6****Instrucciones**

- En cada ejercicio que construya un intervalo de confianza para la diferencia de medias debe interpretar la relación que existe entre éstas.
  - En cada ejercicio que construya un intervalo de confianza para la razón entre varianzas debe interpretar la relación que existe entre éstas.
  - I.C., intervalo de confianza;  $1 - \alpha$ , nivel de confianza del intervalo.
1. En un experimento se compararon las economías de combustible de dos tipos de vehículos. Se utilizaron 12 automóviles VG y 10 NS en pruebas de velocidad fija de 90 km/h. Si para los autos VG se obtuvo un promedio de 12.5 km/l con una desviación estándar de 2.0 km/l y para los autos NS fue de 14.2 km/l, con una desviación estándar de 1.8 km/l, suponga que la distancia recorrida por litro para cada modelo de vehículo se distribuye aproximadamente en forma normal. Con 95% de confianza.
    - a) Construya un intervalo para la diferencia del rendimiento promedio por litro, de los dos automóviles y suponga que  $\sigma_V^2 = 5$  y  $\sigma_N^2 = 4$ .
    - b) Determine el tamaño mínimo de la muestra que debe elegirse para que el error de la estimación de la diferencia de medias de los autos sea menor a 1 km.
    - c) Si el límite inferior del I.C. vale  $-3$ , ¿cuánto corresponde al límite superior y cuál es el nivel de significancia  $1 - \alpha$ ?
  2. Del ejercicio anterior sobre los autos VG y NS con 95% de confianza:
    - a) Construya un intervalo para la diferencia del rendimiento promedio por litro, de los dos automóviles suponga que  $\sigma_V^2 = \sigma_N^2$  desconocidas.
    - b) Si el límite inferior del I.C. vale  $-3$ , ¿cuánto corresponde al límite superior y cuál es el nivel de significancia  $1 - \alpha$ ?
  3. Del ejercicio anterior sobre los autos VG y NS con 95% de confianza:
    - a) Construya un intervalo para la diferencia del rendimiento promedio por litro, suponga  $\sigma_V^2 \neq \sigma_N^2$  desconocidas.
    - b) Si el límite inferior del I.C. vale  $-3$ , ¿cuánto corresponde al límite superior y cuál es el nivel de significancia  $1 - \alpha$ ?
  4. Del ejercicio anterior sobre los autos VG y NS con 95% de confianza:
    - a) Construya un intervalo para la razón entre varianzas del rendimiento por litro para los dos tipos de automóviles. ¿Qué suposición  $\sigma_V^2 = \sigma_N^2$  o  $\sigma_V^2 \neq \sigma_N^2$  debe considerarse válida?
    - b) Si el límite inferior del I.C. en a) vale 0.5 ¿cuánto corresponde al límite superior y cuál es el nivel de significancia  $1 - \alpha$ ?

*Sugerencia.* Efectúe los cálculos en Excel.

5. Una revista publicó los resultados de un estudio sobre la relación entre la participación en los deportes y la destreza manual. De una muestra aleatoria de 37 alumnos de segundo grado que participaron en los deportes, se obtuvo una calificación media de destreza manual de 32.19 y una desviación estándar de 3.34. De una muestra aleatoria independiente de 40 alumnos de segundo grado que no participaron en los deportes, se calculó una calificación media de destreza manual de 31.68 y una desviación estándar de 4.56.
- a) Para resolver los siguientes incisos, ¿es necesario suponer la normalidad en las dos poblaciones anteriores? Explique su respuesta.  
Con 90% de confianza:
- b) Construya un intervalo para la diferencia de la destreza manual media, suponga que  $\sigma_d^2 = 10$  y  $\sigma_{nd}^2 = 18$ .
- c) Determine el tamaño mínimo de la muestra que debe elegirse para que el error de la estimación de la diferencia de medias de la destreza manual sea menor a 1.5.
- d) Si el límite superior del I.C. vale 2, ¿cuánto corresponde al límite inferior y cuál es  $1 - \alpha$ ?
6. Del ejercicio anterior sobre la destreza manual, con 90% de confianza:
- a) Construya un intervalo para la diferencia de la destreza manual media, suponga  $\sigma_d^2 = \sigma_{nd}^2$  desconocidas.
- b) Si el límite superior del I.C. vale 2, ¿cuánto corresponde al límite inferior y cuál es  $1 - \alpha$ ?
7. Del ejercicio anterior sobre la destreza manual, con 90% de confianza:
- a) Construya un intervalo para la diferencia de la destreza manual media, suponga  $\sigma_d^2 < \sigma_{nd}^2$ , desconocidas.
- b) Si el límite superior del I.C. vale 2, ¿cuánto corresponde al límite inferior y cuál es  $1 - \alpha$ ?
8. Del ejercicio anterior sobre la destreza manual, con 90% de confianza:
- a) Construya un intervalo para la razón entre varianzas de la destreza de participantes y no participantes a los deportes. ¿Qué suposición  $\sigma_d^2 = \sigma_{nd}^2$  o  $\sigma_d^2 < \sigma_{nd}^2$  debe considerarse válida?
- b) Si el límite inferior del I.C. vale 0.25 ¿cuánto corresponde al límite superior y cuál es  $1 - \alpha$ ?
- Sugerencia.* Efectúe los cálculos en Excel.
9. Se dice que una nueva dieta reduce el peso de una persona en promedio 4.5 kg en dos semanas. Los pesos de siete mujeres que siguieron esta dieta fueron anotados antes y después de este periodo. Suponga que la distribución de las diferencias de cada persona es aproximadamente normal.

**Tabla 3.10**

Mujer	1	2	3	4	5	6	7
Peso anterior	58.5	60.3	61.7	69.0	64.0	62.6	56.7
Peso posterior	60.0	54.9	58.1	62.1	58.5	59.9	54.4

Con 96% de confianza:

- a) Construya un intervalo para la diferencia de las medias de los pesos y pruebe la eficacia de la dieta.
- b) Si el límite superior del I.C. vale 5, ¿cuál es el límite inferior y con qué  $1 - \alpha$  se obtiene?
10. Del ejercicio anterior sobre la nueva dieta para reducir el peso, con 96% de confianza:
- a) Construya un intervalo para la varianza de las diferencias de los pesos.
- b) Si el límite superior del I.C. es 25, ¿cuál es el límite inferior y con qué  $1 - \alpha$  se obtiene?
- c) ¿Qué puede decir con respecto de la dieta para reducir el peso?
11. Diez animales fueron sometidos a condiciones que simulaban una enfermedad. Se registró el número de latidos del corazón, antes y después del experimento. Si se supone que la distribución de las diferencias de los latidos del corazón antes y después del experimento son normales.

**Tabla 3.11**

Antes	70	84	88	110	105	100	110	67	79	86
Después	115	148	176	191	158	178	179	148	161	157

Con 90% de confianza:

- a) Construya un intervalo para la diferencia de las medias de los latidos y pruebe si aumentan en promedio más de 65.
- b) Si el límite superior del I.C. vale  $-60$ , ¿cuál es el límite inferior y con qué  $1 - \alpha$  se obtiene?
12. Del ejercicio anterior sobre los latidos del corazón de los animales, con 90% de confianza:
- a) Construya un intervalo para la desviación estándar de las diferencias de los latidos.
- b) Si el límite superior del I.C. es 24, ¿cuál es el límite inferior y con qué  $1 - \alpha$  se obtiene?
- c) ¿Qué puede decir con respecto del experimento sobre el aumento de los latidos del corazón?
13. Para averiguar si un nuevo suero detendrá o no a la leucemia se seleccionan 18 ratones que han alcanzado un estado avanzado de la enfermedad, nueve ratones reciben el tratamiento y nueve no. Los tiempos de supervivencia en años desde que se inició el experimento se aprecian en la tabla 3.12.

**Tabla 3.12**

Con tratamiento	0.8	3.2	2.7	5.2	3.7	4.4	5.3	3.4	2.6
Sin tratamiento	1.9	2.1	2.6	4.5	2.2	2.1	1.2	2.8	0.8

Suponga que los tiempos de supervivencia siguen una distribución normal. Con 95% de confianza:

- a) Construya un intervalo para la diferencia de los tiempos promedio de supervivencia de los ratones con y sin tratamiento.
- Al suponer varianzas iguales.
  - Al suponer varianzas diferentes.
- b) Con los resultados anteriores concluya con 95% de confianza si el nuevo suero prolonga el tiempo de supervivencia de los ratones.
14. Del ejercicio anterior del nuevo suero, ¿será posible suponer que son muestras pareadas? En caso afirmativo, agregue los supuestos y resuelva con 95% de confianza para la diferencia de medias y varianzas.
15. Del ejercicio anterior del nuevo suero con 95% de confianza:
- a) Construya un intervalo para la razón de varianzas de los tiempos de supervivencia de los ratones con y sin tratamiento.
- b) Se puede asegurar que el nuevo suero prolonga el tiempo de supervivencia de los ratones en forma más homogénea.
16. Los siguientes datos fueron recabados en un experimento diseñado para verificar si existe una diferencia sistemática en los pesos obtenidos con dos escalas. Es decir, cada objeto es pesado en ambas escalas. Suponga normalidad en las diferencias de los pesos para ambas escalas.

**Tabla 3.13**

Escala 1	11.23	14.36	8.33	10.5	23.42	9.15	13.47	60.47	12.40	19.38
Escala 2	11.27	14.41	8.35	10.52	23.41	9.17	13.52	60.46	12.45	19.35

Con 92% de confianza.

- a) Construya un intervalo para la diferencia de las medias de las escalas y pruebe si no hay diferencias entre éstas.
- b) Si el límite inferior del I.C. vale  $-0.04$ , ¿cuál es el límite superior y con qué  $1 - \alpha$  se obtiene?
17. Del ejercicio anterior de las dos escalas de medición con 95% de confianza:
- a) Construya un intervalo para la desviación estándar de las diferencias de los pesos en ambas escalas.
- b) Si el límite inferior del I.C. es 0.02, ¿cuál es el límite superior y con qué  $1 - \alpha$  se obtiene?
18. Cierta metal se produce, por lo regular, mediante un proceso estándar. Se desarrolla un nuevo proceso en el que se añade una aleación a la producción del metal. Los fabricantes se encuentran interesados en estimar la

verdadera diferencia entre las tensiones de ruptura de los metales producidos por los dos procesos. Para cada metal se seleccionan al azar ocho especímenes y cada uno se somete a una tensión hasta que se rompa. La tabla 3.14 muestra las tensiones de ruptura de los especímenes en kilogramos por centímetro cuadrado.

Tabla 3.14

Proceso estándar	428	419	458	439	441	456	463	429
Proceso nuevo	462	448	435	465	429	472	453	459

Si se supone que el muestreo se llevó a cabo en dos distribuciones normales e independientes, calcule un intervalo de confianza de 98% para la razón de varianzas de los procesos de producción de metales.

19. Con base en el resultado del ejercicio anterior:

- Construya un intervalo con 98% de confianza para la diferencia promedio de las tensiones.
- ¿Se estaría inclinando a concluir con 98% de confianza que existe una diferencia real entre  $\mu_e$  y  $\mu_n$  o que el nuevo proceso tiene una producción más homogénea de metales?

20. Un centro de investigación en medicina del deporte informó en mayo de 2015, respecto a las diferencias en las tasas de consumo de oxígeno para varones universitarios entrenados con dos métodos diferentes e independientes, que uno utiliza entrenamiento intermitente y otro continuo con igual duración. En la tabla 3.15 se registran los resultados, expresados en ml por kg/min de dos muestras aleatorias independientes.

Suponga que las poblaciones tienen distribución normal. Calcule un I.C. de 99% para la razón de varianzas entre ambos métodos.

Tabla 3.15

Entrenamiento continuo	Entrenamiento intermitente
$n_i = 27$	$n_c = 22$
$\bar{x}_i = 39.63$	$\bar{x}_c = 43.71$
$s_i = 9.68$	$s_c = 4.87$

21. Con base en el resultado del ejercicio anterior:

- Construya un intervalo con 99% de confianza para la diferencia promedio de los dos métodos de entrenamiento.
- ¿Se puede concluir con 99% de confianza que existe una diferencia real entre ambos métodos o si alguno de éstos tiene una preparación más homogénea?

22. Los IPC de las empresas TV y CC se muestran en la tabla 3.16. Suponga que el IPC de las empresas tiene una distribución normal y que las muestras son independientes.

- Calcule un I.C. de 90% para la razón de varianzas entre ambos IPC.
- ¿Es posible afirmar con 90% de confianza qué empresa tiene mayor varianza del IPC? Justifique su respuesta.

Tabla 3.16

Fecha	TV	CC
09/09/2013	27.70	23.31
09/08/2013	27.63	23.50
09/07/2013	27.67	23.70
09/06/2013	27.60	23.25
09/03/2013	27.44	23.50

Tabla 3.16 (Continuación)

Fecha	TV	CC
09/02/2013	27.87	23.40
09/01/2013	27.69	23.20
08/31/2013	27.40	22.90
08/30/2013	27.44	22.80
08/27/2013	27.48	22.46
08/26/2013	27.07	22.46
08/25/2013	26.99	22.46

24. Con base en el resultado del ejercicio anterior:

- a) Construya un intervalo con 90% de confianza para la diferencia promedio del IPC de las dos empresas. Es posible afirmar que el IPC medio de TV es mayor a 3.5 unidades que el de CC.
- b) ¿Se puede suponer que se trata de muestras pareadas? Explique su respuesta.

25. Las personas que remiten trabajos de cálculo a un centro de cómputo necesitan estimar la cantidad de tiempo requerida por la computadora para terminarlo. Este tiempo se mide en la CPU en un centro de cómputo. Se consideró a 11 usuarios para que dieran una estimación y se registró su tiempo real en la tabla 3.17.

Tabla 3.17

#### Número de trabajo

Tiempo de CPU (min)	1	2	3	4	5	6	7	8	9	10	11
Estimado	0.50	1.40	0.95	0.45	0.25	1.20	1.60	2.6	1.30	0.35	0.80
Real	1.46	1.52	0.09	0.33	0.71	1.31	1.49	2.9	1.41	0.83	0.74

Suponga normalidad en los tiempos real y estimado, además de normalidad en sus diferencias. Se desea conocer por medio de una estimación para la diferencia de medias de ambos tiempos y una probabilidad de 0.96 si el cliente subestima el tiempo en CPU necesario para los trabajos de cómputo.

- a) Respecto al material de esta unidad, ¿qué recomendaría al cliente utilizar para resolver el problema?
- b) ¿Será posible suponer muestras pareadas? Justifique su respuesta.
- c) Con base en la respuesta de los incisos a) y b) resuelva el problema.

26. Se realizó un experimento para comparar los tiempos medios necesarios para que dos empleados, A y B, completen el trámite de las cuentas corrientes personales para nuevos clientes. Se asignaron al azar diez clientes a cada empleado y se registraron los tiempos de servicio para cada uno. Los resultados obtenidos fueron:

$$A: \sum_{i=1}^{10} X_i = 222 \text{ y } \sum_{i=1}^{10} X_i^2 = 5,075.6400; B: \sum_{i=1}^{10} Y_i = 285 \text{ y } \sum_{i=1}^{10} Y_i^2 = 8,292.78$$

Si supone que los tiempos de servicio por empleado tienen una distribución normal, calcule un intervalo de confianza de 95% para la razón de varianzas de los tiempos requeridos por los empleados para completar el trámite de las cuentas corrientes personales para nuevos clientes. ¿Parece que la variabilidad del empleado B es mayor? Explique su respuesta.

27. Con base en el resultado del ejercicio anterior construya un intervalo de 95% de confianza para la diferencia de los tiempos medios de servicio por parte de los empleados. ¿Parece que la media del empleado B es mayor? Explique su respuesta.



### 3.5 Intervalos de confianza para proporciones

A lo largo del texto se ha visto que una de las distribuciones más importantes es la normal, de forma similar se ha podido notar que la distribución binomial adquiere una gran importancia en el estudio de poblaciones que se dividen en dos clases. La construcción de los intervalos de confianza para las proporciones con amplitud mínima es un poco “artística”, pero si nos restringimos a muestras grandes del TCL podemos utilizar la aproximación para sumas y promedios.

Se ha visto que las proporciones surgen en situaciones como las de los ejemplos siguientes.

#### Ejemplos 3.34 Estimación de las proporciones

1. En la Ciudad de México es de interés conocer la proporción de habitantes que está a favor del metrobús. En este caso se considera la población en dos clases, a favor o en contra de la construcción de este medio de transporte.
2. Se lleva a cabo un estudio para conocer si la ciudadanía de la Ciudad de México está de acuerdo con la aplicación del alcoholímetro. En este caso se considera la población en dos clases, a favor y en contra del alcoholímetro.
3. La producción de tornillos se puede clasificar en buenos y defectuosos, pero es de mayor interés conocer la proporción de tornillos buenos.
4. En diciembre de 2013 se llevó a cabo una encuesta para determinar si los usuarios del Sistema de Transporte Colectivo Metro de la Ciudad de México estaban a favor o en contra del alza del precio del boleto a \$5.

Se pueden dar muchos ejemplos sobre casos en los que se tiene interés por llevar a cabo una estimación de las proporciones.

### Intervalos de confianza para proporciones de muestras grandes

En la unidad 2 se introdujo el concepto de una proporción como el cociente de una variable aleatoria  $X$  con distribución binomial y parámetros  $n$  y  $p$ , entre  $n$ . A la proporción se le denotó por  $\hat{P}$ . Es decir:

$$\hat{P} = \bar{X}$$

con  $E(\bar{P}) = p$  y  $V(\bar{P}) = \frac{p(1-p)}{n}$ , y la distribución de  $\hat{P}$  no es normal, pero en el caso de tamaños de muestra grande podemos utilizar el teorema central del límite para determinar un intervalo de confianza.

#### Teorema 3.15

Si  $\hat{p}$  es la proporción de éxitos en la realización de una muestra aleatoria de tamaño  $n$  ( $n > 30$ ), el intervalo de confianza de  $(1 - \alpha)$  100% para el parámetro binomial  $p$  se puede aproximar

$$\hat{p} - Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

donde  $Z_{\alpha/2}$  es el valor de la distribución normal estándar a la derecha del cual se tiene un área de  $\alpha/2$ , y  $\hat{p} = x/n$  el valor de la proporción y  $x$  la cantidad de éxitos en una realización.

En los siguientes ejemplos se muestra el uso del teorema 3.15.



**Ejemplo 3.35** Intervalos de confianza para proporciones de muestras grandes

En una muestra aleatoria de 100 posibles clientes, 70 dicen que prefieren el producto A. Calcule un intervalo con 95% de confianza para la proporción de todos los posibles clientes que prefieren el producto A.

**Solución**

Para el intervalo de confianza de la proporción primero se calcula el valor de la proporción de personas que prefieren el producto:

$$\hat{p} = \frac{70}{100} = 0.70 \text{ y } \hat{q} = 1 - \hat{p} = \frac{30}{100} = 0.30$$

El grado de confianza en este caso es de 95%; por tanto,  $1 - \alpha = 0.95$  y al buscar en las tablas porcentuales de la distribución normal se tiene  $Z_{\frac{\alpha}{2}} = 1.96$ . Por último:

$$0.70 - 1.96\sqrt{\frac{0.70 \times 0.30}{100}} < p < 0.70 + 1.96\sqrt{\frac{0.70 \times 0.30}{100}} \Rightarrow 0.6102 < p < 0.7898$$

Por tanto, se dice que la proporción de clientes que prefieren el producto A está entre 61.02 y 78.98% con una confianza de 95%.

**Ejemplos variados para proporciones**

Con base en la distribución normal, se puede utilizar su simetría para el error por estimación:

$$\varepsilon = |p - \hat{p}| = Z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}$$

Con el error se calculan tamaños de muestras y coeficientes de confianza. Pero se tiene el mismo problema que en los intervalos de confianza, se requiere del mismo parámetro  $p$ . En este sentido, ¿cómo calcular el tamaño de muestra en proporciones?

Para calcular el tamaño mínimo de muestra que cumpla con el tamaño de error por estimación establecido de antemano, se puede proceder de diferentes formas.

**Con una estimación puntual preliminar**

Se hace un muestreo preliminar para el valor puntual de estimación,  $\hat{p}$ , que se coloca en la expresión para el error por estimación.

$$\varepsilon = Z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Ahora, se despeja la  $n$  y se considera su parte entera más uno para obtener el tamaño mínimo de muestra que cumpla con el error indicado:

$$n = \left[ \left( \frac{Z_{\alpha/2}}{\varepsilon} \right)^2 \hat{p}(1-\hat{p}) \right] + 1 \quad (3.1)$$

**Con una cota inferior**

Se despeja el tamaño de muestra del error por estimación, sin sustituir la estimación puntual  $\varepsilon = Z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}$ . Lo que resulta:

$$n \geq \left( \frac{Z_{\alpha/2}}{\varepsilon} \right)^2 p(1-p)$$

Por último, se acota la expresión  $p(1-p)$  por su máximo, que se obtiene por máximos relativos, lo que resulta:

$$p(1-p) \leq \frac{1}{2} \left( 1 - \frac{1}{2} \right) = \frac{1}{4} \text{ para todo } p.$$

Entonces, en este caso el tamaño mínimo de muestra que se requiere que cumpla con el error es:

$$n = \left\lceil \frac{1}{4} \left( \frac{Z_{\alpha/2}}{\varepsilon} \right)^2 \right\rceil + 1 \quad (3.2)$$

Veamos algunos ejemplos que ilustren cómo calcular el tamaño óptimo de la muestra que cumpla con un error deseado.

### Ejemplo 3.36 Tamaño de muestra en proporciones

Se efectúa un estudio para estimar la proporción de amas de casa que poseen una secadora automática. ¿Cuál es el tamaño mínimo de la muestra que se requiere si se desea tener una confianza de 99% de que la estimación difiera de la verdadera proporción en 0.01?

- a) Resuelva con el uso de la cota inferior.  
 b) Resuelva con una estimación preliminar. Si de 100 amas de casa entrevistadas 20 tenían una secadora automática.

#### Solución

- a) Para la cota inferior solo se requiere del valor  $Z_{\alpha/2}$ . Para esto se tiene  $1 - \alpha = 0.99$ , donde  $\alpha/2 = 0.005 = 0.5\%$ . Por tanto, de las tablas porcentuales de la distribución  $Z$  resulta:

$$Z_{\alpha/2} = 2.5758$$

Mientras que el error por estimación vale  $\varepsilon = 0.01$ . Si se sustituyen estos dos valores en la fórmula (3.2):

$$n = \left\lceil \frac{1}{4} \left( \frac{Z_{\alpha/2}}{\varepsilon} \right)^2 \right\rceil + 1 = \left\lceil \frac{1}{4} \left( \frac{2.5758}{0.01} \right)^2 \right\rceil + 1 = [16586.86] + 1 = 16587.86$$

Luego, el tamaño mínimo de la muestra que se debe seleccionar es  $n \geq 16587$ .

- b) Si se considera la estimación preliminar  $\hat{p} = \frac{20}{100} = 0.20$  y la fórmula 3.1, tenemos:

$$n = \left\lceil \left( \frac{Z_{\alpha/2}}{\varepsilon} \right)^2 \hat{p}(1-\hat{p}) \right\rceil + 1 = \left\lceil \left( \frac{2.5758}{0.01} \right)^2 (0.2)(0.8) \right\rceil + 1 = [10615.59] + 1 = 10616.59$$

Luego, el tamaño mínimo de la muestra que se debe seleccionar es  $n \geq 10616$ . Tamaño menor a la cota inferior resultante sin considerar una estimación preliminar para  $p$ .

Ahora bien, ¿cómo determinar el grado de confianza cuando se tienen los límites del intervalo?

En el caso de las proporciones, el grado de confianza cuando se dan los límites del intervalo está dado en el teorema 3.16.

**Teorema 3.16**

Si  $\hat{p}$  es la proporción de éxitos en la realización de una muestra aleatoria de tamaño  $n$  ( $n > 30$ ), y  $a < p < b$  un intervalo del  $(1 - \alpha)$  100% de confianza para el parámetro  $p$ , entonces:

$$1 - \alpha = 1 - 2\Phi\left((a - \hat{p})\sqrt{\frac{n}{\hat{p}(1 - \hat{p})}}\right) = 2\Phi\left((b - \hat{p})\sqrt{\frac{n}{\hat{p}(1 - \hat{p})}}\right) - 1$$

donde  $\Phi$  es la distribución acumulada normal estándar.

**Ejemplo 3.37 Aplicaciones del teorema 3.16**

Un ingeniero de control de calidad de una línea de producción desea entregar un reporte basado en un intervalo de confianza para la estimación de la proporción de artículos defectuosos producidos en la línea. Elige una muestra de 200 artículos y encuentra que 10 son defectuosos; dados los resultados de la muestra y las exigencias de la gerencia, quiere entregar en su reporte un intervalo en el que la proporción de defectuosos sea menor a 7%. El problema es que no sabe el grado de confianza que satisface sus necesidades. ¿Cuál será el límite inferior del reporte y con qué grado de confianza lo justificaría?

**Solución**

Tenemos  $n = 200$  y la proporción de defectuosos para la muestra es  $p = 10/200 = 0.05$ , se desea que el reporte indique  $p < 0.07$ . Ahora  $\hat{p} = 0.05$  es el centro del intervalo de confianza, en el cual la distancia de  $\hat{p}$  a 0.07 es 0.02, debe ser la misma distancia de  $\hat{p}$  al límite inferior,  $0.05 - 0.02 = 0.03$  límite inferior, donde el intervalo de confianza  $0.03 < p < 0.07$ . Para el grado de confianza utilizamos el teorema anterior:

$$\begin{aligned} 1 - \alpha &= 2\Phi\left((b - \hat{p})\sqrt{\frac{n}{\hat{p}(1 - \hat{p})}}\right) - 1 = 2\Phi\left((0.07 - 0.05)\sqrt{\frac{200}{0.05(0.95)}}\right) - 1 \\ &= 2\Phi(1.2977) - 1 = 0.8064 \end{aligned}$$

La proporción de artículos defectuosos de la línea  $p \in (0.03, 0.07)$  con una confianza de 80.64%.

**Intervalo de confianza de diferencia de proporciones para muestras grandes**

Con frecuencia se utilizan las proporciones para estudiar las preferencias entre dos productos, o la mejora entre dos procesos de producción, entre otros fines. En todos los casos en los que las poblaciones se dividen en dos clases se puede utilizar la diferencia de proporciones para muestras grandes (véase TCL).

**Teorema 3.17**

Si  $\hat{p}_1$  y  $\hat{p}_2$  son las proporciones de éxitos de las realizaciones de dos muestras aleatorias independientes de tamaños  $n_1$  y  $n_2$  ( $n_1 > 30$  y  $n_2 > 30$ ), entonces el intervalo de confianza del  $(1 - \alpha)$  100% para  $p_1 - p_2$  está dado por:

$$(\hat{p}_1 - \hat{p}_2) - Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

donde  $Z_{\alpha/2}$  es el valor de la distribución normal estándar con área derecha de  $\alpha/2$ .

Para la comparación de diferencia de proporciones se tienen los mismos casos que en la comparación de diferencia de medias.

**Ejemplo 3.38 Intervalo de confianza de diferencia de proporciones**

Se lleva a cabo una prueba clínica para conocer si determinada inoculación afecta la incidencia de una enfermedad. Se conservó una muestra de 1 000 ratas en un ambiente controlado durante un año, 500 de las cuales fueron inoculadas. En el grupo al que no se le aplicó la droga hubo 120 casos de esta enfermedad, mientras que del grupo tratado con la droga, 90 la contrajeron. Si  $p_1$  representa la probabilidad de incidencia de la enfermedad en las ratas no tratadas y  $p_2$  la probabilidad de incidencia después de que recibieron la droga, calcule un intervalo de 95% de confianza para  $p_1 - p_2$  e indique si la proporción de incidencias en la enfermedad es mayor en las ratas no tratadas.

**Solución**

Como  $p_1$  es la probabilidad de incidencia de la enfermedad en las ratas no tratadas se tiene que en 500 ratas no tratadas 120 contrajeron la enfermedad, es decir:

$$\hat{p}_1 = \frac{120}{500} = 0.24, \text{ de manera que } \hat{q}_1 = 0.76 \text{ con } n_1 = 500$$

De la misma manera, para las ratas que fueron tratadas se tiene:

$$\hat{p}_2 = \frac{90}{500} = 0.18, \text{ de manera que } \hat{q}_2 = 0.82 \text{ con } n_2 = 500$$

Por último, para el intervalo de 95% de confianza tenemos que  $\alpha/2 = 0.025$  y de las tablas porcentuales para la distribución normal  $Z_{0.025} = 1.96$ . Luego, con la fórmula del teorema tenemos para 3.16  $p_1 - p_2$ :

$$(0.24 - 0.18) - 1.96\sqrt{\frac{0.24 \times 0.76}{500} + \frac{0.18 \times 0.82}{500}} < p_1 - p_2 < (0.24 - 0.18) + 1.96\sqrt{\frac{0.24 \times 0.76}{500} + \frac{0.18 \times 0.82}{500}}$$

$$0.0096 < p_1 - p_2 < 0.1104$$

Del intervalo de confianza para la diferencia de proporciones, se observa que no hay valores negativos. Luego, a 95% de confianza se puede asegurar que la proporción de ratas que contrajeron la enfermedad fue mayor en el caso en que no se aplicó la droga.

**Tamaño de muestras en diferencia de proporciones**

Con base en la distribución normal, se utiliza su simetría para el error por estimación, de lo que se obtiene el error en la diferencia de proporciones

$$\varepsilon = |(p_1 - p_2) - (\hat{p}_1 - \hat{p}_2)| = Z_{\alpha/2} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Con el error es posible calcular los tamaños de muestra y coeficientes de confianza. Pero se tiene el mismo problema que en los intervalos de confianza para las proporciones, se requiere de los mismos parámetros  $p_1$  y  $p_2$ . En este sentido, ¿cómo calcular el tamaño mínimo de las muestras en diferencia de proporciones que cumplan con un error dado por estimación?

Para calcular el tamaño mínimo de la muestra, si se conoce el error por estimación se puede hacer de forma similar que en las proporciones, pero solo para los casos en que las dos muestras sean del mismo tamaño  $n_1 = n_2 = n$ .

**Con una estimación puntual preliminar**

Al hacer un muestreo preliminar para el valor puntual de estimación,  $\hat{p}_1$  y  $\hat{p}_2$  y el mismo tamaño de muestra, los cuales se colocan en la expresión para el error por estimación.

$$\varepsilon = Z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{n}} = Z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1) + \hat{p}_2(1-\hat{p}_2)}{n}}$$

Ahora, se despeja  $n$  y se obtiene el tamaño mínimo de la muestra que cumple con el error por estimación  $\varepsilon$ :

$$n = \left[ \left( \frac{Z_{\alpha/2}}{\varepsilon} \right)^2 (\hat{p}_1(1-\hat{p}_1) + \hat{p}_2(1-\hat{p}_2)) \right] \quad (3.3)$$

### Con una cota inferior

Si se despeja el tamaño de muestra del error por estimación, sin sustituir la estimación puntual,  $\varepsilon = Z_{\alpha/2}$

$\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{n}}$ , queda:

$$n = \left[ \left( \frac{Z_{\alpha/2}}{\varepsilon} \right)^2 (p_1(1-p_1) + p_2(1-p_2)) \right]$$

Por último, se acota la expresión  $p_1(1-p_1) + p_2(1-p_2)$  por su máximo, que resulta,  $p_1(1-p_1) + p_2(1-p_2) \leq 1/4 + 1/4 = 1/2$  para todo  $p_1$  y  $p_2$ .

Entonces, el tamaño mínimo de la muestra que cumple con el error por estimación  $\varepsilon$ :

$$n = \left[ \frac{1}{2} \left( \frac{Z_{\alpha/2}}{\varepsilon} \right)^2 \right] \quad (3.4)$$

Ahora bien, ¿cómo determinar el grado de confianza cuando se tienen los límites del intervalo?

En el caso de la diferencia de proporciones, el grado de confianza cuando se dan los límites del intervalo está dado en el teorema 3.18.

#### Teorema 3.18

Si  $p_1$  y  $p_2$  son las proporciones de éxitos de las realizaciones de dos muestras aleatorias independientes de tamaños  $n_1$  y  $n_2$  ( $n_1 > 30$  y  $n_2 > 30$ ), y  $a < p_1 - p_2 < b$  un intervalo del  $(1 - \alpha)$  100% de confianza para la diferencia de  $p_1 - p_2$ , entonces:

$$1 - \alpha = 1 - 2\Phi\left(\frac{a - (\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}}\right) = 2\Phi\left(\frac{b - (\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}}\right) - 1$$

Donde  $\Phi$  es la distribución acumulada normal estándar.

#### Ejemplo 3.39 Cota inferior

Se está analizando la fracción de productos defectuosos provenientes de dos líneas de producción. Con una muestra aleatoria previa de 400 unidades provenientes de la línea 1 se obtuvo 7.5% defectuosas, se toma otra muestra aleatoria previa de 500 unidades provenientes de la línea 2 y 6.3% resultan defectuosas.

a) Con 90% de confianza calcule los tamaños de muestra mínimos que se requieren para que la diferencia de las proporciones muestrales se desvíe de la verdadera diferencia de proporciones en menos de 2%.

- b) Si el límite inferior del intervalo de confianza para la diferencia de proporciones es  $-0.021$ , ¿cuál es el nivel de confianza que cumple con este valor?
- c) Si el límite superior del intervalo de confianza para la diferencia de proporciones es  $0.05$ , ¿cuál es el nivel de confianza que cumple con este valor?

### Solución

- a) Tenemos  $1 - \alpha = 0.90$ , donde  $Z_{\alpha/2} = 1.645$  y  $\varepsilon = 0.02$ , con  $\hat{p}_1 = 0.075$  y  $\hat{p}_2 = 0.063$ . Luego, con las proporciones previas:

$$n = \left\{ \left( \frac{Z_{\alpha/2}}{\varepsilon} \right)^2 [\hat{p}_1(1 - \hat{p}_1) + \hat{p}_2(1 - \hat{p}_2)] \right\} = \left\{ \left( \frac{1.645}{0.02} \right)^2 [0.075(0.925) + 0.063(0.937)] \right\} = 869$$

Con la cota inferior:

$$n = \left[ \frac{1}{2} \left( \frac{Z_{\alpha/2}}{\varepsilon} \right)^2 \right] = \left[ \frac{1}{2} \left( \frac{1.645}{0.02} \right)^2 \right] = 3382 \text{ mayor que sin la cota inferior.}$$

- b) Para el límite inferior:

$$1 - \alpha = 1 - 2\Phi \left( \frac{a - (\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}_1(1 - \hat{p}_1)/n_1 + \hat{p}_2(1 - \hat{p}_2)/n_2}} \right) = 1 - 2\Phi \left( \frac{-0.021 - (0.075 - 0.063)}{\sqrt{0.075(0.925)/400 + 0.063(0.937/500)}} \right) \\ = 0.9467$$

- c) Para el límite superior:

$$1 - \alpha = 2\Phi \left( \frac{b - (\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}_1(1 - \hat{p}_1)/n_1 + \hat{p}_2(1 - \hat{p}_2)/n_2}} \right) - 1 = 2\Phi \left( \frac{0.05 - (0.075 - 0.063)}{\sqrt{0.075(0.925)/400 + 0.063(0.937/500)}} \right) - 1 \\ = 0.9740$$

## Ejercicios 3.7

### Instrucciones

- I.C. denota intervalo de confianza  $1 - \alpha$ , nivel de confianza.
  - En cada ejercicio el tamaño mínimo de la muestra debe encontrarse para la cota inferior.
1. En cierto proceso industrial la proporción de un muestreo previo de artículos defectuosos indica que es de 5%. Con una confianza de 98% obtenga:
    - a) Un I.C. para la verdadera proporción de artículos defectuosos  $n = 120$ .
    - b) El límite inferior del intervalo de confianza de  $p$  es  $0.02$ , determine el límite superior y  $1 - \alpha$ .
  2. Del ejercicio anterior calcule el tamaño mínimo de la muestra que debe seleccionarse para que la proporción de artículos defectuosos se aleje de la verdadera proporción en menos de 4%.
    - a) Al considerar la proporción previa del ejercicio anterior.
    - b) Para la cota inferior.
  3. Los propietarios de una empresa que fabrica baterías para linterna detectaron cierta cantidad de artículos defectuosos. Para estimar la proporción de artículos buenos realizan un muestreo de tamaño 250, del que obtienen 10 baterías defectuosas. Con una confianza de 98% obtenga:
    - a) Un I.C. para la verdadera proporción de baterías buenas.
    - b) El límite inferior del intervalo de confianza de  $p$  es  $0.94$ , determine el límite superior y  $1 - \alpha$ .

4. Del ejercicio anterior calcule el tamaño mínimo de la muestra que debe seleccionarse para que la proporción muestral de baterías buenas se aleje de la verdadera proporción en menos de 1%.
  - a) Al considerar la proporción previa del ejercicio anterior.
  - b) Para la cota inferior.
5. En la Ciudad de México se quiere obtener una estimación de la proporción de amas de casa que poseen una secadora automática. Se eligió una muestra de 150 amas de casa y 60 contestaron que sí. Con una confianza de 90% obtenga:
  - a) Un I.C. para la verdadera proporción de amas de casa que poseen una secadora automática.
  - b) El límite inferior del intervalo de confianza de  $p$  es 0.02, determine el límite superior y  $1 - \alpha$ .
6. Del ejercicio anterior calcule el tamaño mínimo de la muestra que debe seleccionarse para que la proporción muestral de amas de casa que poseen una secadora automática se aleje de la verdadera proporción en menos de 5%.
  - a) Al considerar la proporción previa del ejercicio anterior.
  - b) Para la cota inferior.
7. El gobierno de México afirma que el porcentaje de desempleados en el país de personas en edad económicamente activa es de 20% con una probabilidad de 0.98. Para probar estadísticamente si la afirmación del gobierno es cierta se elige una muestra de 400 personas en edad económicamente activa, de lo que resultan 90 desempleadas.
  - a) ¿Es cierta la afirmación del gobierno de México? Justifique su respuesta.
  - b) El límite superior del intervalo de confianza de  $p$  es 0.26, determine el límite inferior y  $1 - \alpha$ .
8. Del ejercicio anterior calcule el tamaño mínimo de la muestra que debe seleccionarse para que la proporción muestral se desvíe máximo 5% de la verdadera proporción en 98% de los casos.
  - a) Considerando la proporción previa del ejercicio anterior.
  - b) Para la cota inferior.
9. Si un estudio preliminar revela que 56 de 200 fumadores prefieren la marca  $A$  de cigarrillos, ¿qué tan grande debe ser una muestra para la preferencia de la marca  $A$ , si se desea tener una confianza de 99% de que esté dentro del 0.05 de la proporción real de fumadores que prefieren la marca  $A$  de cigarrillos?
10. Los fabricantes de un nuevo refresco de cola afirman que en la actualidad más de 20% de los habitantes de la Ciudad de México y área metropolitana consumen su producto. Para verificar de manera estadística y con una confianza de 95% la afirmación de los fabricantes, fue seleccionada una muestra aleatoria de 400 ciudadanos, de los cuales 70 contestaron que sí lo consumen.
  - a) ¿Es cierta la afirmación de los fabricantes? Justifique su respuesta.
  - b) El límite inferior del intervalo de confianza de  $p$  es 0.145. Determine el límite superior y  $1 - \alpha$ .
11. Del ejercicio anterior calcule el tamaño mínimo de la muestra que debe seleccionarse para que la proporción muestral se desvíe máximo 2% de la verdadera proporción en 90% de los casos.
  - a) Considerando la proporción previa del ejercicio anterior.
  - b) Para la cota inferior.
12. Para evaluar la efectividad de un nuevo medicamento para tratar cierta enfermedad, se les administró a 650 pacientes que la padecían. Al término de tres días, se habían recuperado 420 pacientes. Determine un intervalo de 99% de confianza para la proporción de pacientes en los que es efectivo el nuevo medicamento.
13. En la zona centro de México se generó una gran polémica por la liquidación de los trabajadores del SME. El gobierno afirma que después de dar a conocer varias irregularidades en el sindicato la proporción de personas a favor de la liquidación es mayor a 60%. Para confirmar esta afirmación fue seleccionada una muestra aleatoria de 800 personas y 460 contestaron estar a favor de la liquidación. Con una confianza de 98% obtenga:
  - a) Un I.C. para la verdadera proporción de personas a favor de la liquidación.
  - b) ¿Se puede considerar válida la afirmación del gobierno mexicano? Justifique su respuesta.
  - c) El límite inferior del intervalo de confianza de  $p$  es 0.54, determine el límite superior y  $1 - \alpha$ .
14. Del ejercicio anterior calcule el tamaño mínimo de la muestra que debe seleccionarse para que la proporción muestral de personas a favor de esta liquidación se aleje de la verdadera proporción en menos de 0.025.

- a) Considerando la proporción previa del ejercicio anterior.  
 b) Para la cota inferior.

### Diferencia de proporciones

15. El gerente de la marca  $A$  de cigarros asegura que sobrepasa en ventas a su competencia, la marca  $B$ , en al menos 11% con una probabilidad de 0.95. Para comprobar de manera estadística la afirmación, el gerente realiza encuestas de forma independiente a dos grupos de fumadores. En el grupo 1 la pregunta fue: ¿prefiere la marca  $A$  de cigarros?, mientras que en el grupo 2 se preguntó: ¿prefiere la marca  $B$  de cigarros? En el grupo 1 de 200 personas 41 contestaron que sí, mientras que en el grupo 2, 18 de 150 respondieron de la misma manera. Con una confianza de 95% obtenga:
- a) Un I.C. para la verdadera diferencia de proporciones. ¿Se justifica la aseveración del gerente?  
 b) El límite inferior del I.C. de  $p_A - p_B$  es 0.02, determine el límite superior y  $1 - \alpha$ .
16. Del ejercicio anterior calcule el tamaño mínimo de la muestra que debe seleccionar el gerente para que la diferencia de proporciones muestrales de los fumadores se aleje de la verdadera diferencia de proporciones en menos de 0.04.
- a) Si se consideran las proporciones previas del ejercicio anterior.  
 b) Para la cota inferior.
17. En un estudio de dos tipos de crímenes cometidos por jóvenes delincuentes confinados en algunas instituciones correccionales, durante un periodo de 10 años se obtuvo.

Tabla 3.18

Tipo de crimen	Tamaño de la muestra	Cantidad de jóvenes que los cometieron
$A$	200	40
$B$	220	35

- a) Construya un intervalo con 95% de confianza para la diferencia de proporciones del tipo de crimen. Indique si es posible asegurar con 95% de confianza que la proporción del tipo de crimen  $A$  es mayor a la del  $B$ . Explique su respuesta.  
 b) El límite superior del I.C. de  $p_A - p_B$  es 0.10, determine el límite inferior y  $1 - \alpha$ .
18. Del ejercicio anterior calcule el tamaño mínimo de la muestra que debe seleccionar el gerente para que la diferencia de proporciones muestrales de los fumadores se aleje de la verdadera diferencia de proporciones en menos de 0.05.
- a) Considerando las proporciones previas del ejercicio anterior.  
 b) Para la cota inferior.
19. Antes de aprobar los matrimonios entre personas del mismo género, el gobierno de la Ciudad de México aseguraba que la proporción de personas de 18 a 24 años de edad a favor de estos matrimonios era de 66.3%, mientras que para las personas de 25 a 34 años era de 56.2%. Con esta información, los representantes del gobierno afirmaron que con 80% de confianza la proporción de personas a favor de 18 a 24 años superaba en más de 0.10 a la proporción de persona a favor de entre 25 y 34 años. Para probar esta afirmación fueron seleccionadas 500 personas de 18 a 24 años y 318 contestaron que estaban a favor, de forma independiente fueron elegidas otras 500 personas de 25 a 34 años y 296 contestaron lo mismo.
- a) Construya un intervalo con 80% de confianza para la diferencia de proporciones de personas con edades entre 18-24 y 25-34 años.
- Indique si es posible asegurar con 80% de confianza que la proporción de personas a favor de 18 a 24 es mayor que la proporción de personas entre 25 y 34 años.
  - ¿Será válida la afirmación del gobierno? Explique su respuesta.
- b) El límite inferior del I.C. de  $p_1 - p_2$  es  $-0.02$ , determine el límite superior y  $1 - \alpha$ .



20. Del ejercicio anterior calcule el tamaño mínimo de la muestra que debe seleccionar el gerente para que la diferencia de proporciones muestrales de las personas con edades entre 18 y 24 y entre 24 y 34 años se aleje de la verdadera diferencia de proporciones en menos de 0.04.
- a) Si considera las proporciones previas del ejercicio anterior.  
b) Para la cota inferior.
21. En una muestra de 250 personas con cierta enfermedad que utilizan medicamentos genéricos la proporción en la que resultó efectivo en los primeros 4 días es de 60%. Mientras que en otra muestra independiente de 320 personas con este padecimiento se les aplicó otro medicamento (no genérico) y resultó ser efectivo en 70% de los pacientes. Con 98% de confianza, ¿se podrá decir que los medicamentos genéricos son menos efectivos? Justifique su respuesta.
22. Se piensa que dos drogas son igual de efectivas para reducir el nivel de ansiedad en ciertas personas perturbadas emocionalmente. En una muestra de 160 personas con este padecimiento la proporción en que la droga A resultó ser efectiva es de 70%. Mientras que en una muestra de 180 personas con este padecimiento se les aplicó la droga B y resultó ser efectiva en 60% de los casos. Con 90% de confianza, ¿es válida la afirmación de que ambas drogas son efectivas? Justifique su respuesta.

## Ejercicios de repaso

### Preguntas de autoevaluación

- 3.1 ¿Una estadística podrá contener al parámetro?
- 3.2 ¿Será cierto que una estimación puntual siempre es mejor que una estimación por intervalos?
- 3.3 ¿Qué es el sesgo de un estimador?
- 3.4 ¿A partir de qué tamaño de la muestra se aconseja utilizar el TCL?
- 3.5 ¿Qué diferencia existe entre un estimador y un estadístico?
- 3.6 ¿Es posible aplicar a cualquier distribución las fórmulas que se determinaron para los intervalos de confianza de la media?
- 3.7 ¿En qué situación de los intervalos de confianza para la diferencia de medias se pide que las observaciones sean dependientes?
- 3.8 ¿Por qué es necesario que las muestras sean independientes en los intervalos de confianza de la razón de varianzas?
- 3.9 ¿Qué significa el nivel de confianza?
- 3.10 ¿Cuáles son las características deseables del intervalo de confianza de un parámetro?
- 3.11 ¿Qué significa que el parámetro  $\theta$  pertenezca al intervalo  $(a, b)$  con una confianza de 95%?

### Ejercicios complementarios con grado de dificultad uno

- 3.12 Sea  $X_1, X_2, \dots, X_n$  una muestra aleatoria de variables con media  $\mu$  y  $\sigma^2$ , muestre que los estimadores siguientes son insesgados de  $\mu$  e indique cuál es más eficiente.

$$a) T_1 = \frac{1}{10}(X_1 + 8X_2 + X_n)$$

$$b) T_2 = \frac{1}{20}(8X_2 + 6X_4 + 4X_6 + 2X_n)$$

$$c) T_3 = \frac{1}{n}(X_1 + (n-1)X_n)$$

- 3.13 Sea una población con función de densidad de Bernoulli con parámetro  $p$ :

$$f(x; p) = \begin{cases} 0, & \text{en caso de fracaso} \\ 1, & \text{en caso de éxito} \end{cases}$$

Por otro lado, sea  $X_1, X_2, \dots, X_5$  una muestra aleatoria para estimar al parámetro  $p$ , de la que se elige una realización dada por 1, 0, 0, 1 y 1. Utilice los valores de la realización para estimar al parámetro  $p$ . ¿Cree que sea confiable la estimación?

### Ejercicios complementarios con grado de dificultad dos

- 3.14 Sea  $X_1, X_2, \dots, X_n$  una muestra aleatoria de la distribución de Bernoulli con  $P(X = 1) = p = 1 - P(X = 0)$ , pruebe que  $T = S_{n-1}^2$  es un estimador insesgado  $p(1 - p)$ .
- 3.15 Sea  $X_1, X_2, \dots, X_{10}$  una muestra aleatoria de la densidad binomial con parámetros  $n = 10$  y  $p$ , pruebe que  $T = \frac{1}{10}\bar{X}$  es un estimador insesgado para  $p$ .
- 3.16 Sea  $X_1, X_2, \dots, X_n$  una muestra aleatoria de la densidad geométrica con parámetros  $p$  ( $P(X = x) = p(1 - p)^{x-1}$ ,  $x = 1, 2, 3, \dots$ ), pruebe que  $T = \bar{X}$  es un estimador insesgado para  $1/p$ .
- 3.17 Una muestra aleatoria de  $n = 16$  meses de los gastos de operación de una compañía tiene un promedio de \$5474 USD con una varianza de \$490 000 USD<sup>2</sup>, si la distribución de todos los gastos operacionales es normal. Con 95% de confianza:

- a) Construya un intervalo para la media de todos los gastos mensuales de la compañía, suponga que  $\sigma = \$750$  USD.
- b) Si el límite superior del intervalo de confianza es de \$6000 USD, ¿cuál es el límite inferior y el nivel de confianza?
- 3.18** Del ejercicio anterior con 95% de confianza:
- a) Determine el tamaño mínimo de la muestra que debe elegirse para que la estimación media esté dentro de un intervalo de longitud de \$500 USD.
- b) Construya un intervalo para la media de todos los gastos mensuales de la compañía, si no se conoce  $\sigma$ .
- 3.19** Del ejercicio anterior sobre gastos operacionales, con 95% de confianza:
- a) Construya un intervalo de confianza para la desviación estándar y decida si fue válida la suposición de que  $\sigma = 750$ .
- b) Si el límite superior del intervalo de confianza para  $\sigma$  es de \$1000 USD, ¿cuál es el límite inferior y el nivel de confianza?
- 3.20** Una muestra aleatoria de 41 cigarrillos de una marca determinada tiene un contenido promedio de nicotina de 1.3 mg y una desviación estándar de 0.17 mg. Si supone que las mediciones están normalmente distribuidas, con 98% de confianza:
- a) Construya un intervalo para la media del contenido de nicotina de los cigarrillos, suponga que  $\sigma = 0.15$  mg.
- b) Si el límite superior del intervalo de confianza es 1.34, ¿cuál es el límite inferior y el nivel de confianza?
- 3.21** Del ejercicio anterior con 98% de confianza:
- a) Determine el tamaño mínimo de la muestra que debe elegirse para que el error de la estimación de la media del contenido de nicotina sea menor a 0.05 mg.
- b) Construya un intervalo para la media del contenido de nicotina de los cigarrillos, si no se conoce  $\sigma$ .
- 3.22** Del ejercicio anterior sobre los cigarrillos, con 98% de confianza:
- a) Construya un intervalo para la desviación estándar y decida si fue válida la suposición de que  $\sigma = 0.15$ .
- b) Si el límite inferior del intervalo de confianza para  $\sigma$  es 0.15 mg, ¿cuál es el límite superior y el nivel de confianza?
- 3.23** Un fabricante de pilas asegura que duran en promedio 40 horas, con una desviación estándar de 1 hora. Si nueve de estas pilas tienen duraciones de 40.5, 38, 38.5, 41, 38.6, 40.5, 37.9, 39.1 y 39 horas, suponga que la población de la duración de las pilas se distribuye aproximadamente en forma normal. Con 95% de confianza:
- a) Construya un intervalo para la duración promedio de las pilas, suponga que  $\sigma = 1$  hora e indique si en estas condiciones es válida la suposición del fabricante de  $\mu = 40$  horas.
- b) Si el límite superior del intervalo de confianza es 40, ¿cuál es el límite inferior y el nivel de confianza?
- 3.24** Del ejercicio anterior con 95% de confianza:
- a) Determine el tamaño mínimo de la muestra que debe elegirse para que el error de la estimación de la media de la duración de las pilas sea menor a 0.5 horas.
- b) Construya un intervalo para la duración promedio de las pilas, si no se conoce  $\sigma$  e indique si en estas condiciones es válida la suposición del fabricante de que  $\mu = 40$  horas.
- 3.25** Del ejercicio anterior sobre las pilas con 95% de confianza:
- a) Construya un intervalo para la varianza y decida si fue válida la suposición de que  $\sigma = 1$  hora.
- b) Si el límite superior del intervalo de confianza para  $\sigma$  es 2 horas, ¿cuál es el límite inferior y el nivel de confianza?

Tabla 3.19

Fecha	BMB
09/06/2013	25.02
09/03/2013	24.84
09/02/2013	25.19
09/01/2013	24.80
08/31/2013	24.83
08/30/2013	24.60
08/27/2013	24.44
08/26/2013	24.30
08/25/2013	24.31
08/24/2013	24.54
08/23/2013	24.12
08/20/2013	24.09
08/19/2013	24.19
08/18/2013	23.85
08/17/2013	23.52
08/16/2013	23.51
08/13/2013	23.35

- 3.26** El IPC de la empresa BMB se muestra en la tabla 3.19 y se supone que tiene una distribución normal durante el año. Con 99% de confianza:
- a) Construya un intervalo para el IPC medio, suponga que  $\sigma = 1$ .
- b) Si el límite inferior del intervalo de confianza es 23.9, ¿cuál es el límite superior y el nivel de confianza?
- 3.27** Del ejercicio anterior con 99% de confianza:
- a) Determine el tamaño mínimo de la muestra que debe elegirse para que el error de la estimación de la media del IPC sea menor a 0.25.
- b) Construya un intervalo para el IPC medio, si no se conoce  $\sigma$ .
- 3.28** Del ejercicio anterior sobre la empresa BMB con 99% de confianza:
- a) Construya un intervalo para la varianza y decida si fue válida la suposición de que  $\sigma = 1$ .

b) Si el límite inferior del intervalo de confianza para  $\sigma$  es 0.40, ¿cuál es el límite superior y el nivel de confianza?

**3.29** Una realización de una muestra aleatoria de tamaño 10 de una mezcla de aluminios dio las siguientes proporciones en peso de un cierto tipo de polvo. El fabricante afirma que las proporciones en peso de la mezcla de aluminio tienen una  $\mu = 0.525$  y  $\sigma = 0.02$ :

0.50 0.54 0.49 0.53 0.52 0.56 0.48 0.50 0.52 0.51

Suponga una población normal, con 99% de confianza:

a) Construya un intervalo para la media de las proporciones de la mezcla, suponga que  $\sigma = 0.02$  e indique si en estas condiciones es válida la suposición del fabricante de  $\mu = 0.525$ .

b) Si el límite superior del intervalo de confianza es 0.525, ¿cuál es el límite inferior y el nivel de confianza?

**3.30** Del ejercicio anterior con 99% de confianza:

a) Determine el tamaño mínimo de la muestra que debe elegirse para que el error de la estimación de la media de las proporciones de la mezcla sea menor a 0.01.

b) Construya un intervalo para la media de las proporciones de la mezcla, si no se conoce  $\sigma$  e indique si en estas condiciones es válida la suposición del fabricante de  $\mu = 0.525$ .

**3.31** Del ejercicio anterior sobre las mezclas de aluminio con 99% de confianza:

a) Construya un intervalo para la desviación estándar y decida si fue válida la suposición de que  $\sigma = 0.02$ .

b) Si el límite inferior del intervalo de confianza para  $\sigma$  es 0.02, ¿cuál es el límite superior y el nivel de confianza para este intervalo?

**3.32** Los resultados del IPC de la empresa CM para una muestra se agruparon por medio de clases de frecuencia, resultando la distribución de frecuencias de la derecha. Si supone que el IPC de CM para este año tiene una distribución aproximadamente normal. Con 98% de confianza:

a) Construya un intervalo para el IPC medio, suponga que  $\sigma = 0.6$ .

b) Si el límite inferior del intervalo de confianza es 12.1, ¿cuál es el límite superior y el nivel de confianza?

**Tabla 3.20**

Intervalos de clase	Frecuencias
[11.19, 11.58]	4
(11.58, 11.97]	15
(11.97, 12.36]	20
(12.36, 12.75]	18
(12.75, 13.14]	9
(13.14, 13.53]	5

**3.33** Del ejercicio anterior con 98% de confianza:

a) Determine el tamaño mínimo de la muestra que debe elegirse para que el error de la estimación de la media del IPC sea menor a 0.2.

b) Construya un intervalo para el IPC medio, si no se conoce  $\sigma$ .

**3.34** Del ejercicio anterior sobre el IPC de la empresa CM con 98% de confianza:

a) Construya un intervalo para la desviación estándar y decida si fue válida la suposición de que  $\sigma = 0.6$ .

b) Si el límite inferior del intervalo de confianza para  $\sigma$  es 0.45, ¿cuál es el límite superior y el nivel de confianza?

**3.35** Se lleva a cabo un estudio para comparar las horas que tardan hombres y mujeres en armar un producto determinado. Las experiencias anteriores indican que la distribución de tiempos tanto para hombres como para mujeres es aproximadamente normal. Una muestra aleatoria de tiempos para 11 hombres y 14 mujeres arrojan los siguientes valores  $\bar{x}_h = 35$ ,  $s_h = 6.1$ ,  $\bar{x}_m = 40$ ,  $s_m = 5.3$ . Con 90% de confianza:

a) Construya un intervalo de confianza para la diferencia del rendimiento medio de armado, suponga que  $\sigma_h^2 = 35$  y  $\sigma_m^2 = 30$ .

b) Determine el tamaño mínimo de la muestra que debe elegirse para que el error de la estimación de la diferencia de medias de los tiempos de armado sea menor a 2.5 horas.

c) Si el límite inferior del I.C. vale  $-8$ , ¿cuánto corresponde al límite superior y cuál es  $1 - \alpha$ ?

**3.36** Del ejercicio anterior sobre los rendimientos medios de armado entre hombres y mujeres con 90% de confianza:

a) Construya un intervalo de confianza para la diferencia del rendimiento medio de armado, suponga que  $\sigma_h^2 = \sigma_m^2$  son desconocidas.

b) Si el límite inferior del I.C. vale  $-8$ , ¿cuánto corresponde al límite superior y cuál es  $1 - \alpha$ ?

**3.37** Del ejercicio anterior sobre los rendimientos medios de armado entre hombres y mujeres con 90% de confianza:

a) Construya un intervalo de confianza para la diferencia del rendimiento medio de armado, suponga  $\sigma_h^2 > \sigma_m^2$ , desconocidas.

b) Si el límite inferior del I.C. vale  $-8$ , ¿cuánto corresponde al límite superior y cuál es  $1 - \alpha$ ?

**3.38** Del ejercicio anterior sobre los rendimientos medios de armado entre hombres y mujeres con 90% de confianza:

a) Construya un intervalo de confianza para la razón entre desviaciones estándar de los tiempos de armado entre hombres y mujeres. ¿Qué suposición  $\sigma_h^2 = \sigma_m^2$  o  $\sigma_h^2 > \sigma_m^2$  debe considerarse válida?

b) Si el límite inferior del I.C. en a) vale 0.8, ¿cuánto corresponde el límite superior y cuál es  $1 - \alpha$ ?

*Sugerencia.* Efectúe los cálculos en Excel.

**3.39** Se quiere comparar la vida media de los cinescopios para receptores de televisión de dos fabricantes A y B; para esto se toma una muestra aleatoria de 26 cinescopios

prios del fabricante  $A$ , de los que se obtuvo una vida media de 6.5 años con una desviación estándar de 0.9 años; en tanto que la vida media de 20 cinescopios del fabricante  $B$  fue de 6.0 años con una desviación estándar de 0.8 años. Considere poblaciones normalmente distribuidas y con 98% de confianza.

- Construya un intervalo de confianza para la diferencia de la vida media de los cinescopios, suponga que  $\sigma_A^2 = 0.5$  y  $\sigma_B^2 = 0.4$ .
- Determine el tamaño mínimo de la muestra que debe elegirse para que el error de la estimación de la diferencia de las vidas medias de los cinescopios sea menor a 0.35 años.
- Si el límite superior del I.C. vale un año, ¿cuánto corresponde al límite inferior y cuál es  $1 - \alpha$ ?

**3.40** Del ejercicio anterior sobre la vida media de los cinescopios, con 98% de confianza:

- Construya un intervalo de confianza para la diferencia de la vida media de los cinescopios, suponga  $\sigma_A^2 = \sigma_B^2$  desconocidas.
- Si el límite superior del I.C. vale un año, ¿cuánto corresponde al límite inferior y cuál es  $1 - \alpha$ ?

**3.41** Del ejercicio anterior sobre la vida media de los cinescopios, con 98% de confianza:

- Construya un intervalo de confianza para la diferencia de la vida media de los cinescopios,  $\sigma_A^2 \neq \sigma_B^2$ , desconocidas.
- Si el límite superior del I.C. vale un año, ¿cuánto vale el límite inferior y cuál es  $1 - \alpha$ ?

**3.42** Del ejercicio anterior sobre la vida media de los cinescopios, con 98% de confianza:

- Construya un intervalo de confianza para la razón entre varianzas de los tiempos medios de vida de los cinescopios. ¿Qué suposición  $\sigma_A^2 = \sigma_B^2$  o  $\sigma_A^2 \neq \sigma_B^2$  debe considerarse válida?
- Si el límite superior del I.C. vale tres años, ¿cuánto vale el límite inferior y cuál es  $1 - \alpha$ ?

Sugerencia. Efectúe los cálculos de este inciso en Excel.

**3.43** En un proceso químico, se comparan dos catalizadores para verificar su efecto en el resultado de la reacción del proceso. Se preparó una muestra de 12 procesos utilizando el catalizador 1 y una de 10 con el 2. En el primer caso se obtuvo un rendimiento promedio de 84.5 con una desviación estándar muestral de 3, mientras que el promedio para la segunda muestra fue de 81 y la desviación estándar muestral de 5.5. Suponga que las poblaciones están distribuidas aproximadamente en forma normal. Construya un intervalo de 90% confianza para la razón de varianzas entre ambos catalizadores.

**3.44** Con base en el resultado del ejercicio anterior construya un intervalo con 90% de confianza para la diferencia promedio de los catalizadores en la reacción del proceso. ¿Se puede concluir con 90% de confianza que existe una diferencia real entre  $\mu_1$  y  $\mu_2$ ? ¿Los efectos entre ambos catalizadores son homogéneos?

**3.45** El IPC de las empresas Bimbo® y Grupo Modelo® se muestran en la tabla de abajo. Suponga que el IPC de las empresas tiene una distribución normal y que las muestras son independientes.

- Calcule un I.C. de 95% para la razón de varianzas entre ambos IPC.
- ¿Es posible afirmar con 95% de confianza qué empresa tiene mayor varianza del IPC? Justifique su respuesta.

**Tabla 3.21**

Fecha	BMB	GMD
09/06/2013	25.02	27.79
09/03/2013	24.84	27.73
09/02/2013	25.19	27.70
09/01/2013	24.80	27.06
08/31/2013	24.83	27.17
08/30/2013	24.60	27.06
08/27/2013	24.44	27.14
08/26/2013	24.30	27.32
08/25/2013	24.31	27.41
08/24/2013	24.54	27.66
08/23/2013	24.12	27.80
08/20/2013	24.09	27.78
08/19/2013	24.19	28.10
08/18/2013	23.85	28.25
08/17/2013	23.52	28.30
08/16/2013	23.51	28.00
08/13/2013	23.35	28.04

**3.46** Con base en el resultado del ejercicio anterior.

- Construya un intervalo con 95% de confianza para la diferencia promedio del IPC de las dos empresas. Es posible afirmar que el IPC de la empresa GMD es mayor en cuatro unidades al de BMB.
- ¿Se puede suponer que se trata de muestras pareadas? Explique su respuesta.

**3.47** En Guadalajara y Monterrey se llevó a cabo una investigación sobre el costo de la vida, para estimar el costo promedio en alimentación en familias de cuatro personas. De cada una de estas ciudades se seleccionaron muestras aleatorias independientes de 21 familias, los resultados obtenidos fueron:

$$\sum_{i=1}^{21} X_i = 139,150, \quad \sum_{i=1}^{21} X_i^2 = 1,103,192,500,$$

$$\sum_{i=1}^{21} Y_i = 139,720 \text{ y } \sum_{i=1}^{21} Y_i^2 = 1,114,254,400.$$

Si se supone que la distribución sobre el costo de vida en ambas ciudades es normal.

Construya un intervalo de confianza de 90% para estimar  $\sigma_1/\sigma_2$ . ¿Parece que la variabilidad en ambas ciudades es igual? Explique su respuesta.

- 3.48** Con base en el resultado del ejercicio anterior construya un intervalo con 90% de confianza para estimar  $\mu_1 - \mu_2$ . ¿Parece que el costo de vida promedio en ambas ciudades es igual? Explique su respuesta.
- 3.49** En una zona de la Ciudad de México se quiere estimar la proporción de residentes que están en contra de la construcción de la línea 3 del metrobús. Se toma una muestra aleatoria de 120 residentes y 75 manifestaron estar en contra. Con una confianza de 85% obtenga lo que se pide a continuación:
- Un I.C. para la verdadera proporción de residentes que están en contra de la construcción de la línea 3 del metrobús en su zona.
  - El límite inferior del intervalo de confianza de  $p$  es 0.50, determine el límite superior y  $1 - \alpha$ .
- 3.50** Del ejercicio anterior calcule el tamaño mínimo de la muestra que debe seleccionarse para que la proporción muestral de residentes que están en contra de la construcción de la línea 3 del metrobús se aleje de la verdadera proporción en menos de 5%.
- Si considera la proporción previa del ejercicio anterior.
  - Para la cota inferior.
- 3.51** El gobierno de la Ciudad de México indicó que durante el primer semestre de 2013 el porcentaje de ciudadanos que sufrió un robo fue de 7% y que su estudio es correcto con una probabilidad de 90%. Para probar de manera estadística esta información fue seleccionada una muestra aleatoria de 500 ciudadanos, de los cuales 48 sufrieron algún tipo de robo.
- ¿Es cierta la indicación del gobierno de la Ciudad de México? Justifique su respuesta.
  - El límite inferior del intervalo de confianza de  $p$  es 0.06. Determine el límite superior y  $1 - \alpha$ .
- 3.52** Del ejercicio anterior calcule el tamaño mínimo de la muestra que debe seleccionarse para que la proporción muestral se desvíe máximo 2% de la verdadera proporción en 90% de los casos.
- Considerando la proporción previa del ejercicio anterior.
  - Para la cota inferior.
- 3.53** ¿De qué tamaño debe seleccionarse una muestra si se desea estimar la proporción de habitantes de una comunidad que tiene una escolaridad básica de seis años, con un error de estimación máximo de 2% y si se usa un grado de confianza de 92%? Suponga que a partir de un estudio anterior se encontró que la proporción de habitantes con esta característica es de 31.25%.
- 3.54** Antes de aprobar los matrimonios entre personas del mismo género, el gobierno de la Ciudad de México aseguraba que la proporción de personas de 18 a 24 años de edad a favor de estos matrimonios era de 66.3%. Para probar esta afirmación fueron seleccionadas 500 personas de 18 a 24 años y 318 contestaron que estaban de acuerdo. Con una confianza de 95% obtenga:
- Un I.C. para la verdadera proporción de personas a favor de estos matrimonios, ¿se puede considerar válida la afirmación del gobierno de la Ciudad de México? Justifique su respuesta.
  - El límite inferior del intervalo de confianza de  $p$  es 0.60 Determine el límite superior y  $1 - \alpha$ .
- 3.55** Del ejercicio anterior calcule el tamaño mínimo de la muestra que debe seleccionarse para que la proporción muestral de personas a favor de estos matrimonios se aleje de la verdadera proporción en menos de 0.025.
- Considerando la proporción previa del ejercicio anterior.
  - Para la cota inferior.
- 3.56** Un fabricante de insecticidas en presentación de aerosol desea comparar dos nuevos productos. Se emplean en el experimento dos habitaciones del mismo tamaño, cada una con 1000 moscas. En uno se rocía el insecticida  $A$  y en el otro el insecticida  $B$  en igual cantidad, resultando que los insecticidas son efectivos en 85 y 76% de los casos, respectivamente. Con esta prueba y 99% de confianza, ¿qué puede determinar el fabricante con respecto a los insecticidas? Justifique su respuesta.
- 3.57** Suponga que en un estudio acerca del uso de internet se observa que en una muestra de 220 alumnos de escuelas particulares, 85% tiene que usar la red para sus trabajos al menos tres veces por semana, mientras que en una muestra de 280 alumnos de las escuelas públicas 40% usan la red con la misma periodicidad.
- Con esta información es posible afirmar con una confianza de 95% que la proporción de estos alumnos de escuelas particulares es mayor a la de alumnos de escuelas públicas en más de 50%. Explique su respuesta.
  - Resuelva el inciso a) con 75% de confianza.
  - ¿A partir de qué grado de confianza es verdadera la afirmación del inciso a)?
- 3.58** Del ejercicio anterior calcule el tamaño mínimo de la muestra que debe seleccionarse para que la diferencia de proporciones muestrales de los alumnos mencionados se aleje de la verdadera diferencia de proporciones en menos de 0.05. Con un nivel de confianza de 95%.
- Si considera las proporciones previas del ejercicio anterior.
  - Para la cota inferior.
- 3.59** Un fabricante de rodamientos realizó un muestreo de 640 artículos y encontró que 15% de los artículos producidos por la máquina  $A$  presentan un defecto menor. Mientras que en una muestra de 760 rodamientos por la máquina  $B$  solo 13% de los rodamientos también presentan este defecto.
- ¿Con esta información es válido suponer con una confianza de 95% que la proporción de artículos defectuosos de ambas máquinas son iguales? Explique su respuesta.



b) ¿Cuál es el nivel de confianza máximo que hace falsa la afirmación del inciso a)?

## Ejercicios complementarios con grado de dificultad tres

**3.60** Sea  $X_1, X_2, \dots, X_n$  una muestra aleatoria de variables con media  $\mu$  y varianza  $\sigma^2$ , muestre que  $T = \frac{2}{n(n+1)}(X_1 + 2X_2 + 3X_3 + \dots + nX_n)$  es un estimador insesgado de  $\mu$  y calcule su varianza.

**3.61** Suponga que de una población con varianza  $\sigma^2$  se seleccionan dos muestras aleatorias independientes de tamaños,  $n_1$  y  $n_2$ . Por otro lado, sean  $S_1^2$  y  $S_2^2$  dos estimadores insesgados de  $\sigma^2$ , demuestre que

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

es un estimador insesgado  $\mu\sigma^2$ . ¿Cambia la afirmación anterior si las muestras se consideran de dos poblaciones con una misma varianza  $\sigma^2$ ? ¿Cómo podría ser el estimador que se propone insesgado para  $\sigma^2$  en el caso de tres muestras aleatorias independientes?, y ¿en el caso de cuatro muestras?

**3.62** Suponga que se tiene una población con parámetro  $\theta$  y que  $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_n$  son estimadores insesgados de  $\theta$ .

a) Demuestre que cualquier media ponderada de los estimadores insesgados es otro estimador de este tipo.

b) ¿Ocurrirá lo mismo que en el inciso anterior para la media geométrica?, y ¿para la media armónica?

**3.63** La medición del radio de un círculo presenta errores aleatorios, cuya distribución es normal con media cero y varianza desconocida  $\sigma^2$ . Suponga que tenemos una muestra aleatoria de  $n$  medidas. Determine un estimador insesgado para el área del círculo.

**3.64** Un fabricante de dos aleaciones de magnesio quiere probar de manera estadística con una probabilidad de 0.95, si las durezas de ambas son iguales o cuál es más dura. Para esto toma dos muestras aleatorias de tamaño 10 cada una y mide los grados de dureza Brinell de las dos aleaciones.

Tabla 3.22

Aleación 1	66.3	63.5	64.9	61.8	64.3	64.7	65.1	64.5	68.4	63.2
Aleación 2	71.3	60.4	62.6	63.9	68.8	70.1	64.8	68.9	65.8	66.9

Por datos históricos se sabe que ambas aleaciones provienen de poblaciones aproximadamente normales.

a) ¿Qué le recomendaría hacer al fabricante?

b) ¿Será posible suponer que se trata de muestras pareadas? Justifique su respuesta.

c) Con base en las respuestas de los incisos a) y b) resuelva el problema.

Resuelva el problema anterior, pero ahora suponga que las dos aleaciones de magnesio tienen características muy similares y que las diferencias de cada pareja son independientes con distribución normal.

## Proyectos de la unidad 3

En la hoja "IPC-42 Emp" del archivo "Datos IPC Divisas.xlsx" que se encuentra en la página del libro en SALI, se encuentra una base de datos extraída de la página: <http://economia.terra.com.mx/mercados/acciones/cambios.aspx?idtel=IB032MEXBOL> de todos los IPC de 41 empresas que cotizan en México, a partir de febrero de 2013. Con esta información realice los siguientes proyectos:

- I. Con los valores del IPC de las empresas encuentre dos que tengan comportamientos normales y en cada uno construya intervalos con 95% de confianza para la media.
- II. Con las empresas encontradas en el ejercicio anterior, realice una comparación de medias por medio de intervalos de confianza.
- III. Con los valores del IPC de la empresa CC lleve a cabo una prueba de bondad de ajuste. Con todos los valores de la base de datos para esta empresa. Calcule un intervalo de confianza de 95%.

# Pruebas de hipótesis

UNIDAD

4



## Competencia específica a desarrollar

- Inferir el comportamiento de las características poblacionales, mediante la aplicación de los métodos de prueba de hipótesis, para el proceso de toma de decisiones.

## ¿Qué sabes?

- ¿Qué es una prueba de hipótesis?
- ¿Por qué es importante una prueba de hipótesis en la toma de decisiones?
- ¿Cuántos tipos de errores en una prueba de hipótesis conoces?

## Introducción

En la unidad anterior revisamos la estimación de parámetros por medio de estimadores puntuales e intervalos de confianza; ahora veremos otro tipo de teoría para aumentar el conocimiento sobre los parámetros de la distribución de la población en estudio, la cual consiste en llevar a cabo suposiciones o conjeturas sobre los parámetros que se analizan.

Suponga que tenemos una población en estudio de la que desconocemos los parámetros de su distribución y acerca de la que hacemos conjeturas, y quisiéramos saber si estas suposiciones son o no válidas. Por ejemplo, suponga que la población en estudio consiste en las calificaciones de los alumnos de ingeniería en la materia de mecánica. Además, conocemos la distribución de la población, pero desconocemos su parámetro media sobre el que formulamos la conjetura de que es mayor a 7.0. Para probar la verdad o falsedad de la conjetura elegimos la realización de una muestra aleatoria de las calificaciones de tamaño 15, con lo que se obtienen los resultados 8, 4, 6, 7, 6, 9, 5, 7, 8, 10, 7, 3, 9, 8, 4.

Se calcula la calificación promedio de estos datos para tener una estimación puntual del parámetro media, de lo que resulta  $\bar{x}_1 = 6.7$ . Entonces, ¿se podrá decir que la conjetura es falsa?

Suponga que elegimos otra realización del mismo tamaño de la muestra aleatoria, de la que se obtiene 6, 10, 7, 5, 7, 9, 10, 8, 7, 9, 7, 8, 4, 6, 8. Pero, en este caso la estimación puntual de la calificación promedio resulta  $\bar{x}_2 = 7.4$ ; por tanto, ¿se puede decir que la conjetura es verdadera?

Uno de los principales problemas al iniciar el estudio de las pruebas de hipótesis reside en comprender que el incumplimiento o cumplimiento de la conjetura por parte de los datos de la realización que se llevó a cabo no es suficiente para que estadísticamente digamos que es falsa o verdadera. Así, en la primera realización de una muestra aleatoria de tamaño 15 no se puede decir estadísticamente que la conjetura,  $\mu > 7$  es falsa debido a que  $\bar{x}_1 = 6.7$ . De igual forma, no podemos afirmar con la segunda realización que la conjetura  $\mu > 7$  sea verdadera.

En esta unidad estudiamos los conceptos básicos de la teoría de pruebas de hipótesis, y después tratamos la parte metodológica para las pruebas de hipótesis, que ayude en la toma de decisiones en problemas relacionados con poblaciones muy difíciles o imposibles de analizar en su totalidad. Por ejemplo, para concluir con cierta significancia sobre la vida promedio de las bombillas de luz de cierta marca, se puede formular una hipótesis, sobre la que deberá probarse su validez o falsedad. Es decir, buscaremos evidencias que ayuden a decidir si la debemos *rechazar* o *no*.

De forma similar a como se realizó con los intervalos de confianza, analizaremos la parte metodológica sobre poblaciones con distribución normal o aproximadamente normal y poblaciones con distribución tipo Bernoulli o binomial. Es decir, partiremos de una realización de la muestra aleatoria con base en las fórmulas que vamos a construir y obtendremos las reglas de decisión para los contrastes de hipótesis requeridos.

La parte metodológica para determinar reglas de decisión en los contrastes de hipótesis se aplicará a los parámetros que hasta ahora se estudiaron:

- Media y diferencia de medias de poblaciones aproximadamente normales.
- Varianza y razón entre varianzas de poblaciones aproximadamente normales.
- Proporciones y diferencia de proporciones de poblaciones con distribución de Bernoulli.
- Algunas otras pruebas para variables discretas con distribución: de Poisson y geométrica, entre otras.

### 4.1 Conceptos básicos sobre pruebas de hipótesis

En la unidad 3 estudiamos los intervalos de confianza para realizar estimaciones sobre los parámetros de la distribución de una población y poder llevar a cabo una mejor toma de decisiones al analizar la población de interés. En la presente unidad veremos otro método estadístico que nos ayude a la toma de decisiones en problemas relacionados con poblaciones que resultan muy difíciles o imposibles de analizar en su totalidad. Por ejemplo, para estimar con un valor de significancia, dado de antemano, la vida media de las bombillas de luz de cierta marca podemos formular una hipótesis, cuya validez deberá ser probada. Es decir, buscaremos evidencias que ayuden a decidir si la hipótesis se rechaza o no.



Llamaremos **hipótesis estadística** a cualquier afirmación o conjetura referente a los parámetros de una o más poblaciones.

Probar una *hipótesis estadística* consiste en buscar evidencias para decidir sobre el rechazo o no de la afirmación formulada. En el ejemplo de las bombillas de luz podemos conjeturar que su vida promedio está por arriba de las 750 horas, ahora suponga que se elige una muestra de estas bombillas y resulta que su vida promedio fue de 730 horas, la primera pregunta que surge en nuestra mente después de analizar la muestra es: ¿será evidencia suficiente el resultado de la realización para indicar que la conjetura realizada no es correcta?

En la prueba de hipótesis la verdad con respecto a la decisión tomada de rechazar o no la afirmación realizada solo se puede conocer al estudiar a toda la población. Por tanto, en las pruebas efectuadas debemos acostumbrarnos a comprender que *no rechazar* una afirmación basándonos en una realización indica que a partir de los datos obtenidos *no existen evidencias para hacerlo*.

Al formular una afirmación sobre un suceso y realizar una prueba de rechazo o no, es lógico preguntarse, ¿con base en qué se rechazará o no la afirmación realizada?

Note que al establecer una hipótesis siempre existirá, de forma implícita, otra que se le contrapone de manera que a las hipótesis formuladas se les da el nombre de **hipótesis nula** y **alterna**, que denotaremos por  $H_0$  y  $H_1$  o  $H_a$ , respectivamente. Así, inicia uno de los primeros problemas en el estudio de las pruebas de hipótesis, ¿cómo determinar la hipótesis nula y la alterna?

Para poder dar respuesta a esta pregunta necesitamos algunos conceptos más, que tratamos en las siguientes subsecciones.

## Regiones de rechazo y no rechazo

Suponga que nos encontramos ante el problema de la duración promedio de las bombillas de luz, en el que la población tiene un comportamiento que se puede describir por la función de densidad  $f(x; \mu)$  y el parámetro  $\mu$  tiene un espacio paramétrico  $\Omega = [0, \infty)$  ( $\mu$  tiempo de vida de las bombillas,  $\mu$  no puede ser negativo). Entonces, podemos establecer el siguiente contraste de hipótesis (en este momento aún no se explica cómo establecer las hipótesis nula y alterna):

$$H_0: \mu \leq 750$$

$$H_1: \mu > 750$$

Es decir, el espacio paramétrico  $\Omega = [0, \infty)$  se divide en dos regiones que denotamos con  $\omega$  para la región correspondiente al parámetro en la hipótesis nula y  $\Omega - \omega$  a la región correspondiente al parámetro en la hipótesis alterna. De esta forma, podemos establecer el contraste de hipótesis anterior en forma más general y equivalente a:

$$H_0: \mu \in \omega$$

$$H_1: \mu \in \Omega - \omega$$

Hasta el momento no hemos hablado acerca del problema que será de interés en forma práctica; es decir, qué hacer cuando solo se tengan datos para decidir cuál de las hipótesis es válida o qué entenderemos por una prueba de hipótesis.

Suponga que tenemos el problema de la prueba de hipótesis para la vida media de las bombillas y que cada una tiene un tiempo de vida descrito por una variable aleatoria con función de densidad  $f(x; \mu)$ . Por otro lado, tenemos una muestra aleatoria de estas variables  $X_1, X_2, \dots, X_n$  denotada por el vector  $\mathbf{X}$ . Como en todo problema concreto requeriremos trabajar no con la muestra aleatoria, sino con sus realizaciones, al representar  $R$  como el conjunto de todas las realizaciones de  $\mathbf{X}$ :

$$R = \{\mathbf{x} | \mathbf{x} \text{ es una realización de } \mathbf{X}\}$$

Así, que en un problema práctico podemos hacer una partición del conjunto  $R$  y tomar una decisión sobre la validez de la hipótesis nula basándonos en los resultados de las observaciones.

Se llama **prueba de hipótesis** para probar  $H_0$  contra  $H_a$  a una partición de  $R$  en dos conjuntos, que denotamos por  $R_a$  y  $R_r$ , y **región de no rechazo**, **región de rechazo** o **región crítica**, respectivamente.

Observe que  $\omega$  y  $\Omega - \omega$  no son iguales a  $R_a$  y  $R_r$ , porque los primeros dos conjuntos forman una partición del espacio de parámetros, mientras que los segundos una partición del conjunto de realizaciones. Entonces la regla de decisión se basa en estas últimas.

Note en la definición anterior que es posible establecer un procedimiento con el que sea factible decidir a partir de una realización de la muestra aleatoria si  $H_0$  es verdadera o no. En este sentido, ¿cómo determinar si una hipótesis nula debe ser rechazada o no?

Para el contraste de hipótesis de un parámetro  $\theta$ , en general:

$$H_0: \theta \in \omega$$

$$H_1: \theta \in \Omega - \omega$$

La regla de decisión estará dada con base en la realización  $x$  como:

- Rechazar  $H_0$  si  $\mathbf{x} \in R_r$ ,  $R_r$ , región de rechazo o región crítica.
- No rechazar  $H_0$  si  $\mathbf{x} \in R_a$ ,  $R_a$ , región de no rechazo.

Por ejemplo, para el caso particular del tiempo de vida promedio de las bombillas, tenemos:

$$H_0: \mu \leq 750$$

$$H_1: \mu > 750$$

Luego,  $R = \{\mathbf{x} | \mathbf{x} \text{ es una realización } \mathbf{x}\}$ , y podemos elegir un valor  $\mu_0 \in \Omega$ , no solo  $\mu_0 = 750$ , de tal forma que la partición de  $R$  estará dada por:

$$R_a = \{\mathbf{x} | \mathbf{x} \in R \text{ y } \bar{x} \leq \mu_0\}$$

$$R_r = \{\mathbf{x} | \mathbf{x} \in R \text{ y } \bar{x} > \mu_0\}$$

Observe que para cada valor elegido de  $\mu_0 \in \Omega$  tenemos una prueba de hipótesis o partición de  $R$ . Al valor  $\mu_0$  que divide las regiones de rechazo y no rechazo se llama **valor crítico**. Por ejemplo, podemos considerar el valor crítico  $\mu_0 = 760$  horas, con lo cual quedan establecidas las regiones de no rechazo (para  $\bar{x} \leq 760$ ) y rechazo o región crítica (para  $\bar{x} > 760$ ). Lo más probable es que nos preguntemos, ¿cómo se determinaron las regiones? En este caso, la cantidad de 760 se eligió como un ejemplo ilustrativo de qué son las regiones de rechazo y no rechazo. Así, vimos que podemos tener una infinidad de pruebas y desde luego, surgen las siguientes preguntas: ¿cuál de todas las pruebas sería buena?, ¿se podrá establecer una prueba que sea la mejor?

## Tipos de errores en una prueba de hipótesis

En la subsección anterior tratamos la definición de prueba de hipótesis, como una partición del conjunto de realizaciones. Después, con base en una regla de decisiones, se decide rechazar o no la hipótesis nula, pero sabemos que en estadística los resultados encontrados no son 100% confiables puesto que siempre dependen de las condiciones aleatorias de la variabilidad del fenómeno en estudio. Por este motivo, al tomar una decisión con respecto a la validez de la hipótesis nula estamos propensos a cometer uno de los dos, errores que trataremos sean lo más pequeños posibles.

Llamamos **error tipo I** cuando se rechaza la hipótesis nula, aunque en realidad es verdadera y **error tipo II** cuando no se rechaza la hipótesis nula, aunque en realidad es falsa.

Dadas las definiciones de los dos errores más importantes al rechazar o no una hipótesis nula, surge la pregunta: ¿cuál es la probabilidad de cometer los errores tipo I o II?

Con la definición anterior podemos dar respuesta a la pregunta sobre una buena prueba, ya que será razonable considerar que es aquella que minimiza las probabilidades de ambos tipos de errores. De igual forma, la mejor prueba, si es que existe, será la que minimiza las probabilidades de ambos errores con respecto a todas las otras pruebas posibles.

Del párrafo anterior parece que encontrar una buena prueba es tarea simple; sin embargo, de manera general cuando se minimiza la probabilidad de uno de los errores, la probabilidad del otro tipo de error aumenta. De hecho, dar respuesta a las preguntas anteriores es uno de los principales problemas que existen en la teoría de prueba de hipótesis y requiere de cierto tiempo y mayor comprensión del problema. Para esto iniciaremos con la introducción de la siguiente notación:

$$P(\text{Error tipo I usando } R_r | H_0) = \text{Probabilidad de cometer el error tipo I, con } H_0 \text{ verdadera}$$

Del mismo modo, tenemos:

$$P(\text{Error tipo II usando } R_a | H_1) = \text{Probabilidad de cometer el error tipo II, con } H_1 \text{ verdadera}$$

Note que el cálculo de probabilidades para el error tipo II también se puede llevar a la región de rechazo por medio del complemento.

$$P(\text{Error tipo II usando } R_a | H_1) = 1 - P(\text{Error tipo II usando } R_r | H_1)$$

En los ejemplos 4.1, 4.2 y 4.3 mostraremos los cálculos para conocer los errores tipo I y II.

#### Ejemplo 4.1 Errores tipos I y II

Suponga que en el caso del tiempo de vida de las bombillas de luz, éstas tienen una desviación estándar de vida igual a 50 horas, si se considera una muestra de 49 bombillas y las hipótesis:

$$H_0: \mu \leq 750$$

$$H_1: \mu > 750$$

con la región de rechazo establecida para promedios de vida mayores a 760 horas, calcule:

- La probabilidad de cometer un error tipo I para el caso en que  $\mu = 740$ .
- La probabilidad de cometer el error tipo II para el caso en que  $\mu = 755$ .

#### Solución

- Para calcular las probabilidades tenemos que la región de rechazo está dada por  $\bar{X} > 760$ ; por otro lado, el tamaño de la muestra es  $n = 49$ , entonces podemos utilizar el teorema central del límite:

$$\begin{aligned} P(\text{error tipo I al usar } R_r | H_0) &= P(\bar{X} > 760 | \mu \leq 750) = P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > \frac{760 - \mu}{50/\sqrt{49}} \mid \mu = 740\right) \\ &= P\left(Z > \frac{760 - 740}{50/\sqrt{49}}\right) = P(Z > 2.8) = 0.0026 \end{aligned}$$

- De la misma forma para calcular la probabilidad del error tipo II hacemos uso del teorema central del límite y el hecho de que  $R_a$  está dada por  $\bar{X} \leq 760$ .

$$\begin{aligned} P(\text{II al usar } R_a | H_1) &= P(\bar{X} \leq 760 | \mu > 750) = P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq \frac{760 - \mu}{50/\sqrt{49}} \mid \mu = 755\right) = P\left(Z \leq \frac{760 - 755}{50/\sqrt{49}}\right) \\ &= P(Z \leq 0.70) = 0.7580 \end{aligned}$$

También se puede calcular por medio del complemento:

Antes de resolver el problema, observe que en el cálculo del error tipo I, la hipótesis nula tiene que ser verdadera,  $\mu \leq 750$ . Es decir, para calcular la probabilidad del error tipo I tenemos una infinidad de valores del parámetro media, por tal razón elegimos uno en particular:  $\mu = 740$ . De manera similar para calcular la probabilidad del error tipo II elegimos el valor  $\mu = 755$ , que cumple con la condición  $\mu > 750$ , cuando  $H_1$  es verdadera.

$$P(\text{error tipo II si se usa } R_a | H_1) = 1 - P(\text{error tipo II al usar } R_r | H_1)$$

Luego:

$$\begin{aligned} P(\text{error tipo II al usar } R_a | H_1) &= 1 - P(\text{error tipo II al usar } R_r | H_1) = 1 - P(\bar{X} > 760 | \mu > 750) \\ &= 1 - P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > \frac{760 - \mu}{50/\sqrt{49}} \mid \mu = 755\right) = 1 - P\left(Z > \frac{760 - 755}{50/\sqrt{49}}\right) = 0.7580 \end{aligned}$$

El valor del error tipo II es demasiado grande, en consecuencia lo más probable es que la prueba utilizada, o partición, del conjunto de realizaciones no sea la más adecuada.

#### Ejemplo 4.2 Errores tipos I y II

Suponga que en el ejemplo anterior se considera la región de rechazo para promedios de vida mayores a 752 horas. Calcule:

- La probabilidad de cometer un error tipo I para el caso en que  $\mu = 740$ .
- La probabilidad de cometer el error tipo II para el caso en que  $\mu = 755$ .

#### Solución

- Si se continúa con la misma metodología de cálculos de probabilidades, se tendrá:

$$\begin{aligned} P(\text{error tipo I al usar } R_r | H_0) &= P(\bar{X} > 752 | \mu \leq 750) = P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > \frac{752 - \mu}{50/\sqrt{49}} \mid \mu = 740\right) \\ &= P\left(Z > \frac{752 - 740}{50/\sqrt{49}}\right) = P(Z > 1.68) = 0.0465 \end{aligned}$$

- De manera similar, para calcular la probabilidad del error tipo II con  $R_a$  dada por  $\bar{X} > 752$ .

$$\begin{aligned} P(\text{II si se usa } R_a | H_1) &= P(\bar{X} \leq 752 | \mu > 750) = P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq \frac{752 - \mu}{50/\sqrt{49}} \mid \mu = 755\right) = P\left(Z \leq \frac{752 - 755}{50/\sqrt{49}}\right) \\ &= P(Z \leq -0.42) = 0.3372 \end{aligned}$$

Al comparar las dos pruebas anteriores,  $\bar{X} > 760$  con  $\bar{X} > 752$ , concluimos que es mejor la prueba para la partición  $\bar{X} > 752$  y  $\bar{X} \leq 752$  de  $R$ . Puesto que la probabilidad del error tipo I es pequeña (alrededor de 5%), mientras que la probabilidad del error tipo II disminuyó de manera considerable, comparada con la correspondiente de la partición  $\bar{X} > 760$  y  $\bar{X} \leq 760$  (véase figura 4.1).

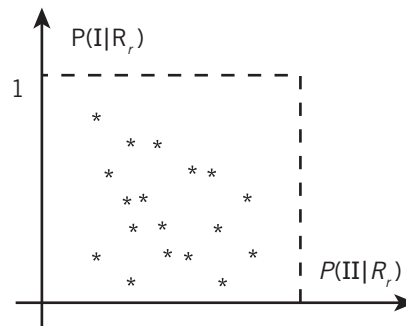


Figura 4.1 Valores de las probabilidades de los errores tipo I y II.

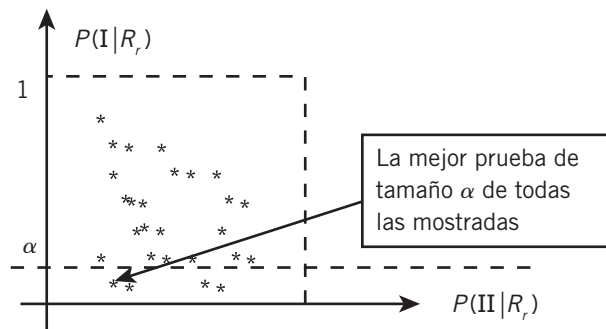
Del ejemplo anterior y las definiciones de pruebas de hipótesis y la partición del conjunto  $R$ , tenemos que el problema de encontrar una buena prueba se puede trabajar como la determinación de una buena partición del conjunto  $R$ , incluso es posible notar que una prueba queda especificada de manera total con la sola definición de la región de  $R_r$ . Así a cada prueba le corresponde una pareja de probabilidades  $P(I|R_r)$  y  $P(II|R_r)$ , las cuales se grafican en el plano cartesiano (véase figura 4.1).

Luego, la *mejor prueba* es aquella cuya región de rechazo  $R_r^*$  es tal que la pareja de probabilidades  $(P(I|R_r^*), P(II|R_r^*))$  se encuentra más próxima al origen de coordenadas.

Por último, llegamos al momento de definir algunos de los conceptos de mayor uso e importancia en las pruebas de hipótesis.

Se conocerá como **prueba de  $\alpha$** , a una prueba  $R_r$  que satisface  $P(I|R_r) \leq \alpha$  para algún valor  $\alpha \in (0, 1)$ . Si además la prueba de tamaño  $\alpha$  tiene la mínima probabilidad del error tipo II, se llamará entonces **prueba más potente**.

En forma gráfica, la definición anterior representa lo que se observa en la figura 4.2. Sea  $\alpha \in (0, 1)$  y tracemos una línea horizontal, que deje por debajo a todas las pruebas de tamaño  $\alpha$ , elegimos la que tenga menor valor de la probabilidad del error tipo II  $P(II|R_r)$ .



**Figura 4.2** Valores de las probabilidades de los errores I y II, y tamaño de la prueba.

Podemos observar que si tenemos una prueba de tamaño  $\alpha$  y  $\alpha^* \in (0, 1)$ , tal que  $\alpha < \alpha^*$ , entonces la prueba también es de tamaño  $\alpha^*$ . Esto se deduce de forma inmediata de:

$$P(I|R_r) \leq \alpha < \alpha^*$$

### Ejemplo 4.3 Errores tipos I y II

Determine el tipo de prueba que se trata con respecto al tamaño  $\alpha$  resultante en cada uno de los dos ejemplos anteriores.

#### Solución

En el ejemplo 4.1 se encontró para la partición  $\bar{X} > 760$  y  $\bar{X} \leq 760$ , que la probabilidad del error tipo I fue 0.0026. Luego, podemos considerar que se trata de una prueba de tamaño  $\alpha$ ,  $\alpha \in (0.0026, 1)$ .

En el ejemplo 4.2 se encontró para la partición  $\bar{X} > 752$  y  $\bar{X} \leq 752$ , que la probabilidad del error tipo I fue 0.0465. Luego, podemos considerar que se trata de una prueba de tamaño  $\alpha$ ,  $\alpha \in (0.0465, 1)$ .

De los ejemplos anteriores observamos que cada cálculo de probabilidades se basa en la partición del conjunto de realizaciones, la cual queda determinada por una acotación con respecto a una estadística. En los ejemplos anteriores se usó la estadística  $\bar{X}$ .

Sea el contraste de hipótesis para el parámetro  $\theta$ :

$$\begin{aligned} H_0: \theta &\in \omega \\ H_1: \theta &\in \Omega - \omega \end{aligned}$$

La estadística que se usa para determinar la región de rechazo se denomina estadística de prueba y al valor que acota la región de rechazo le denominamos valor crítico.

En los ejemplos anteriores, la estadística de prueba fue  $\bar{X}$  (debido a que el parámetro de estudio era  $\mu$ ), mientras que en el ejemplo 4.1 el valor crítico utilizado fue de 760 y en el 4.2 de 752. Después, veremos que uno de los principales problemas de las pruebas de hipótesis reside en determinar la estadística de prueba y sus valores críticos.

De lo anterior concluimos:

- La **prueba es la regla de decisión**; rechazar o no la hipótesis nula.
- Al definir la partición, también lo hacemos con la regla de decisión y requerimos, tanto de la estadística de prueba como del valor crítico.

## Función de potencia y tamaño de la prueba

En un principio se definieron las pruebas de hipótesis con regiones de rechazo y no rechazo, esto dio una base para comprender mejor cómo establecer los contrastes de hipótesis.

Se llama **función de potencia de la prueba** a  $\beta(\theta): \Omega \rightarrow [0, 1]$  cuando:

$$\beta(\theta) = P(\text{rechazar } H_0 | \theta)$$

Observe que en la definición de la función de potencia de la prueba, ésta dependerá de la región en donde se encuentre el parámetro de estudio  $\theta$ . Por ejemplo, si  $\theta \in \omega$  la función de potencia de la prueba define a la probabilidad del error tipo I, puesto que:

$$\beta(\theta) = P(\text{rechazar } H_0 | \theta \in \omega) = P(\text{rechazar } H_0 | H_0)$$

Una prueba de tamaño  $\alpha$  es equivalente al valor  $\alpha \in (0, 1)$  que cumple  $\sup_{\theta \in \omega} \beta(\theta) \leq \alpha$ , cuando se cumple la igualdad, entonces  $\alpha$  se llama **nivel de significancia**.

De igual manera, cuando se definió la prueba de tamaño  $\alpha$  para una partición, si la prueba  $\phi$  es de tamaño  $\alpha$ , entonces también será de tamaño  $\alpha^*$  para toda  $\alpha^* \geq \alpha$ .

Por otro lado, si  $\theta \in \Omega - \omega$ , entonces:

$$\beta(\theta) = P(\text{rechazar } H_0 | \theta \in \Omega - \omega) = 1 - P(\text{no rechazar } H | \theta \in \Omega - \omega) = 1 - P(\text{error tipo II} | H_1)$$

En general, a la probabilidad del error tipo II se le suele denotar por  $\beta$ . Es decir,

$$\beta = P(\text{error tipo II} | H_1)$$

Se llama **potencia de la prueba** a  $1 - \beta$ , en donde  $\beta = P(\text{error tipo II} | H_1)$ . Es decir:

$$\text{potencia de la prueba} = 1 - \beta = \beta(\theta) \text{ para } \theta \in \Omega - \omega$$

De la definición anterior, podemos notar que la potencia de la prueba es buena cuando la probabilidad del error tipo II es pequeña.

Note que la potencia de la prueba coincide con el valor de la función de potencia cuando el parámetro  $\theta \in \Omega - \omega$ . Entonces, ¿qué representa la potencia de la prueba?

La potencia de la prueba cuantifica la probabilidad de rechazar la hipótesis nula cuando ésta es falsa. Es decir, rechazar la hipótesis nula de manera acertada.

Del comentario anterior vemos que al llevar a cabo una prueba de hipótesis será recomendable trabajar, no solo con su tamaño, sino también con su potencia.

#### Ejemplo 4.4 Potencia de la prueba

Suponga que en el caso de la vida promedio de las bombillas de luz, éstas tienen una distribución normal  $N(\mu, 50^2)$ , considere una muestra de 49 bombillas y las hipótesis:

$$H_0: \mu \leq 750$$

$$H_1: \mu > 750$$

con la región de rechazo establecida para medias mayores a 760. Es decir, la estadística de prueba  $T(\mathbf{X}) = \bar{X}$  y valor crítico  $a = 760$ , con región de rechazo  $\bar{x} > 760$ . Determine los siguientes aspectos:

- La función de potencia de la prueba y trace su gráfica.
- Nivel de significancia.
- Una expresión para la probabilidad del error tipo II.
- Calcule la potencia de la prueba cuando  $\mu = 755 \in \Omega - \omega$  e interprete el resultado.
- Calcule la potencia de la prueba cuando  $\mu = 770 \in \Omega - \omega$  e interprete el resultado.

#### Solución

- Calculemos la función de potencia, para lo cual necesitamos la región crítica  $R_r = \{\mathbf{x} | \mathbf{x} \in \mathbf{R} \text{ y } \bar{x} > 760\}$ , en la que la estadística de prueba es  $T(\mathbf{X}) = \bar{X}$ , cuyo resultado es:

$$\beta(\mu) = P(\text{rechazar } H_0 \text{ usando } \phi | \mu) = P(\bar{X} > 760 | \mu) = P\left(Z > \frac{760 - \mu}{50/\sqrt{49}} | \mu\right) = 1 - \Phi\left(\left(\frac{760 - \mu}{50}\right)7\right)$$

$$\text{Es decir, la función de potencia es } \beta(\mu) = 1 - \Phi\left(\left(\frac{760 - \mu}{50}\right)7\right) = 1 - \int_{-\infty}^{\left(\frac{760 - \mu}{50}\right)7} \frac{e^{-z^2/2}}{\sqrt{2\pi}} dz$$

La gráfica de la función de potencia se obtiene asignando valores a  $\mu$ ,  $0 \leq \mu < \infty$  (véase tabla 4.1 y figura 4.3).

**Tabla 4.1** Valores de la potencia del ejemplo 4.4.

$\mu$	Potencia	$\mu$	Potencia	$\mu$	Potencia
740.0	0.00256	760.0	0.50000	780.0	0.99744
744.0	0.01255	764.0	0.71226	784.0	0.99961
748.0	0.04648	768.0	0.86864	790.0	0.99999
752.0	0.13136	772.0	0.95352		
756.0	0.28774	776.0	0.98745		



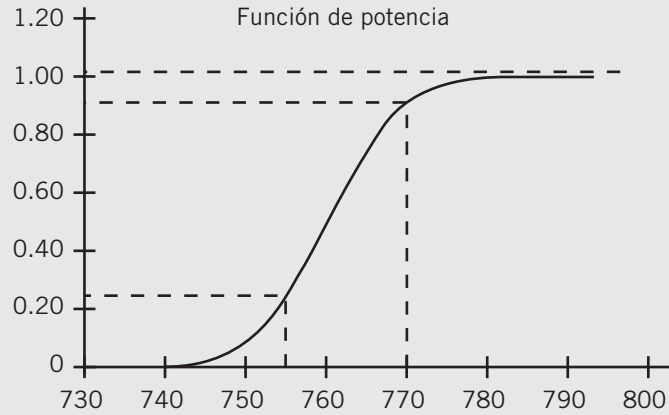


Figura 4.3 Función de potencia del ejemplo 4.4.

Con ayuda de la gráfica de la figura 4.3 aproximamos las potencias solicitadas en los incisos *d*) y *e*).

*b*) El nivel de significancia:

$$\alpha = \sup_{\mu \in \omega} \beta(\mu) = \sup_{\mu \leq 750} \left\{ 1 - \Phi \left( \left( \frac{760 - \mu}{50} \right) 7 \right) \right\} = 1 - \Phi \left( \left( \frac{760 - 750}{50} \right) 7 \right) = 0.081$$

Note que el supremo se obtiene en el valor máximo de  $\mu$ , porque  $\Phi$  es creciente,  $-\Phi$  será decreciente, entonces obtiene su máximo cuando el argumento tiene el valor más pequeño, pero está evaluada en  $-\mu$ . El valor más pequeño de  $-\mu$  es cuando  $\mu$  es más grande, como  $\mu \leq 750$ . Entonces la función de potencia se evalúa en 750.

*c*) La probabilidad del error tipo II.

Sabemos que la función de potencia,  $\beta(\mu) = 1 - \beta$  para  $\mu > 750$ , despejamos a  $\beta$ .

$$\beta = 1 - \beta(\mu) = 1 - \left[ 1 - \Phi \left( \left( \frac{760 - \mu}{50} \right) 7 \right) \right] = \Phi \left( \left( \frac{760 - \mu}{50} \right) 7 \right) \text{ para } \mu > 750$$

*d*) La potencia de la prueba cuando  $\mu = 755 \in \Omega - \omega$ , será:

$$1 - \beta = \beta(755) = 1 - \Phi \left( \left( \frac{760 - 755}{50} \right) 7 \right) = 1 - \Phi(0.7) = 0.2420$$

Si la región de rechazo se establece para valores mayores a 760 horas y la verdadera vida promedio de las bombillas es de 755, entonces la potencia de la prueba tendrá que ser baja, debido a que habrá muchas realizaciones con un promedio por debajo de 760 horas, pero mayor a 750 puesto que si la vida promedio de la población fuera de 755, se esperaría que el promedio de vida de las realizaciones estuviera alrededor de este valor, luego habrá muchas realizaciones con promedio de vida por debajo del valor crítico (véase figura 4.4). Por estas razones, no será rechazada la hipótesis nula, aunque tendría que serlo. La situación que ocurre se muestra de manera gráfica en la figura 4.4.

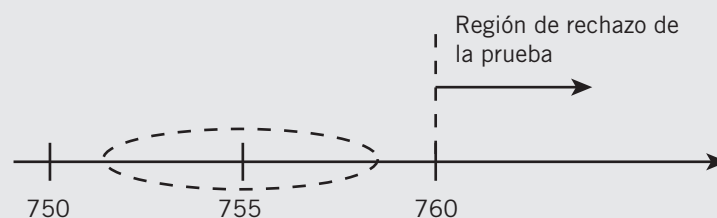


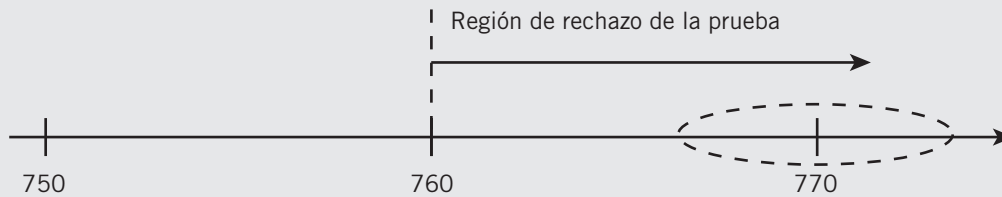
Figura 4.4 Regiones en donde se rechaza la prueba con parámetro 755 y valor crítico 760.

Se espera que la mayoría de los promedios muestrales estén alrededor de  $\mu = 755$ . En estas situaciones se dice que la prueba elegida no fue la más adecuada.

e) La potencia de la prueba cuando  $\mu = 770 \in \Omega - \omega$  es:

$$1 - \beta = \beta(770) = 1 - \Phi\left(\left(\frac{760 - 770}{50}\right)7\right) = 1 - \Phi(-1.40) = 0.9192$$

La interpretación es similar a la del inciso anterior, pero si se considera que la verdadera vida promedio de las bombillas de luz es de 770 horas, quiere decir que al tomar realizaciones de la muestra su vida promedio estará en las proximidades de 770 y en la mayoría de los casos se tendrá que rechazar la hipótesis nula al calcular la potencia. Por consiguiente, la potencia de la prueba será alta (véase figura 4.5). Lo anterior resulta debido a que en la mayoría de las realizaciones que se calcule la vida promedio de las bombillas de luz resulta por arriba de 760 horas (valor crítico).



**Figura 4.5** Regiones en donde se rechaza la prueba con parámetro 770 y valor crítico 760.

En estas situaciones se dice que la prueba elegida fue buena para un nivel de significancia de 8.1 por ciento.

## Elección de la hipótesis nula y alterna

En los textos metodológicos sobre pruebas de hipótesis, la formulación de las hipótesis nula y alterna es un poco confusa, porque se establecen reglas que no consideran la esencia de la teoría sobre la que están cimentadas. En este sentido, ¿cómo establecer la hipótesis nula y alterna en un problema particular?

Para dar respuesta a la elección de la hipótesis nula y alterna tenemos que tener en cuenta los requisitos que debe cumplir la mejor prueba:

1. La prueba se encuentra a partir de la filosofía de rechazar la hipótesis nula (véase tamaño de la prueba). Puesto que se busca la probabilidad de cometer el error tipo I, rechazar la hipótesis nula aunque es verdadera.
2. Para determinar el nivel de significancia de la prueba lo definimos con base en el supremo,  $\alpha = \sup_{\theta \in \omega} \beta(\theta)$ , aunque en la literatura algunos autores emplean el máximo en lugar del supremo, en cuyos casos es importante establecer la parte del espacio paramétrico bajo la hipótesis nula,  $\omega$ , puesto que si el conjunto es abierto y la función de potencia  $\beta(\theta)$  es monótona, entonces no tendrá máximo, ya que éste se encuentra en la frontera del conjunto, situación que no ocurre con el supremo.
3. Después de dichas pruebas se elige la que tenga mayor potencia.

Así, de (1), el investigador trabaja con la filosofía de rechazar  $H_0$ , por consiguiente, en la hipótesis alterna plantea su conjetura que quiere probar y supone verdadera. Por otro lado, de (2) debemos tener cuidado de establecer un conjunto cerrado en  $\omega$ , razón que da origen a proponer la hipótesis nula con la aseveración del parámetro con el signo de relación  $\leq, \geq$  o  $=$ .

## Ejemplos 4.5 Hipótesis nula y alterna

1. Suponga que un fabricante de bombillas de luz asegura que su producto tiene una duración promedio mayor a 810 horas. Plantee un juego de hipótesis para probar la afirmación del fabricante.	$H_0: \mu \leq 810$ $H_1: \mu > 810$
2. El gerente de mercadotecnia de una empresa asegura que los costos de publicidad promedio semanal para el siguiente año que ayuden a mantener ventas por arriba de las del actual serán menores que \$250 000. Plantee un juego de hipótesis para probar la afirmación del fabricante.	$H_0: \mu \geq 250\,000$ $H_1: \mu < 250\,000$
3. Suponga que un fabricante de lavadoras asegura que su producto tiene un promedio de fallas que está fuera del intervalo [4, 10] años. Plantee un juego de hipótesis para probar la afirmación del fabricante.	$H_0: 4 \leq \mu \leq 10$ $H_1: \mu < 4 \text{ o } \mu > 10$
4. Suponga que un fabricante de lavadoras asegura que la proporción de lavadoras que tienen fallas en menos de 4 años es menor a 3%. Plantee un juego de hipótesis para probar la afirmación del fabricante.	$H_0: p \geq 0.03$ $H_1: p < 0.03$
5. Sea una variable aleatoria con distribución binomial $X \sim \text{Bin}(2, \theta)$ con espacio $\Omega = \{0.25, 0.60\}$ y el investigador asegura que $\theta = 0.60$ . Plantee el juego de hipótesis del problema. Note que el parámetro solo puede tomar dos valores.	$H_0: \theta = 0.25$ $H_1: \theta = 0.60$
6. Sea una variable aleatoria con distribución binomial $X \sim \text{Bin}(2, \theta)$ con espacio $\Omega = [0, 1]$ el investigador asegura que $\theta = 0.60$ . Plantee el juego de hipótesis del problema. En estas situaciones la afirmación se pone en la hipótesis nula, para conservar la relación de igualdad en la hipótesis nula.	$H_0: \theta = 0.60$ $H_1: \theta \neq 0.60$

## Cálculo de las probabilidades para los dos tipos de errores

Los cálculos con respecto a las probabilidades de los dos tipos de errores que podemos cometer al probar una hipótesis se realizan con las siguientes expresiones:

$$P(\text{error tipo I al usar } R_r | H_0) = P(\text{rechazar } H_0 | H_0 \text{ es válida})$$

$$P(\text{error tipo II al usar } R_a | H_1) = P(\text{no rechazar } H_0 | H_1 \text{ es válida})$$

$$\text{Potencia de la prueba} = P(\text{rechazar } H_0 | H_1 \text{ es válida})$$

Los siguientes tres ejemplos muestran diferentes distribuciones para el cálculo de las probabilidades de los errores tipo I y II.

## Ejemplo 4.6 Cálculo de probabilidades

Suponga que en el caso del tiempo de vida de las bombillas de luz, éstas tienen una desviación estándar de vida igual a 55 horas, considere una muestra de 60 bombillas y las hipótesis:

$$H_0: \mu \leq 810$$

$$H_1: \mu > 810$$

con la región de rechazo establecida para promedios de vida mayores a 820 horas, calcule:

- La probabilidad de cometer un error tipo I para el caso en que  $\mu = 800$ .
- La probabilidad de cometer el error tipo II para el caso en que  $\mu = 815$ .

**Solución**

- Para calcular las probabilidades, tenemos que la región de rechazo está dada por  $\bar{X} > 820$  por otro lado, el tamaño de la muestra es  $n = 60$ . Luego utilizamos el teorema central del límite:

$$\begin{aligned}
 P(\text{error tipo I al usar } R_r | H_0) &= P(\bar{X} > 820 | \mu \leq 810) = P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > \frac{820 - \mu}{55/\sqrt{60}} \mid \mu = 800\right) \\
 &= P\left(Z > \frac{820 - 800}{55/\sqrt{60}}\right) = P(Z > 2.82) \approx \Phi(-2.82) = 0.0024
 \end{aligned}$$

b) De igual forma, para calcular la probabilidad del error tipo II, hacemos uso del teorema central del límite y el hecho de que  $R_a$  está dada por  $\bar{X} \leq 820$ :

$$P(\text{II al usar } R_a | H_1) = P(\bar{X} \leq 820 | \mu > 810) = P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq \frac{820 - \mu}{55/\sqrt{60}} \mid \mu = 815\right) = P\left(Z \leq \frac{820 - 815}{55/\sqrt{60}}\right)$$

$$P(\text{II al usar } R_a | H_1) = P(Z \leq 0.70) \approx \Phi(0.70) = 0.7580$$

El valor del error tipo II es demasiado grande. En consecuencia, lo más probable es que la prueba utilizada, o lo que es lo mismo, la partición del conjunto de realizaciones, no lo sea.

#### Ejemplo 4.7 Cálculo de probabilidades

Suponga que en el ejemplo anterior se considera la región de rechazo para promedios de vida mayores a 812 horas. Calcule las probabilidades de los incisos anteriores.

#### Solución

a) Si continúa con la misma metodología de cálculos de probabilidades tendremos:

$$P(\text{error tipo I al usar } R_r | H_0) = P(\bar{X} > 812 | \mu \leq 810) = P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > \frac{812 - \mu}{55/\sqrt{60}} \mid \mu = 800\right)$$

$$P(\text{error tipo I al usar } R_r | H_0) = P\left(Z > \frac{812 - 800}{55/\sqrt{60}}\right) = P(Z > 1.69) \approx \Phi(-1.69) = 0.0455$$

b) De igual manera, para calcular la probabilidad del error tipo II, con  $R_a$  dada por  $\bar{X} \leq 812$ :

$$\begin{aligned}
 P(\text{II si se usa } R_a | H_1) &= P(\bar{X} \leq 812 | \mu > 810) = P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq \frac{812 - \mu}{55/\sqrt{60}} \mid \mu = 815\right) = P\left(Z \leq \frac{812 - 815}{55/\sqrt{60}}\right) \\
 &= P(Z \leq -0.42) \approx \Phi(-0.42) = 0.3372
 \end{aligned}$$

#### Ejemplo 4.8 Cálculo de probabilidades

Suponga que el representante de la delegación Iztacalco de la Ciudad de México afirma que menos de 30% de sus habitantes están en contra de su nuevo proyecto para combatir la delincuencia. Si considera una muestra de 25 habitantes y las hipótesis:

$$H_0: p \geq 0.30$$

$$H_1: p < 0.30$$

calcule:

a) La probabilidad de cometer un error tipo I para el caso en que  $p = 0.32$ .

b) La probabilidad de cometer el error tipo II para el caso en que  $p = 0.28$ .

Defina las variables aleatorias  $X_i$  para  $i = 1, 2, \dots, 25$ , la persona entrevistada está en contra del nuevo proyecto. Además, utilice la región de rechazo  $T = \sum_{i=1}^{25} X_i < 7$ .

### Solución

a) Para calcular las probabilidades tenemos que la región de rechazo está

dada por  $T = \sum_{i=1}^{25} X_i < 7$  y que la distribución de  $T$  es binomial con parámetros  $n = 25$  y  $p$ . Luego:

$$P(\text{error tipo I al usar } R_r | H_0) = P(T < 7 | p \geq 0.30) = P(T < 7 | p = 0.32) = F_T(6) = 0.2657$$

Para los cálculos utilizamos la distribución binomial acumulada:

$$P(T < 7 | p = 0.32) = C_0^{25}(0.32)^0(0.68)^{25} + C_1^{25}(0.32)^1(0.68)^{24} + \dots + C_6^{25}(0.32)^6(0.68)^{19} \approx 0.2657$$

b) De igual manera, para calcular la probabilidad del error tipo II con  $R_a$

dada por  $T = \sum_{i=1}^{25} X_i \geq 7$ :

$$P(\text{II al usar } R_a | H_1) = P(T \geq 7 | p < 0.30) = 1 - P(T \leq 6 | p = 0.28) = 0.5753$$

De manera similar al inciso a), para los cálculos utilizamos la distribución binomial acumulada:

$$1 - P(T < 7 | p = 0.28) = 1 - [C_0^{25}(0.28)^0(0.72)^{25} + \dots + C_6^{25}(0.28)^6(0.72)^{19}] \approx 1 - 0.4247 = 0.5753$$

El resultado también se puede obtener de tablas con  $n = 25$  y una interpolación para  $F_T(6 | p = 0.30) = 0.3407$  y  $F_T(6 | p = 0.35) = 0.1734$ , con lo que se obtiene  $0.2738 \approx F_T(6 | p = 0.32) = 0.2657$ .

El resultado también se puede obtener de tablas con  $n = 25$  y una interpolación para  $F_T(6 | p = 0.25) = 0.5611$  y  $F_T(6 | p = 0.30) = 0.3407$ , cuyo resultado  $0.4289 \approx F_T(6 | p = 0.28) = 0.4247$ .

En este sentido, ¿cómo calcular el tamaño de la muestra cuando se indican los tamaños de los errores tipo I y II?

Cuando se considera que los tamaños de las probabilidades de los errores tipo I y II son conocidos, junto con la distribución de la población, se puede determinar el tamaño de la muestra.

### Ejemplo 4.9 Cálculo de probabilidades

Suponga que tenemos una población con distribución normal de la que conocemos su varianza igual a  $30 \text{ u}^2$  y el contraste de hipótesis;  $H_0: \mu \geq 54$  contra  $H_1: \mu < 54$ . Además, se estableció la región de rechazo para  $\bar{x} < a$ . ¿Cuál debe ser el valor crítico de la prueba  $a$ , y de qué tamaño se debe seleccionar la muestra aleatoria si se quiere un nivel de significancia igual a 0.05 y una probabilidad de error tipo II de 0.01 cuando  $\mu = 50$ ?

### Solución

a) Este tipo de problemas tienen un grado de complejidad un poco superior a los anteriores, pero se resuelve con el planteamiento de las dos probabilidades que se dan como datos y con éstas se obtiene un sistema de dos ecuaciones con dos incógnitas (valor crítico y tamaño de la muestra). Así, en este caso:

$$0.05 = \sup_{\mu \geq 54} P(\text{rechazar } H_0 | \mu \geq 54) = P(\bar{X} < a | \mu = 54) = P\left(Z < \frac{a - 54}{\sqrt{30}/\sqrt{n}}\right) = \Phi\left(\frac{a - 54}{\sqrt{30}} \sqrt{n}\right)$$

Note que el supremo de la probabilidad lo alcanza en el extremo (justo por eso algunos autores piden que los espacios paramétricos en la hipótesis nula sean cerrados), por esta razón, se usa  $\mu = 54$ . De igual manera, para la probabilidad del error tipo II:

$$0.01 = 1 - P(\text{rechazar } H_0 \mid \mu < 54) \Big|_{\mu=50} = 1 - P(\bar{X} < a \mid \mu = 50) = 1 - P\left(Z < \frac{a - 50}{\sqrt{30}/\sqrt{n}}\right) = 1 - \Phi\left(\frac{a - 50}{\sqrt{30}}\sqrt{n}\right)$$

Es decir:

$$0.01 = 1 - \Phi\left(\frac{a - 50}{\sqrt{30}}\sqrt{n}\right) \Rightarrow \Phi\left(\frac{a - 50}{\sqrt{30}}\sqrt{n}\right) = 1 - 0.01 = 0.99$$

Al simplificar las dos ecuaciones anteriores tendremos el siguiente sistema de dos ecuaciones con dos incógnitas:

$$0.05 = \Phi\left(\frac{a - 54}{\sqrt{30}}\sqrt{n}\right) \text{ y } 0.99 = \Phi\left(\frac{a - 50}{\sqrt{30}}\sqrt{n}\right)$$

En ambas ecuaciones extraemos la inversa de la función acumulada y despejamos a  $\sqrt{n}$ :

$$\begin{cases} \Phi^{-1}(0.05) = \frac{a - 54}{\sqrt{30}}\sqrt{n} \\ \Phi^{-1}(0.99) = \frac{a - 50}{\sqrt{30}}\sqrt{n} \end{cases} \Rightarrow \begin{cases} \frac{\Phi^{-1}(0.05)\sqrt{30}}{a - 54} = \sqrt{n} \\ \frac{\Phi^{-1}(0.99)\sqrt{30}}{a - 50} = \sqrt{n} \end{cases}$$

Si se igualan ambas ecuaciones resulta:

$$\frac{\Phi^{-1}(0.05)\sqrt{30}}{a - 54} = \frac{\Phi^{-1}(0.99)\sqrt{30}}{a - 50}$$

Y al despejar la variable  $a$ , tenemos:

$$\begin{aligned} \Phi^{-1}(0.05)(a - 50) &= \Phi^{-1}(0.99)(a - 54) \\ \Phi^{-1}(0.05)a - 50\Phi^{-1}(0.05) &= \Phi^{-1}(0.99)a - 54\Phi^{-1}(0.99) \\ \Phi^{-1}(0.05)a - \Phi^{-1}(0.99)a &= 50\Phi^{-1}(0.05) - 54\Phi^{-1}(0.99) \\ (\Phi^{-1}(0.05) - \Phi^{-1}(0.99))a &= 50\Phi^{-1}(0.05) - 54\Phi^{-1}(0.99) \end{aligned}$$

Por último:

$$a = \frac{50\Phi^{-1}(0.05) - 54\Phi^{-1}(0.99)}{\Phi^{-1}(0.05) - \Phi^{-1}(0.99)} = \frac{50(-1.6449) - 54(2.3263)}{-1.6449 - 2.3263} = 52.34$$

Ahora, para el tamaño de la muestra utilizamos alguna de las dos ecuaciones originales, por ejemplo:

$$\sqrt{n} = \frac{\Phi^{-1}(0.99)\sqrt{30}}{a - 50} \Rightarrow n = \left(\frac{2.3263 \times \sqrt{30}}{52.34 - 50}\right)^2 = 29.65$$

Concluimos que la constante crítica debe ser igual a 52.34, mientras que el tamaño de la muestra es 30.

## Ejercicios 4.1

1. El gerente de mercadotecnia de una red comercial se queja sobre cierto producto perecedero, afirma que menos de 80% de este producto dura el tiempo que el distribuidor dice. De manera que el gerente establece las hipótesis  $H_0: p \geq 0.80$  y  $H_1: p < 0.80$ . Elige una muestra aleatoria del tamaño de 20 artículos y establece la re-

gión crítica, rechazar  $H_0: p \geq 0.80$  si menos de 15 duran el tiempo que el distribuidor dice. Defina las variables aleatorias  $X_i$  si el producto  $i$  dura el tiempo que el distribuidor dice para  $i = 1, 2, \dots, 20$ .

a) Evalúe  $\alpha$ , al suponer que  $p = 0.8$ .

b) Evalúe  $\beta$  para la alternativa  $p = 0.7$ .

2. Del ejercicio anterior calcule la potencia de la prueba si  $p = 0.6$ .

3. El titular de la delegación Álvaro Obregón de la Ciudad de México afirma que los robos que ocurren en la demarcación cada día es menor a 4. Para comprobar esta afirmación pide revisar de manera aleatoria siete reportes diarios y establecer las hipótesis  $H_0: \lambda_T \geq 28$  y  $H_1: \lambda_T < 28$ . En la que  $\lambda_T$  es la razón de los robos en la delegación durante los siete días y  $T$  la variable aleatoria que representa el total de robos durante el periodo de análisis. Para tomar decisiones, el director establece la región crítica para  $T < 25$  y define las variables aleatorias  $X_i$  cantidad de robos que ocurrieron en la delegación en el día  $i$  para  $i = 1, 2, \dots, 7$ . Suponga que las variables  $X_i$  tienen una distribución de Poisson con  $\lambda = 4$  robos por día. *Sugerencia.* Recuerde que cuando

$X_i \sim P(\lambda)$ , entonces la distribución de  $T = \sum_{i=1}^n X_i$  está dada por  $T \sim P(n\lambda)$ .

a) Evalúe  $\alpha$ , al suponer que  $\lambda_T = 30$ .

b) Evalúe  $\beta$  para la alternativa  $\lambda_T = 24$ .

4. Del ejercicio anterior calcule la potencia de la prueba si  $\lambda_T = 26$ .

5. El gerente de publicidad de una televisora en México afirma que introducir los programas de series aumentó la audiencia a más de 60% del total de televidentes, de manera que las hipótesis son  $H_0: p \leq 0.60$  y  $H_1: p > 0.60$ . Para comprobar esta afirmación el gerente lleva a cabo una encuesta a 500 personas elegidas al azar. Si más de 320 contestan que sí ven el programa de las series, se aceptará la hipótesis  $H_1: p > 0.60$  (se rechaza la hipótesis nula), de otra forma se concluye que  $p \leq 0.60$ . Defina las variables aleatorias  $X_i$  si la persona  $i$  ve los programas de series, para  $i = 1, 2, \dots, 500$ . *Sugerencia.* Utilice el teorema central del límite.

a) Evalúe  $\alpha$ , al suponer que  $p = 0.6$ .

b) Evalúe  $\beta$  para la alternativa  $p = 0.65$ .

6. Del ejercicio anterior calcule la potencia de la prueba para  $p = 0.70$ .

7. Para verificar la proporción de habitantes de la Ciudad de México que están a favor de la forma de gobernar del PRD se lleva a cabo una muestra aleatoria de 400 habitantes de la ciudad y se les pregunta si están a favor de ésta. Si entre 220 y 260 están a favor del gobierno, se concluirá que 60% de los habitantes están a favor. Defina las variables aleatorias  $X_i$  si la persona  $i$  de la Ciudad de México está a favor de la forma de gobernar del PRD, para  $i = 1, \dots, 400$ . *Sugerencia.* Utilice el teorema central del límite.

a) Establezca el contraste de hipótesis.

b) Encuentre la probabilidad de cometer un error tipo I si 60% de los habitantes está a favor de la forma de gobernar del PRD.

8. En el ejercicio anterior, ¿cuál es la probabilidad de cometer un error tipo II al utilizar este procedimiento de prueba si solo 50% de los habitantes están a favor de la manera de gobernar del PRD?

9. Suponga que  $X$  es una variable aleatoria normal con varianza 100. Si se toma una muestra aleatoria de tamaño 16 pruebe la hipótesis  $H_0: \mu = 10$  y  $H_1: \mu \neq 10$ . Si considera la región de rechazo  $\bar{X} < 9$  o  $\bar{X} > 11$ , encuentre la expresión para la función de potencia de la prueba, y con la ayuda de algún paquete matemático trace su gráfica y escriba sus conclusiones.

10. Del ejercicio anterior:

a) Calcule el nivel de significancia de la prueba.

b) Evalúe  $\beta$  para la alternativa  $\mu = 12$ .

11. Del ejercicio anterior de la variable normal con varianza 100:

a) Calcule la potencia de la prueba para el caso en que la verdadera media sea  $\mu = 8$ , e interprete el resultado.



b) Suponga que se establece un nivel de significancia de 5% y se desea determinar los valores críticos correspondientes. ¿Será factible encontrarlos? En caso afirmativo determine dichos valores.

12. Un fabricante de llantas desconoce la proporción  $p$  de artículos defectuosos, supone que dicha proporción es menor a 15%. Para este efecto establece el siguiente juego de hipótesis:

$$H_0: p \geq 0.15$$

$$H_1: p < 0.15$$

El fabricante desea conocer los tamaños de los errores tipo I y II, así como la potencia de la prueba, para lo cual elige una muestra aleatoria de 25 llantas. Define una variable aleatoria  $X$  cantidad de llantas defectuosas en la muestra y considere una región crítica para  $X < 3$  o  $\hat{p} = \frac{X}{n} < \frac{3}{25}$ .

a) Encuentre la expresión para la función de potencia de la prueba y con la ayuda de algún paquete matemático trace su gráfica y escriba sus conclusiones.

b) Calcule el nivel de significancia de la prueba.

13. Del ejercicio anterior:

a) Obtenga una expresión para la probabilidad del error tipo II.

b) Evalúe  $\beta$  para la alternativa  $p = 0.10$ .

14. Del ejercicio anterior sobre la fabricación de llantas:

a) Calcule la potencia de la prueba para el caso en que la verdadera proporción de artículos defectuosos sea  $p = 0.14$ , e interprete el resultado.

b) Con los resultados encontrados y una realización de 25 llantas, 0 para defectuosos y 1 para buenas

0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0

Decida si rechaza la hipótesis nula.

a) Si considera el valor de la realización para la estadística de prueba, como el valor del parámetro, calcule su potencia.

## Conceptos básicos sobre los tipos de pruebas de hipótesis

Esta subsección la dedicaremos a un breve listado de conceptos básicos sobre los tipos de pruebas de hipótesis que utilizaremos en las siguientes secciones.

1. Cuando el espacio paramétrico correspondiente a una hipótesis tiene un solo elemento se llama hipótesis simple, y cuando tiene dos o más se llama hipótesis compuesta. Por ejemplo:  $H_0: \theta \in \omega$  con  $\omega = \{\theta_0\}$  es una hipótesis simple y  $H_0: \theta \in \omega$  con  $\omega = \{\theta_0, \theta_1, \theta_2, \dots\}$  es una hipótesis compuesta.
2. Cuando la hipótesis alterna es simple y la prueba tiene la mayor potencia se llama **prueba más potente de tamaño  $\alpha$** ,  $MP(\alpha)$ . Es decir,  $H_0: \theta \in \omega$  vs.  $H_1: \theta \in \Omega - \omega$  con  $\Omega - \omega = \{\theta_0\}$ .
3. Cuando la hipótesis alterna es compuesta y la prueba tiene la mayor potencia de todas las otras pruebas del mismo tamaño, se llama **prueba uniformemente más potente de tamaño  $\alpha$** ,  $UMP(\alpha)$ . Es decir,  $H_0: \theta \in \omega$  vs.  $H_1: \theta \in \Omega - \omega$  con  $\Omega - \omega = \{\theta_0, \theta_1, \theta_2, \dots\}$ .
4. Clasificación elemental de pruebas de hipótesis:
  - Una prueba se llama **unilateral** o **de una cola** si la hipótesis alterna es de la forma  $H_1: \theta > \theta_0$  o  $H_1: \theta < \theta_0$ .
  - Una prueba se llama **bilateral** o **de dos colas** si la hipótesis alterna es de la forma  $H_1: \theta < \theta_0$  o  $\theta > \theta_0$ , es común su uso cuando  $\theta_1 = \theta_2 = \theta_0$ . Es decir,  $H_1: \theta \neq \theta_0$ .
  - Una prueba se llama **simple contra simple** si la hipótesis nula y la alterna son simples. Aquí se busca la prueba  $MP(\alpha)$ .



- Una prueba se llama **simple contra compuesta** si la hipótesis nula es simple y la alterna es compuesta. Aquí se busca la prueba  $UMP(\alpha)$ .
  - Una prueba se llama **compuesta contra compuesta** si ambas hipótesis nula y alterna son compuestas. En caso de tratarse de una prueba unilateral se busca la prueba  $UMP(\alpha)$ . Pero en el caso de una prueba bilateral se busca la prueba  $UMPI(\alpha)$ , **prueba uniformemente más potente insesgada de tamaño  $\alpha$** .
  - Una prueba se llama **compuesta contra simple** si la hipótesis nula es compuesta y la alterna es simple. Aquí se busca la prueba  $MP(\alpha)$ .
5. Las combinaciones de hipótesis más comunes que se presentan en la práctica y que trabajaremos durante el texto son:
- a)  $H_0: \theta = \theta_0$  contra  $H_1: \theta \neq \theta_0$  (simple contra compuesta), se busca la  $UMPI(\alpha)$ .
  - b)  $H_0: \theta \leq \theta_0$  contra  $H_1: \theta > \theta_0$  (compuesta contra compuesta), se busca la  $UMP(\alpha)$ .
  - c)  $H_0: \theta \geq \theta_0$  contra  $H_1: \theta < \theta_0$  (compuesta contra compuesta), se busca la  $UMP(\alpha)$ .
  - d)  $H_0: \theta_0 \leq \theta \leq \theta_1$  contra  $H_1: \theta < \theta_0$  o  $\theta > \theta_1$  (compuesta contra compuesta), se busca la  $UMPI(\alpha)$ .

Además, para la elección de la hipótesis alterna se planteará la afirmación de que el investigador presupone como verdadera. Excepto en el caso cuando la afirmación esté dada por  $\theta = \theta_0$  y las hipótesis sean simple contra compuesta.

Los resultados de las pruebas que se establezcan en los casos que tratamos están dadas para el caso *compuesta contra compuesta* o *simple contra compuesta*, pero se pueden usar para las situaciones restantes de *simple contra simple*.

## Metodología para probar una hipótesis estadística

Para la prueba de hipótesis que realizaremos en las siguientes secciones se recomienda seguir los siguientes pasos.

1. **Formulación de hipótesis.** Establecer la hipótesis nula y la hipótesis alterna. La hipótesis nula y alterna se establecen con base en las reglas mencionadas.
2. **Nivel de significancia.** Fijar el nivel de significancia  $\alpha$ , el cual propone el investigador.
3. **EP y CC.** Con el valor de  $\alpha$  realizar los cálculos para determinar los cuantiles correspondientes, su estadística de prueba (EP) y la constante crítica (CC parte derecha de la regla de decisión) que se requieren en la regla de decisión dada en la fórmula.
4. **Decisión.** Aplicar la regla de decisión de la metodología, calculando el valor de la estadística de prueba para la realización dada y decidir si se rechaza o no la hipótesis nula.

## 4.2 Pruebas de hipótesis para los parámetros de una distribución normal

Así como en el caso de intervalos de confianza, analizaremos los parámetros de una distribución normal, media ( $\mu$ ) y la varianza ( $\sigma^2$ ). Iniciamos el desarrollo metodológico para las pruebas de hipótesis con el parámetro media, para el cual se analizan dos situaciones: cuando conocemos la varianza poblacional y cuando es desconocida.

### Pruebas de hipótesis para la media de poblaciones aproximadamente normales cuando se conoce $\sigma$

En esta subsección se formula el resultado correspondiente a los cuatro casos de la prueba de hipótesis para el parámetro media de una población normal con el parámetro varianza conocido.

## Teorema 4.1

Sea  $X_1, X_2, \dots, X_n$  una muestra aleatoria de  $N(\mu, \sigma_0^2)$ , entonces podemos tener alguno de los siguientes contrastes de hipótesis, con **EP**  $\bar{X} \sim N\left(\mu, \frac{\sigma_0^2}{n}\right)$  o  $Z = \frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}} \sim N(0, 1)$

- a)  $H_0: \mu \geq \mu_0$  contra  $H_1: \mu < \mu_0$ , luego la prueba UMP( $\alpha$ ), para  $\alpha \in (0, 1)$  estará dada por la siguiente regla de decisión para una realización  $x_1, x_2, \dots, x_n$ :

$$\text{Rechazar } H_0: \mu \geq \mu_0 \text{ si } \bar{x} < \mu_0 + \frac{\sigma_0}{\sqrt{n}} \Phi^{-1}(\alpha) \text{ o } \bar{x} < \mu_0 - \frac{\sigma_0}{\sqrt{n}} Z_\alpha$$

- b)  $H_0: \mu \leq \mu_0$  contra  $H_1: \mu > \mu_0$ , luego la prueba UMP( $\alpha$ ), para  $\alpha \in (0, 1)$  estará dada por la siguiente regla de decisión para una realización  $x_1, x_2, \dots, x_n$ :

$$\text{Rechazar } H_0: \mu \leq \mu_0, \text{ si } \bar{x} > \mu_0 + \frac{\sigma_0}{\sqrt{n}} \Phi^{-1}(1 - \alpha) \text{ o } \bar{x} > \mu_0 + \frac{\sigma_0}{\sqrt{n}} Z_\alpha$$

- c)  $H_0: \mu = \mu_0$  contra  $H_1: \mu \neq \mu_0$ , luego la prueba UMPI( $\alpha$ ), para  $\alpha \in (0, 1)$  estará dada por la siguiente regla de decisión para una realización  $x_1, x_2, \dots, x_n$ :

$$\text{Rechazar } H_0: \mu = \mu_0, \text{ si } \bar{x} < \mu_0 + \frac{\sigma_0}{\sqrt{n}} \Phi^{-1}(\alpha/2) \text{ o } \bar{x} > \mu_0 + \frac{\sigma_0}{\sqrt{n}} \Phi^{-1}(1 - \alpha/2)$$

$$\text{rechazar } H_0: \mu = \mu_0, \text{ si } \bar{x} < \mu_0 - \frac{\sigma_0}{\sqrt{n}} Z_{\alpha/2} \text{ o } \bar{x} > \mu_0 + \frac{\sigma_0}{\sqrt{n}} Z_{\alpha/2}$$

- d)  $H_0: \mu_0 \leq \mu \leq \mu_1$  contra  $H_1: \mu < \mu_0$  o  $\mu > \mu_1$ , luego la prueba UMPI( $\alpha$ ), para  $\alpha \in (0, 1)$  estará dada por la siguiente regla de decisión para una realización  $x_1, x_2, \dots, x_n$ :

$$\text{Rechazar } H_0: \mu_0 \leq \mu \leq \mu_1, \text{ si } \bar{x} < \mu_0 + \frac{\sigma_0}{\sqrt{n}} \Phi^{-1}(\alpha/2) \text{ o } \bar{x} > \mu_1 + \frac{\sigma_0}{\sqrt{n}} \Phi^{-1}(1 - \alpha/2)$$

$$\text{rechazar } H_0: \mu_0 \leq \mu \leq \mu_1, \text{ si } \bar{x} < \mu_0 - \frac{\sigma_0}{\sqrt{n}} Z_{\alpha/2} \text{ o } \bar{x} > \mu_1 + \frac{\sigma_0}{\sqrt{n}} Z_{\alpha/2}$$

Con  $\mu_0, \mu_1 \in \mathbb{R}$  y  $\sigma_0^2 > 0$  valores conocidos de antemano, en donde,  $\Phi^{-1}(\gamma)$  representa el cuantil  $\gamma$  de la distribución normal estándar para  $\gamma \in (0, 1)$  y  $Z_\gamma$  representa el valor de la variable normal estándar cuya área a la derecha es  $\gamma \in (0, 1)$ , es decir  $\Phi^{-1}(\gamma) = Z_{1-\gamma}$ .

Ahora, vamos a revisar cuatro ejemplos para ilustrar el uso de este caso de pruebas de hipótesis.

## Ejemplo 4.10 Pruebas de hipótesis

Una máquina produce piezas metálicas de forma cilíndrica. Se toma una muestra de nueve piezas cuyos diámetros son 9.8, 9.5, 9.8, 11.5, 9.0, 10.4, 9.8, 10.1 y 11.2 cm. Suponga que los diámetros tienen una distribución aproximadamente normal con una varianza de  $0.64 \text{ cm}^2$ . Si el fabricante de las piezas afirma que su diámetro promedio es de 10 cm.

- a) Plantee el contraste de hipótesis adecuado al problema.  
b) Aplique la metodología indicada y determine la prueba UMPI ( $\alpha$ ) para  $\alpha = 0.01$ .

- c) ¿Qué puede decir respecto a la afirmación del fabricante con un nivel de significancia de 0.01?  
 d) Si el valor de  $\mu = 10.5$  cm, calcule la potencia de la prueba con la regla de decisión b). Después de calcular la potencia de la prueba cambiaría su impresión de la afirmación sobre la prueba.

### Solución

a) Se pide una prueba de hipótesis para la media, en la que se desea probar que la media de las piezas metálicas es igual a 10 cm (espacio paramétrico cerrado),  $H_0: \mu = 10$ . Entonces, la hipótesis alterna será el opuesto, es decir diferente de 10,  $H_1: \mu_1 \neq 10$ .

b) Al seguir los pasos para una prueba de hipótesis:

- $H_0: \mu = 10$  contra  $H_1: \mu_1 \neq 10$ .
- Nivel de significancia  $\alpha = 0.01$ .
- Estamos ante una situación como la del inciso c) del teorema 4.1. Requerimos calcular la CC:

$$\mu_0 + \frac{\sigma_0}{\sqrt{n}} \Phi^{-1}(\alpha/2) \text{ y } \mu_0 + \frac{\sigma_0}{\sqrt{n}} \Phi^{-1}(1 - \alpha/2) \text{ y comparar con el valor de la EP } \bar{x}.$$

Al calcular cada elemento, tenemos  $\mu_1 = 10$ ,  $\sigma_0 = \sqrt{0.64} = 0.80$ ,  $n = 9$ ,  $\alpha = 0.01$ , con las tablas porcentuales para la distribución normal estándar,  $\Phi^{-1}(0.01/2) = \Phi^{-1}(0.5\%) = -2.5758$  y  $\Phi^{-1}(1 - 0.01/2) = \Phi^{-1}(99.5\%) = 2.5758$ . Al final, la regla de decisión.

$$\text{Rechazar: } H_0: \mu = 10, \text{ si } \bar{x} < \mu_0 + \frac{\sigma_0}{\sqrt{n}} \Phi^{-1}(\alpha/2) = 10 - \frac{0.80}{\sqrt{9}}(2.5758) = 9.31 \text{ o}$$

$$\bar{x} > \mu_0 + \frac{\sigma_0}{\sqrt{n}} \Phi^{-1}(1 - \alpha/2) = 10 + \frac{0.80}{\sqrt{9}}(2.5758) = 10.69$$

Es decir, rechazar  $H_0: \mu = 10$  si  $\bar{x} < 9.31$  o  $\bar{x} > 10.69$ . De manera gráfica podemos observar en la figura 4.6 qué ocurre con las regiones de la prueba de hipótesis.



Figura 4.6 Regiones de la prueba del ejemplo 4.10.

- Por último, aplicamos la regla de decisión, para lo cual calculamos el valor de la estadística de prueba con los valores de la realización, de lo que se obtiene  $\bar{x} = 10.12$  que resulta fuera de la región de rechazo,  $\bar{x} < 9.31$  o  $\bar{x} > 10.69$ . Luego, concluimos que  $H_0: \mu = 10$  no se rechaza con 1% de significancia.
- c) **Conclusión:** con 1% de significancia y la realización obtenida no existen evidencias para rechazar la hipótesis nula  $H_0: \mu = 10$ .
- d) Para calcular la potencia de la prueba utilizamos la región de rechazo,

$$\begin{aligned} P(\text{rechazar } H_0 \mid \mu \neq 10) &= P(\bar{X} < 9.31 \text{ o } \bar{X} > 10.69 \mid \mu = 10.5) = P\left(Z < \frac{9.31 - 10.50}{0.80/\sqrt{9}} \text{ o } Z > \frac{10.69 - 10.50}{0.80/\sqrt{9}}\right) \\ &= P(Z < -4.46) + P(Z > 0.71) = 0.2389 \end{aligned}$$

Aunque la prueba es la UMPI (0.01), su potencia es baja, para el caso de que la verdadera media sea de 10.5 cm.

Del ejemplo anterior surgen las preguntas: ¿podremos aumentar la potencia?, en caso afirmativo, ¿cómo hacerlo?

Resulta que la potencia puede aumentarse de varias formas. Una es mediante el incremento del nivel de significancia y la otra si se acumula más información, ya que mientras más información se tiene de un problema la incertidumbre disminuye.

### Ejemplo 4.11 Prueba de potencia

En el ejemplo 4.10 aumente la potencia de la prueba.

- a) Con un nivel de significancia de 10%.  
 b) Al agregar información: 9.8, 9.5, 9.8, 11.5, 9.0, 10.4, 9.8, 10.1, 11.2, 10.4, 9.8, 10, 10.2, 9.7, 9.9, 10.1, 10.2, 10.1, 9.5, 9.9, 9.9, 10.5, 10.7, 10.1 y 10.2 cm.

### Solución

Del ejemplo anterior tenemos planteadas las hipótesis, falta agregar los cambios en los cálculos.

a) En este caso al aumentar el nivel de significancia los incisos *i*) e *ii*) se conservan y los cálculos son:

- iii*) Requerimos calcular la CC:  $\mu_0 + \frac{\sigma_0}{\sqrt{n}} \Phi^{-1}(\alpha/2)$  y  $\mu_0 + \frac{\sigma_0}{\sqrt{n}} \Phi^{-1}(1 - \alpha/2)$ . Por tablas porcentuales para la distribución normal estándar,  $\Phi^{-1}(0.10/2) = \Phi^{-1}(5\%) = -1.6449$  y  $\Phi^{-1}(95\%) = 1.6449$ . Por otro lado,  $\sigma_0 = 0.80$ ,  $n = 9$  y  $\mu_0 = 10$ . Por último, la regla de decisión.

$$\text{Rechazar: } H_0: \mu = 10, \text{ si } \bar{x} < 10 - \frac{0.80}{\sqrt{9}}(1.6449) = 9.561 \text{ o } \bar{x} > 10 + \frac{0.80}{\sqrt{9}}(1.6449) = 10.439$$

Es decir, rechazar  $H_0: \mu = 10$  si  $\bar{x} < 9.561$  o  $\bar{x} > 10.439$ . En la figura 4.7 se aprecia lo que sucede con las regiones de la prueba de hipótesis.



Figura 4.7 Regiones de la prueba del ejemplo 4.11a.

- iv*) Por último, se aplica la regla de decisión. Para esto calculamos  $\bar{x} = 10.12$ , valor que queda fuera de la región de rechazo,  $\bar{x} < 9.561$  o  $\bar{x} > 10.439$ .

**Conclusión:** con 10% de significancia no hay evidencias para rechazar  $H_0: \mu = 10$ , no se rechaza. La potencia en estas condiciones será:

$$\begin{aligned} P(\text{rechazar } H_0 \mid \mu \neq 10) &= P(\bar{X} < 9.561 \text{ o } \bar{X} > 10.439 \mid \mu = 10.5) = P\left(Z < \frac{9.561 - 10.5}{0.80/\sqrt{9}} \text{ o } Z > \frac{10.439 - 10.5}{0.80/\sqrt{9}}\right) \\ &= P(Z < -3.52) + P(Z > -0.23) = 0.5912 \end{aligned}$$

b) Ahora, aumentamos la información, los incisos *i*) y *ii*) se conservan y los cálculos serán:

- iii*) En el ejemplo anterior encontramos  $\Phi^{-1}(0.01/2) = -2.5758$ ,  $\Phi^{-1}(99.5\%) = 2.5758$ . Por otro lado,  $\sigma_0 = 0.80$ ,  $n = 25$  y  $\mu_0 = 10$ . Por último, la regla de decisión.

$$\text{Rechazar: } H_0: \mu = 10, \text{ si } \bar{x} < 10 - \frac{0.80}{\sqrt{25}}(2.5758) = 9.59 \text{ o } \bar{x} > 10 + \frac{0.80}{\sqrt{25}}(2.5758) = 10.41$$

Es decir, rechazar  $H_0: \mu = 10$  si  $\bar{x} < 9.59$  o  $\bar{x} > 10.41$ . En la figura 4.8 podemos apreciar de forma gráfica las regiones de la prueba de hipótesis.



Figura 4.8 Regiones de la prueba del ejemplo 4.11b.

iv) Por último aplicamos la regla de decisión, para esto calculamos el valor de la estadística de prueba correspondiente a la realización, de lo que se obtiene  $\bar{x} = 10.09$  valor que queda fuera de la región de rechazo  $\bar{x} < 9.59$  o  $\bar{x} > 10.41$ .

**Conclusión:** con 1% de significancia no hay evidencias para rechazar  $H_0: \mu = 10$ , no se rechaza.

Para calcular la potencia de la prueba utilizamos la región de rechazo:

$$\begin{aligned} P(\text{rechazar } H_0 \mid \mu \neq 10) &= P(\bar{X} < 9.59 \text{ o } \bar{X} > 10.41 \mid \mu = 10.5) = P\left(Z < \frac{9.59 - 10.5}{0.80/5}\right) + P\left(Z > \frac{10.41 - 10.5}{0.80/5}\right) \\ &= P(Z < -5.69) + P(Z > -0.56) = 0.7123 \end{aligned}$$

En el ejemplo anterior comprobamos que la potencia de la prueba la podemos aumentar al incrementar el nivel de significancia o la información.

#### Ejemplo 4.12 Prueba de hipótesis

Los encargados de un centro de atención ciudadana, donde está instalada una máquina de refrescos, recibieron constantes quejas por parte de los usuarios, quienes indican que se despacha menos líquido que el estipulado en las instrucciones (240 ml de refresco en promedio). Por tal motivo, deciden cambiar la máquina si al revisar una muestra aleatoria y llevar a cabo una prueba de hipótesis a 5% de significancia se decide de manera estadística que es válida la afirmación de los usuarios. Si supone que la cantidad de líquido despachada por la máquina tiene una distribución aproximadamente normal con una desviación estándar igual a 15 ml:

- Plantee un contraste de hipótesis adecuado para el problema y lleve a cabo la prueba si una muestra aleatoria de 36 refrescos arroja un contenido promedio de 225 ml.
- Calcule la potencia de la prueba suponga que  $\mu = 230$  ml.
- Explique cómo apoya el resultado obtenido de manera estadística a una toma de decisiones sobre el retiro o no de la máquina.

#### Solución

a) Se pide una prueba de hipótesis para la media, en donde los usuarios afirman que la máquina en promedio despacha menos de 240 ml,  $\mu < 240$ . Luego, tendremos:

- $H_0: \mu \geq 240$  contra  $H_1: \mu < 240$ .
- Nivel de significancia  $\alpha = 0.05$ .

iii) Estamos ante una situación similar a la del inciso a) del teorema 4.1. Requerimos calcular la CC:  $\mu_0 + \frac{\sigma_0}{\sqrt{n}}\Phi^{-1}(\alpha)$  y comparar con el valor de la EP  $\bar{x}$ .

Si se calcula cada elemento, tenemos  $\mu_0 = 240, \sigma_0 = 15, n = 36, \alpha = 0.05$ , con las tablas porcentuales para la distribución normal estándar,  $\Phi^{-1}(0.05) = \Phi^{-1}(5\%) = -1.6449$ . Por último, la regla de decisión.

$$\text{Rechazar: } H_0: \mu = 240, \text{ si } \bar{x} < \mu_0 + \frac{\sigma_0}{\sqrt{n}}\Phi^{-1}(\alpha) = 240 - \frac{15}{\sqrt{36}}(1.6449) = 235.9$$

Es decir, rechazar  $H_0: \mu = 240$  si  $\bar{x} < 235.9$ . En la figura 4.9 se puede apreciar de forma gráfica las regiones de la prueba de hipótesis.



Figura 4.9 Regiones de la prueba del ejemplo 4.12a.

iv) Por último, aplicamos la regla de decisión, si se recuerda que  $\bar{x} = 225 \in R_r$ . Así, concluimos que  $H_0: \mu = 240$  se rechaza al 5% de significancia.

**Conclusión:** con 5% de significancia y la realización obtenida no existen evidencias para refutar la afirmación de los clientes de que  $\mu < 240$ .

b) Para calcular la potencia de la prueba utilizamos la región de rechazo:

$$P(\text{rechazar } H_0 \mid \mu < 240) = P(\bar{X} < 235.9 \mid \mu = 230) = P\left(Z < \frac{235.9 - 230}{15/\sqrt{36}}\right) = P(Z < 2.36) = 0.991$$

Si la verdadera media fuera 230 mL la prueba tendría una potencia muy elevada.

c) El resultado ayuda de manera estadística a tomar la decisión de retirar la máquina despachadora de refresco, ya que a un nivel de significancia de 5%, rechazamos la hipótesis nula en apoyo a la afirmación del consumidor. Además, la potencia de la prueba es bastante elevada y de la teoría sabemos que esta prueba es la UMP(0.05), es decir, en estas condiciones no existe otra prueba más potente.

¿Cómo realizar una prueba de hipótesis cuando se tiene una tabla de frecuencias?

**Ejemplo 4.13**

Una máquina de refrescos está ajustada de manera que la cantidad de líquido despachado se distribuye aproximadamente normal con una desviación estándar de 15 ml. Se elige una muestra de tamaño 60 y un trabajador registra el líquido despachado por clases de frecuencia, mostradas a la derecha. El fabricante afirma que la máquina despacha en promedio 240 ml. A un nivel de significancia de 5% pruebe si es válida la afirmación del fabricante.

**Solución**

Para la solución se siguen los mismos pasos que para datos no agrupados, solo cambian las fórmulas para calcular los valores muestrales requeridos, en este caso nos referimos al promedio.

**Tabla 4.2** Frecuencias.

Intervalos de clase	Frecuencias ( $n_i$ )
[239, 241)	4
[241, 243)	10
[243, 245)	20
[245, 247)	11
[247, 249)	12
[249, 251)	3

Si se calcula la media por clases de frecuencias,  $\bar{x}_f = \frac{1}{n} \sum_{i=1}^m x_i^m n_i = \frac{1}{60} \sum_{i=1}^6 x_i^m n_i$ .

Las marcas de clase se obtienen de la tabla 4.2: 240, 242, 244, 246, 248 y 250. Así:

$$\bar{x}_f = \frac{1}{60}(240 \times 4 + 242 \times 10 + 244 \times 20 + 246 \times 11 + 248 \times 12 + 250 \times 3) = 244.8667$$

Para el planteamiento del contraste de hipótesis notamos que la afirmación del fabricante se refiere a una igualdad de media  $\mu < 240$ , el conjunto paramétrico es cerrado, luego la hipótesis nula será  $\mu < 240$ .

Al seguir los pasos de la metodología para pruebas de hipótesis.

i)  $H_0: \mu = 240$  contra  $H_1: \mu \neq 240$ .

ii) Nivel de significancia  $\alpha = 0.05$ .

iii) Estamos ante una situación como la del inciso c) del teorema 4.1. Requerimos calcular la CC:

$$\mu_0 + \frac{\sigma_0}{\sqrt{n}} \Phi^{-1}(\alpha/2) \text{ y } \mu_0 + \frac{\sigma_0}{\sqrt{n}} \Phi^{-1}(1 - \alpha/2) \text{ y comparar con el valor de la EP } \bar{x}.$$

Si se calcula cada elemento,  $\mu_0 = 240$ ,  $\sigma_0 = 15$ ,  $n = 60$ ,  $\alpha = 0.05$ , con las tablas porcentuales para la distribución normal estándar,  $\Phi^{-1}(0.025) = -1.96$  y  $\Phi^{-1}(1 - 0.02) = 1.96$ . Por último, la regla de decisión:

$$\text{Rechazar: } H_0: \mu = 240, \text{ si } \bar{x}_f < \mu_0 + \frac{\sigma_0}{\sqrt{n}} \Phi^{-1}(\alpha/2) = 240 - \frac{15}{\sqrt{60}}(1.96) = 236.20 \text{ o}$$

$$\bar{x}_f > \mu_0 + \frac{\sigma_0}{\sqrt{n}} \Phi^{-1}(1 - \alpha/2) = 240 + \frac{15}{\sqrt{60}}(1.96) = 243.80$$

Es decir, rechazar  $H_0: \mu = 240$  si  $\bar{x}_f < 236.20$  o  $\bar{x}_f > 243.80$ . En la figura 4.10 es posible observar las regiones de la prueba de hipótesis.



Figura 4.10 Regiones de la prueba del ejemplo 4.13.

iv) Por último, aplicamos la regla de decisión, al recordar que  $\bar{x}_f = 244.87$ .

**Conclusión:** con 5% de significancia hay evidencias para rechazar  $H_0: \mu = 240$ , se rechaza.

## Pruebas de hipótesis para la media de poblaciones aproximadamente normales cuando se desconoce $\sigma$

En esta subsección se formula el resultado correspondiente a los cuatro casos de la prueba de hipótesis para el parámetro media de una población normal con el parámetro varianza desconocido.

### Teorema 4.2

Sea  $X_1, X_2, \dots, X_n$  una muestra aleatoria de  $N(\mu, \sigma^2)$  entonces podemos tener alguno de los siguientes contrastes de hipótesis, con **EP**  $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$  o  $T = \frac{\bar{X} - \mu_0}{S_{n-1}/\sqrt{n}} \sim t_{n-1}$ .



a)  $H_0: \mu \geq \mu_0$  contra  $H_1: \mu < \mu_0$ , luego la prueba  $UMP(\alpha)$ , para  $\alpha \in (0, 1)$  estará dada por la siguiente regla de decisión para una realización  $x_1, x_2, \dots, x_n$ :

$$\text{Rechazar } H_0: \mu \geq \mu_0, \text{ si } \bar{x} < \mu_0 + \frac{S_{n-1}}{\sqrt{n}} F_{t_{n-1}}^{-1}(\alpha) \text{ o } \bar{x} < \mu_0 - \frac{S_{n-1}}{\sqrt{n}} t_{\alpha}(n-1)$$

b)  $H_0: \mu \leq \mu_0$  contra  $H_1: \mu > \mu_0$ , luego la prueba  $UMP(\alpha)$ , para  $\alpha \in (0, 1)$  estará dada por la siguiente regla de decisión para una realización  $x_1, x_2, \dots, x_n$ :

$$\text{Rechazar } H_0: \mu \leq \mu_0, \text{ si } \bar{x} > \mu_0 + \frac{S_{n-1}}{\sqrt{n}} F_{t_{n-1}}^{-1}(1-\alpha) \text{ o } \bar{x} > \mu_0 + \frac{S_{n-1}}{\sqrt{n}} t_{\alpha}(n-1)$$

c)  $H_0: \mu = \mu_0$  contra  $H_1: \mu \neq \mu_0$ , luego la prueba  $UMPI(\alpha)$ , para  $\alpha \in (0, 1)$  estará dada por la siguiente regla de decisión para una realización  $x_1, x_2, \dots, x_n$ :

$$\text{Rechazar } H_0: \mu_0 \leq \mu \leq \mu_1, \text{ si } \bar{x} < \mu_0 + \frac{S_{n-1}}{\sqrt{n}} F_{t_{n-1}}^{-1}(\alpha/2) \text{ o } \bar{x} > \mu_0 + \frac{S_{n-1}}{\sqrt{n}} F_{t_{n-1}}^{-1}(1-\alpha/2)$$

$$\text{Rechazar } H_0: \mu_0 \leq \mu \leq \mu_1, \text{ si } \bar{x} < \mu_0 - \frac{S_{n-1}}{\sqrt{n}} t_{\alpha/2}(n-1) \text{ o } \bar{x} > \mu_0 + \frac{S_{n-1}}{\sqrt{n}} t_{\alpha/2}(n-1)$$

d)  $H_0: \mu_0 \leq \mu \leq \mu_1$  contra  $H_1: \mu < \mu_0 \text{ o } \mu > \mu_1$ , luego la prueba  $UMPI(\alpha)$ , para  $\alpha \in (0, 1)$  estará dada por la siguiente regla de decisión para una realización  $x_1, x_2, \dots, x_n$ :

$$\text{Rechazar } H_0: \mu_0 \leq \mu \leq \mu_1, \text{ si } \bar{x} < \mu_0 + \frac{S_{n-1}}{\sqrt{n}} F_{t_{n-1}}^{-1}(\alpha/2) \text{ o } \bar{x} > \mu_1 + \frac{S_{n-1}}{\sqrt{n}} F_{t_{n-1}}^{-1}(1-\alpha/2)$$

$$\text{Rechazar } H_0: \mu_0 \leq \mu \leq \mu_1, \text{ si } \bar{x} < \mu_0 - \frac{S_{n-1}}{\sqrt{n}} t_{\alpha/2}(n-1) \text{ o } \bar{x} > \mu_1 + \frac{S_{n-1}}{\sqrt{n}} t_{\alpha/2}(n-1)$$

Con  $\mu_0, \mu_1 \in \mathbb{R}$  valores conocidos de antemano, donde,  $F_{t_{n-1}}^{-1}(\gamma)$  representa el cuantil  $\gamma$  de la distribución t-Student con  $n-1$  grados de libertad para  $\gamma \in (0, 1)$ ,  $t_{\gamma}$  constituye el valor de la variable t-Student con  $n-1$  grados de libertad cuya área derecha es  $\gamma \in (0, 1)$ .

Revisemos tres ejemplos para ilustrar el uso de este caso de pruebas de hipótesis.

#### Ejemplo 4.14 Pruebas de hipótesis

El gerente de ventas de una empresa productora de bombillas de luz en su reunión con los representantes de diferentes centros comerciales, afirma que su producto tiene una duración promedio mayor a 800 horas. Antes de realizar la compra, cada representante de los centros comerciales decide comprobar de manera estadística la afirmación del fabricante. Con ese objetivo, eligieron una muestra de tamaño 26 y obtuvieron un tiempo de vida promedio de  $\bar{x} = 810$  horas con una desviación estándar muestral de 45 horas.

- Plantee un contraste de hipótesis adecuado para el problema, con un nivel de significancia de 5% y si supone normalidad en el tiempo de vida de las bombillas justifique si es o no válida la afirmación de los fabricantes.
- Calcule la potencia de la prueba al suponer que  $\mu = 820$  horas.
- Explique cómo apoya el resultado obtenido de manera estadística a una toma de decisiones a los representantes de los centros comerciales sobre la afirmación del fabricante de bombillas.

**Solución**

a) Para plantear el contraste de hipótesis tenemos en cuenta que la afirmación del fabricante es  $\mu > 800$  horas y no un conjunto cerrado. Luego,

i)  $H_0: \mu \leq 800$  contra  $H_1: \mu > 800$ .

ii) Nivel de significancia  $\alpha = 0.05$ .

iii) Estamos ante una situación como la del inciso b) del teorema 4.2. Es decir, requerimos calcular la CC:

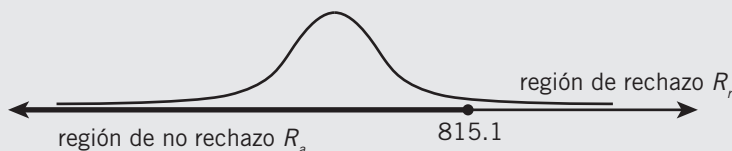
$$\mu_0 + \frac{s_{n-1}}{\sqrt{n}} F_{t_{n-1}}^{-1}(1 - \alpha) \text{ y comparar con el valor de la EP } \bar{x}.$$

Al calcular cada elemento,  $\mu_0 = 800$ ,  $s_{n-1} = 45$ ,  $n - 1 = 26 - 1 = 25$ ,  $\alpha = 0.05$ , con las tablas porcentuales para la distribución t-Student,  $F_{t_{25}}^{-1}(1 - 0.05) = 1.708$ . Por último, la regla de decisión.

Rechazar:  $H_0: \mu \leq 800$ , si

$$\bar{x} > \mu_0 + \frac{s_{n-1}}{\sqrt{n}} F_{t_{n-1}}^{-1}(1 - \alpha) = 800 + \frac{45}{\sqrt{26}}(1.708) = 815.1$$

Es decir, rechazar  $H_0: \mu \leq 800$  si  $\bar{x} > 815.1$ . En la figura 4.11 se muestran las regiones de la prueba de hipótesis.



**Figura 4.11** Regiones de la prueba del ejemplo 4.14.

iv) Por último, aplicamos la regla de decisión, para esto recordamos que  $\bar{x} = 810 < 815.1$ , con lo cual concluimos que al 5% de significancia y la realización considerada no hay evidencias para rechazar  $H_0: \mu \leq 800$ .

**Conclusión:** con 5% de significancia y la realización obtenida no es válida la afirmación del fabricante de que  $\mu > 800$ .

b) Para calcular la potencia de la prueba utilizamos la región de rechazo:

$$P(\text{rechazar } H_0 \mid \mu > 800) = P(\bar{X} > 815.1 \mid \mu = 820) = P\left(T = \frac{815.1 - 820}{45/\sqrt{26}}\right) = P(T > -0.5552) = 0.7082$$

c) En este caso, los compradores deberían tomar la decisión de no comprar las bombillas de luz o tomar varias realizaciones más para corroborar su toma de decisiones. Además, de la teoría sabemos que esta prueba es la UMP(0.05), es decir, en estas condiciones no existe otra prueba más potente.

1. Para calcular el valor de la probabilidad se puede utilizar cualquier paquete estadístico, por ejemplo, Excel. En la pestaña de función escribir: =1-DISTR.T.CD(0.5552,25) (25 son los grados de libertad. 0.5552 es el valor de la variable con el que se va a calcular la probabilidad a la derecha. Para las probabilidades de la distribución acumulada o cola izquierda se utiliza =DISTR.T.CD(0.5552,25)=0.70812
2. En caso de no tener un paquete estadístico se usan las tablas con los valores más próximos a 0.5552, obtenemos con 25 grados de libertad  $P(T > 0.6844) = 0.25$  y  $P(T > 0.5312) = 0.30$ . Si se interpola  $P(T > 0.5552) \approx 0.2922$ . Al final,  $P(T > -0.5552) \approx 0.7078$ , valor muy próximo al encontrado con Excel, 0.7082 (un error de 4 diez milésimas).

**Ejemplo 4.15** Prueba de hipótesis

Los fabricantes de máquinas despachadoras de bebida afirman que sus máquinas despachan entre [225, 245] mililitros de bebida. Para probar la afirmación los consumidores toman una muestra aleatoria de 30 servicios

de la máquina de bebidas, de lo que se obtiene un contenido promedio de 229 ml, con una desviación estándar de 25 ml. Si supone normalidad en la cantidad de líquido despachada por la máquina y con base en una prueba estadística al nivel de significancia de 0.04.

- Plantee un contraste adecuado de hipótesis para el problema y justifique si es o no válida la afirmación de los fabricantes de máquinas despachadoras de bebida.
- Explique cómo apoya el resultado obtenido de manera estadística a una toma de decisiones sobre la afirmación del fabricante de estas máquinas.

### Solución

a) Para el planteamiento del contraste de hipótesis notamos que la afirmación del fabricante es que  $\mu \in [225, 245]$ , el conjunto paramétrico es cerrado en ambos extremos, luego la hipótesis nula será  $225 \leq \mu \leq 245$ . De forma que tendremos los pasos para la prueba:

i)  $H_0: 225 \leq \mu \leq 245$  contra  $H_1: \mu < 225 \cup \mu > 245$ .

ii) Nivel de significancia  $\alpha = 0.04$ .

iii) Estamos ante una situación como la del inciso d) del teorema 4.2. Requerimos calcular la CC:

$$\mu_0 + \frac{s_{n-1}}{\sqrt{n}} F_{t_{n-1}}^{-1}(\alpha/2) \text{ y } \mu_1 + \frac{s_{n-1}}{\sqrt{n}} F_{t_{n-1}}^{-1}(1 - \alpha/2) \text{ y comparar con el valor de la EP } \bar{x}.$$

Si se calcula cada elemento,  $\mu_0 = 225$ ,  $\mu_1 = 245$ ,  $s_{n-1} = 25$ ,  $n - 1 = 30 - 1 = 29$ ,  $\alpha = 0.04$ , con las tablas porcentuales para la distribución t-Student,  $F_{t_{29}}^{-1}(1 - 0.02) = 2.150$  y  $F_{t_{29}}^{-1}(0.02) = -2.150$ . Al final, la regla de decisión.

$$\text{Rechazar: } H_0: 225 \leq \mu \leq 245, \text{ si } \bar{x} < \mu_0 + \frac{s_{n-1}}{\sqrt{n}} F_{t_{n-1}}^{-1}(\alpha/2) = 225 - \frac{25}{\sqrt{30}}(2.150) = 215.2 \text{ o}$$

$$\bar{x} > \mu_1 + \frac{s_{n-1}}{\sqrt{n}} F_{t_{n-1}}^{-1}(1 - \alpha/2) = 245 + \frac{25}{\sqrt{30}}(2.150) = 254.8$$

Es decir, rechazar  $H_0: 225 \leq \mu \leq 245$  si  $\bar{x} < 215.2$  o  $\bar{x} > 254.8$ . De manera gráfica podemos apreciar las regiones de la prueba en la figura 4.12.



Figura 4.12 Regiones de la prueba del ejemplo 4.15.

iv) Por último, aplicamos la regla de decisión, recordando que  $\bar{x} = 229 \in [215.2, 254.8]$ .

**Conclusión:** con 4% de significancia y la realización tomada no hay evidencias para rechazar  $H_0: 225 \leq \mu \leq 245$ .

- Con un nivel de significancia de 4% y la realización elegida podemos concluir que no existe evidencia para refutar la afirmación del fabricante. Además, de la teoría sabemos que esta prueba es la UMPI(0.04), es decir, en estas condiciones no existe otra prueba más potente.

Como puede apreciarse, el uso del teorema 4.2 está limitado a las tablas de distribución t-Student. En general están elaboradas para valores de  $n \leq 30$ , por consiguiente, surge la pregunta, ¿qué hacer cuando se desconoce  $\sigma$  y el tamaño de la muestra es mayor a 30?

En las tablas que utilizamos se tienen valores para grados de libertad de 1 a 30; 40, 50, 60, 70, 80, 90 y 100, para otros grados de libertad podemos utilizar la interpolación o aproximar la distribución t-Student con la distribución normal para grados de libertad grandes.

#### Ejemplo 4.16 Pruebas de hipótesis

De acuerdo con las normas establecidas para un examen de aptitud mecánica, las personas de 18 años deberían promediar menos de 73.2. Si 45 personas de esa edad elegidas de manera aleatoria promediaron 66.7 con desviación estándar de 8.6, pruebe el contraste de hipótesis para la media poblacional:

$$H_0: \mu \geq 73.2 \text{ contra } H_1: \mu < 73.2$$

Encuentre la prueba UMP (0.05). Suponga normalidad en las calificaciones de los exámenes.

- Si usa una aproximación de la distribución t-Student con la normal (grados de libertad 44).
- Si usa un paquete estadístico para el valor del cuantil de la distribución  $t$ .
- Si usa la interpolación de los valores de las tablas de la distribución  $t$ .

#### Solución

a) Los grados de libertad son 44 mayores a 30, utilizaremos una aproximación de la distribución t-Student por la normal.

i)  $H_0: \mu \geq 73.2$  contra  $H_1: \mu < 73.2$ .

ii) Nivel de significancia  $\alpha = 0.05$ .

iii) Estamos ante una situación como la del inciso a) del teorema 4.2, pero con grados de libertad mayores a 30, luego utilizaremos una aproximación con la normal. Requerimos calcular la CC:  $\mu + \frac{s_{n-1}}{\sqrt{n}} \Phi^{-1}(\alpha)$  y comparar con el valor de la EP  $\bar{x}$ .

Si se calcula cada elemento,  $\mu_0 = 73.2$ ,  $s_{n-1} = 8.6$ ,  $n = 45$ ,  $\alpha = 0.05$ , con las tablas porcentuales para la distribución normal estándar  $\Phi^{-1}(0.05) = -1.6449$ . Por último, la regla de decisión.

$$\text{Rechazar: } H_0: \mu \geq 73.2, \text{ si } \bar{x} < \mu_0 + \frac{s_{n-1}}{\sqrt{n}} \Phi^{-1}(\alpha) = 73.2 - \frac{8.6}{\sqrt{45}}(1.6449) = 71.09$$

Es decir, rechazar  $H_0: \mu \geq 73.2$  si  $\bar{x} < 71.09$ . En la figura 4.16 se muestran las regiones de la prueba.

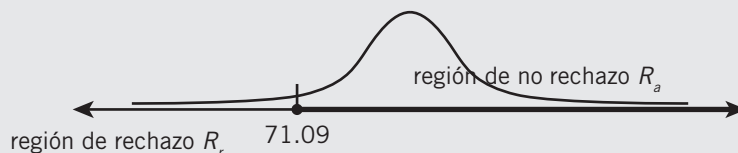


Figura 4.13 Regiones de la prueba del ejemplo 4.16.

iv) Por último, aplicamos la regla de decisión, al recordar que  $\bar{x} = 66.7 < 71.09$ .

**Conclusión:** con 5% de significancia y la realización tomada hay evidencias para rechazar  $H_0: \mu \geq 73.2$ .

b) Los pasos i) y ii) son iguales al inciso a), solo cambiará el inciso iii), ya que en lugar de  $\Phi^{-1}(0.05) = -1.6449$ , se utiliza  $F_{t_{44}}^{-1}(0.05) = -1.6802$ . De forma que la regla de decisión es:

$$\text{Rechazar: } H_0: \mu \geq 73.2, \text{ si } \bar{x} < \mu_0 + \frac{s_{n-1}}{\sqrt{n}} F_{t_{44}}^{-1}(0.05) = 73.2 - \frac{8.6}{\sqrt{45}}(1.6802) = 71.05$$

Es decir, rechazar  $H_0: \mu \geq 73.2$  si  $\bar{x} < 71.05$ , diferente del inciso a) solo en cuatro centésimas, la conclusión no cambiará.

- c) Ahora en lugar de utilizar  $\Phi^{-1}(0.05) = -1.6449$  o  $F_{t_{44}}^{-1}(0.05) = -1.6802$  emplearemos una interpolación, para  $F_{t_{44}}^{-1}(0.05)$  con los valores más próximos de tablas;  $F_{t_{40}}^{-1}(0.05) = -1.684$  y  $F_{t_{50}}^{-1}(0.05) = -1.676$ . Al obtener  $F_{t_{44}}^{-1}(0.05) \approx -1.6808$ , valor que se diferencia del encontrado con el paquete en solo seis diez milésimas y la conclusión no cambiará.

Con este ejemplo mostramos que cuando se trate de pruebas de hipótesis para la media en donde no se conozca la varianza muestral, pero el tamaño sea grande podemos utilizar la aproximación por la normal.

## Pruebas para la varianza de poblaciones normales

El otro parámetro de una población normal, varianza, se estudió en los intervalos de confianza con estadística de prueba  $S_{n-1}^2$  y distribución ji cuadrada con  $n - 1$  grados de libertad dada por:

$$\chi_{n-1}^2 = \frac{(n-1)S_{n-1}^2}{\sigma^2}$$

### Teorema 4.3

Sea  $X_1, X_2, \dots, X_n$  una muestra aleatoria de  $N(\mu, \sigma^2)$  entonces podemos tener alguno de los siguientes contrastes de hipótesis:

- a)  $H_0: \sigma^2 \geq \sigma_0^2$  contra  $H_1: \sigma^2 < \sigma_0^2$ , entonces la prueba de tamaño  $\alpha$ , para  $\alpha \in (0, 1)$  estará dada por la siguiente regla de decisión para una realización  $x_1, x_2, \dots, x_n$ :

$$\text{Rechazar } H_0: \sigma^2 \geq \sigma_0^2, \text{ si } s_{n-1}^2 < \frac{\sigma_0^2}{n-1} F_{\chi_{n-1}^2}^{-1}(\alpha) \text{ o } s_{n-1}^2 < \frac{\sigma_0^2}{n-1} \chi_{n-1}^2(1-\alpha)$$

- b)  $H_0: \sigma^2 \leq \sigma_0^2$  contra  $H_1: \sigma^2 > \sigma_0^2$ , entonces la prueba de tamaño  $\alpha$ , para  $\alpha \in (0, 1)$  estará dada por la siguiente regla de decisión para una realización  $x_1, x_2, \dots, x_n$ :

$$\text{Rechazar } H_0: \sigma^2 \leq \sigma_0^2, \text{ si } s_{n-1}^2 > \frac{\sigma_0^2}{n-1} F_{\chi_{n-1}^2}^{-1}(1-\alpha) \text{ o } s_{n-1}^2 > \frac{\sigma_0^2}{n-1} \chi_{n-1}^2(\alpha)$$

- c)  $H_0: \sigma^2 = \sigma_0^2$  contra  $H_1: \sigma^2 \neq \sigma_0^2$ , entonces la prueba de tamaño  $\alpha$ , para  $\alpha \in (0, 1)$  estará dada por la siguiente regla de decisión para una realización  $x_1, x_2, \dots, x_n$ :

$$\text{Rechazar } H_0: \sigma^2 = \sigma_0^2, \text{ si } s_{n-1}^2 < \frac{\sigma_0^2}{n-1} F_{\chi_{n-1}^2}^{-1}(\alpha/2) \text{ o } s_{n-1}^2 > \frac{\sigma_0^2}{n-1} F_{\chi_{n-1}^2}^{-1}(1-\alpha/2)$$

$$\text{rechazar } H_0: \sigma^2 = \sigma_0^2, \text{ si } s_{n-1}^2 < \frac{\sigma_0^2}{n-1} \chi_{n-1}^2(1-\alpha/2) \text{ o } s_{n-1}^2 > \frac{\sigma_0^2}{n-1} \chi_{n-1}^2(\alpha/2)$$

- d)  $H_0: \sigma_0^2 \leq \sigma^2 \leq \sigma_1^2$  contra  $H_1: \sigma^2 < \sigma_0^2$  o  $\sigma^2 > \sigma_1^2$ , entonces la prueba de tamaño  $\alpha$ , para  $\alpha \in (0, 1)$  estará dada por la siguiente regla de decisión para una realización  $x_1, x_2, \dots, x_n$ :

$$\text{Rechazar } H_0: \sigma_0^2 \leq \sigma^2 \leq \sigma_1^2, \text{ si } s_{n-1}^2 < \frac{\sigma_0^2}{n-1} F_{\chi_{n-1}^2}^{-1}(\alpha/2) \text{ o } s_{n-1}^2 > \frac{\sigma_1^2}{n-1} F_{\chi_{n-1}^2}^{-1}(1-\alpha/2)$$

Con  $\sigma_0^2, \sigma_1^2 \in \mathbb{R}^+$  valores conocidos de antemano, donde,  $F_{\chi_n^2}^{-1}(\gamma)$  representa el cuantil  $\gamma$  de la distribución  $ji$  cuadrada con  $n$  grados de libertad para  $\gamma \in (0, 1)$ , y  $\chi_n^2(\gamma)$  representa el valor de la variable  $ji$  cuadrada con  $n$  grados de libertad cuya área derecha es  $\gamma \in (0, 1)$ .

En los siguientes dos ejemplos mostramos aplicaciones del teorema 4.3.

#### Ejemplo 4.17 Aplicaciones del teorema 4.3

Una máquina produce piezas metálicas de forma cilíndrica. Se toma una muestra de nueve piezas cuyos diámetros son 9.8, 9.5, 9.8, 11.5, 9.0, 10.4, 9.8, 10.1 y 11.2 mm, respectivamente. Suponga que los diámetros de las piezas tienen una distribución aproximadamente normal. Si el fabricante de dichas piezas afirma que su diámetro promedio tiene una varianza menor a  $1 \text{ mm}^2$ .

- Plantee el contraste de hipótesis adecuado al problema para probar la afirmación del fabricante.
- Aplique la metodología indicada y determine la prueba de tamaño  $\alpha = 0.01$ .
- ¿Qué puede indicar con respecto a la afirmación del fabricante con un nivel de significancia de 0.01?
- Si el valor de  $\sigma^2 = 0.5 \text{ mm}^2$ , calcule la potencia de la prueba.

#### Solución

- Se pide una prueba de hipótesis para la varianza, en donde deseamos probar que la varianza de las piezas metálicas es menor a  $1 \text{ mm}^2$  (espacio paramétrico abierto),  $H_1: \sigma^2 < 1$ . Luego, la hipótesis nula será el opuesto, es decir mayor o igual a  $1 \text{ mm}^2$ ,  $H_0: \sigma^2 \geq 1$ .
- Si sigue los pasos para una prueba de hipótesis.

i)  $H_0: \sigma^2 \geq 1$  contra  $H_1: \sigma^2 < 1$ .

ii) Nivel de significancia  $\alpha = 0.01$ .

iii) Estamos ante una situación como la del inciso a) del teorema 4.3. Requerimos calcular la

$$CC: \frac{\sigma_0^2}{n-1} F_{\chi_{n-1}^2}^{-1}(\alpha) \text{ y comparar con el valor de la EP } s_{n-1}^2.$$

Si se calcula cada elemento,  $\sigma_0^2 = 1$  y  $\alpha = 0.01$ , con las tablas porcentuales para la distribución  $ji$  cuadrada con  $n - 1 = 9 - 1 = 8$  g.l.;  $F_{\chi_{n-1}^2}^{-1}(0.01) = 1.6465$ . Por último, la regla de decisión.

$$\text{Rechazar: } H_0: \sigma^2 \geq 1, \text{ si } s_{n-1}^2 < \frac{\sigma_0^2}{n-1} F_{\chi_{n-1}^2}^{-1}(\alpha) = \frac{1}{9-1}(1.6465) = 0.206$$

Es decir, rechazar  $H_0: \sigma^2 \geq 1$ , si  $s_{n-1}^2 < 0.206$ . De manera gráfica, tenemos las regiones de la prueba en la figura 4.14.

- Para calcular el valor de la probabilidad con la distribución  $ji$  cuadrada en Excel-Microsoft 2016 se escribe en la pestaña la función: =DISTR.CHICUAD(3.296, 8, 1)=0.085567.... 3.296 es el valor del cuantil con el que se va a determinar la probabilidad con 8 g.l. y 1 indica que se calcula la probabilidad acumulada.
- Para probabilidades de cola derecha se utiliza =DISTR.CHICUAD.CD(3.296,8) =0.914433...  
Con tablas se buscan los valores más próximos a 3.296 con 8 g.l.  $P(\chi_8^2 < 3.2881) = 0.085$  y  $P(\chi_8^2 < 3.3570) = 0.090$ , interplanos  $P(\chi_8^2 < 3.296)$ ;  $P(\chi_8^2 < 3.296) \approx 0.08557$ , valor muy próximo al encontrado de manera directa con algún paquete, 0.085567.

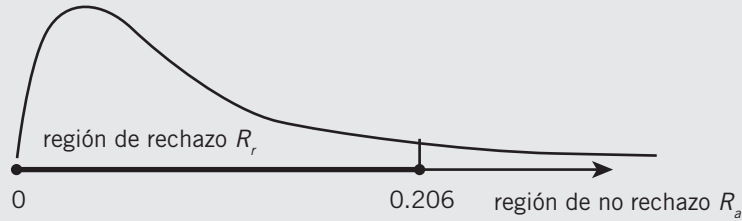


Figura 4.14 Regiones de la prueba del ejemplo 4.17.

- iv) Por último, aplicamos la regla de decisión, para lo cual calculamos el valor de la estadística de prueba según la realización, de lo que se obtiene  $s_{n-1}^2 = 0.637$  valor que queda fuera de la región de rechazo,  $s_{n-1}^2 < 0.206$ .
- c) **Conclusión:** con 1% de significancia en la realización obtenida no existen evidencias para rechazar  $H_0: \sigma^2 \geq 1$ , no se rechaza.
- d) Para calcular la potencia de la prueba utilizamos la región de rechazo y estadística ji cuadrada:

$$P(\text{rechazar } H_0 \mid \sigma^2 < 1) = P(S_{n-1}^2 < 0.206 \mid \sigma^2 = 0.5) = P\left(\frac{(n-1)S_{n-1}^2}{\sigma^2} < \frac{(9-1)0.206}{0.5}\right) = P(\chi_8^2 < 3.296) = 0.0856$$

Si la verdadera varianza es  $0.5 \text{ mm}^2$  la potencia de la prueba es pésima.

### Ejemplo 4.18 Aplicaciones del teorema 4.3

Los fabricantes de máquinas despachadoras de café afirman que sus máquinas sirven la bebida con una desviación estándar igual a 20 ml. Para probar la afirmación, los consumidores toman una muestra aleatoria de 30 servicios de la máquina de café de los que obtienen un contenido promedio de 229 ml, con una desviación estándar de 25 ml. Suponga normalidad en la cantidad de café despachada por la máquina.

- a) Plantee un contraste de hipótesis adecuado para el problema y justifique si es o no válida la afirmación de los fabricantes de máquinas despachadoras a un nivel de significancia de 0.04.
- b) Calcule la potencia de la prueba si  $\sigma = 28$ .

#### Solución

a) Para el planteamiento del contraste de hipótesis note que la afirmación del fabricante es  $\sigma = 20$ , luego  $H_0: \sigma = 20$  contra  $H_1: \sigma \neq 20$ . Si sigue los pasos de la prueba, primero revisamos un estadístico, por lo cual, elevamos al cuadrado para trabajar con la varianza, sobre la que sí conocemos una metodología.

i)  $H_0: \sigma^2 = 20^2$  contra  $H_1: \sigma^2 \neq 20^2$ .

ii) Nivel de significancia  $\alpha = 0.04$ .

iii) Estamos ante una situación como la del inciso c) del teorema 4.3. Requerimos calcular CC:

$$\frac{\sigma_0^2}{n-1} F_{\chi_{n-1}^2}^{-1}(\alpha/2) \text{ y } \frac{\sigma_0^2}{n-1} F_{\chi_{n-1}^2}^{-1}(1-\alpha/2) \text{ y comparar con el valor de la EP } s_{n-1}^2.$$

Al calcular cada elemento,  $\sigma_0^2 = 400$  y  $\alpha = 0.04$ , con las tablas porcentuales para la distribución ji cuadrada con  $n - 1 = 30 - 1 = 29$  grados de libertad  $F_{\chi_{n-1}^2}^{-1}(1 - 0.02) = 46.6926$  y  $F_{\chi_{n-1}^2}^{-1}(0.02) = 15.5745$ . Al final, la regla de decisión.

$$\text{Rechazar: } H_0: \sigma^2 = 400, \text{ si } s_{n-1}^2 < \frac{\sigma_0^2}{n-1} F_{\chi_{n-1}^2}^{-1}(\alpha/2) = \frac{400}{30-1}(15.5745) = 214.82 \text{ o}$$



$$s_{n-1}^2 > \frac{\sigma_1^2}{n-1} F_{\chi_{n-1}^2}^{-1}(1 - \alpha/2) = \frac{400}{30-1} (46.6926) = 644.04$$

Es decir, rechazar  $H_0: \sigma^2 = 400$  si  $s_{n-1}^2 < 214.82$  o  $s_{n-1}^2 > 644.04$ . De manera gráfica se muestran las regiones de la prueba en la figura 4.15.



Figura 4.15 Regiones de la prueba del ejemplo 4.18.

iv) Por último, aplicamos la regla de decisión para  $s_{n-1}^2 = 625$  valor que queda fuera de la región de rechazo,  $s_{n-1}^2 < 214.82$  o  $s_{n-1}^2 > 644.04$ .

**Conclusión:** con 4% de significancia y la realización tomada que no hay evidencias para rechazar  $H_0: \sigma^2 = 400$  o  $H_0: \sigma = 20$ .

b) Para calcular la potencia de la prueba utilizamos la región de rechazo:

$$\begin{aligned} P(\text{rechazar } H_0 \mid \sigma^2 \neq 400) &= P(S_{n-1}^2 < 214.82 \cup S_{n-1}^2 > 644.04 \mid \sigma^2 = 28^2) \\ &= P\left(\frac{(n-1)S_{n-1}^2}{\sigma^2} < \frac{(30-1)214.82}{784}\right) + P\left(\frac{(n-1)S_{n-1}^2}{\sigma^2} > \frac{(30-1)644.04}{784}\right) \\ &= P(\chi_{29}^2 < 7.946) + 1 - P(\chi_{29}^2 > 23.823) = 0.00004 + 0.73761 = 0.73765 \end{aligned}$$

Si la verdadera varianza es 625 ml<sup>2</sup> la potencia de la prueba es aceptable.

## Ejercicios 4.2

- Un fabricante de máquinas despachadoras de refresco asegura que sirven un promedio de 250 ml, pero debido a quejas de los consumidores sobre una máquina en particular, decide verificarla al servir 20 veces la máquina y obtener un promedio de 247 ml con una desviación estándar de 10.5 ml.
  - Plantee el contraste de hipótesis apropiado para este problema.
  - Justifique a un nivel de significancia de 0.10 si es válida la afirmación del fabricante.
- Del ejercicio anterior:
  - Calcule la potencia de la prueba en el inciso b), si supone que  $\mu = 240$  ml.
  - Calcule la potencia de la prueba en el inciso b), si supone que  $\mu = 255$  ml.
  - Explique cómo apoya el resultado obtenido de forma estadística a una toma de decisiones a los fabricantes de las máquinas.
- Del ejercicio anterior sobre las máquinas despachadoras de refresco:
  - El fabricante afirma que la máquina tiene una varianza en el despachado de refresco menor a 120 ml. Decide retirar la máquina si su afirmación resulta errónea con  $\alpha = 0.05$ .
  - Calcule la potencia de la prueba si supone que  $\sigma = 9$  ml.
- Mientras efectúan una tarea determinada en condiciones simuladas de ausencia de gravedad, el ritmo cardiaco de 31 astronautas en adiestramiento, se incrementa en promedio 24 pulsaciones por minuto con una desvia-



ción estándar de 4.28 pulsaciones por minuto. Los médicos aseguran que dicho aumento en promedio  $\mu \in [25, 27]$  pulsaciones por minuto. Suponga normalidad en el aumento de las pulsaciones de los ritmos cardiacos de los astronautas.

- a) Plantee el contraste de hipótesis apropiado para este problema.
  - b) A un nivel de significancia de 0.05, ¿será válida la afirmación de los médicos?
5. Del ejercicio anterior calcule la potencia de la prueba, suponga que  $\mu = 29$  pulsaciones por minuto.
6. Del ejercicio anterior sobre el ritmo cardiaco en ausencia de gravedad:
- a) Los médicos aseguran que la desviación estándar de las pulsaciones es menor a 5. Plantee el juego adecuado de hipótesis y verifique con 0.05 de significancia, si la afirmación es válida.
  - b) Si la verdadera desviación estándar es de 3.9, calcule la potencia de la prueba en el inciso a).
7. Un médico al estudiar cierto medicamento llegó a la conclusión de que el tiempo promedio en que hace efecto en los pacientes es inferior a 40 minutos. Para probar su conclusión, el médico elige una muestra aleatoria de 15 pacientes, con un resultado de  $\bar{x} = 35$  minutos y  $s_{n-1}^2 = 170 \text{ min}^2$ . Por estudios previos se sabe que la población es normal con varianza de  $100 \text{ min}^2$ . El médico ve que ambas tienen una gran diferencia; entonces, decide que para tener una mayor certeza estadística, primero debe realizar una prueba para la varianza. Plantee el contraste de hipótesis apropiado para validar con 0.05 de significancia la suposición de la varianza de los estudios previos.
8. Del ejercicio anterior, suponga que la verdadera varianza vale  $200 \text{ min}^2$  y calcule la potencia de la prueba.
9. Del ejercicio anterior sobre un cierto medicamento:
- a) Plantee el contraste de hipótesis apropiado para la conclusión del tiempo promedio del médico.
  - b) Justifique a un nivel de significancia de 0.05, si es válida la afirmación del médico. Utilice el resultado del ejercicio anterior sobre la varianza.
  - c) Calcule la potencia de la prueba en el inciso b) para  $\mu = 34$  minutos.
10. Una empresa que utiliza botellas de vidrio de 2 L comenzó a quejarse del producto, porque su gerente asegura que las botellas en promedio tienen un espesor menor a 4 mm con una desviación estándar mayor a 0.07 mm. Para hacerlo, el gerente manda a un ingeniero de control de calidad, para que pruebe su aseveración. El ingeniero elige de manera aleatoria una muestra de 25 botellas de vidrio y mide su espesor, al encontrar una media y desviación estándar muestral de 3.9 mm y 0.09 mm, respectivamente. Suponga normalidad en la distribución del espesor de las paredes de las botellas de vidrio de dos litros.
- a) Plantee los contrastes de hipótesis apropiados para este problema.
  - b) Justifique a un nivel de significancia de 0.05, si de manera estadística es válida la afirmación del gerente de la empresa.
11. Del ejercicio anterior:
- a) Calcule la potencia de la prueba en el inciso b), suponga que  $\mu = \bar{x}$  mm.
  - b) Calcule la potencia de la prueba, suponga que  $\sigma = 0.10$  mm.
12. Los instrumentos de precisión utilizados por la Profeco para supervisar están diseñados para proporcionar una medición que se considera solo correcta en promedio. Los clientes que llegan a una gasolinera se han quejado de que reciben en promedio menos gasolina que la marcada en el medidor. Entonces, la institución manda a un supervisor a verificar la bomba de gasolina. En la bomba de la estación de gasolina, el supervisor realiza 10 mediciones en garrafones de 20 L: 20.5, 19.99, 20.0, 20.3, 19.90, 20.05, 19.79, 19.85, 19.95 y 20.15. Si supone normalidad en las mediciones de 20 L, verifique de forma estadística la afirmación de los consumidores.
- a) Plantee el contraste de hipótesis apropiado para este problema.
  - b) Justifique a un nivel de significancia de 0.03, si es válida la afirmación de los consumidores mediante la estadística.

13. Del ejercicio anterior calcule la potencia de la prueba, suponga que  $\mu = 19.7$  L.
14. Del ejercicio anterior sobre la Profeco, para reforzar su decisión la institución medirá también la desviación estándar en los 10 garrafones, si ésta es mayor a 0.2 L, reforzará la queja de los consumidores de gasolina.
- Plantee el contraste de hipótesis apropiado para este problema.
  - Justifique a un nivel de significancia de 0.03, si con estadística esto refuerza la afirmación de los consumidores.
15. Del ejercicio anterior de la desviación estándar sobre la gasolina, calcule la potencia de la prueba, suponga que  $\sigma = 0.3$  L.
16. Se lleva a cabo un estudio sobre los envases de un lubricante específico, para lo cual se toma una muestra aleatoria de 10 envases, y se obtienen sus contenidos: 9.2, 9.7, 10.1, 10.3, 10.1, 9.8, 9.9, 10.4, 10.3 y 9.8 L, respectivamente. El fabricante afirma que el contenido medio de los envases es mayor a 9.5. Suponga una normalidad con  $\sigma = 0.4$  en los contenidos de los envases, pruebe la afirmación del fabricante.
- Plantee el contraste de hipótesis apropiado para este problema.
  - Justifique a un nivel de significancia de 0.05, si de manera estadística es válida la afirmación del fabricante sobre el contenido promedio de los envases.
17. Del ejercicio anterior calcule la potencia de la prueba, suponga que la media poblacional vale 9.45.
18. Del ejercicio anterior sobre los envases de lubricante:
- Realice una prueba de hipótesis con 5% de significancia para la suposición de que  $\sigma = 0.4$ .
  - Decida si fue válida la suposición de que  $\sigma = 0.4$ .
19. El IPC de la empresa BMB se muestra en la tabla 4.3 y se supone que tiene una distribución normal durante el año. Suponga que la desviación estándar del IPC es igual a 1.4 y pruebe dicha suposición a 4% de significancia.

Tabla 4.3

Fecha	BMB	Fecha	BMB
09/06/2013	25.02	08/24/2013	24.54
09/03/2013	24.84	08/23/2013	24.12
09/02/2013	25.19	08/20/2013	24.09
09/01/2013	24.80	08/19/2013	24.19
08/31/2013	24.83	08/18/2013	23.85
08/30/2013	24.60	08/17/2013	23.52
08/27/2013	24.44	08/16/2013	23.51
08/26/2013	24.30	08/13/2013	23.35
08/25/2013	24.31		

20. Con base en el resultado del ejercicio anterior, resuelva lo siguiente:
- Sus dirigentes aseguran que el IPC en promedio ese año fue igual a 24.5. Plantee el contraste de hipótesis apropiado para este problema.
  - Justifique a un nivel de significancia de 0.04, si en la estadística es válida la afirmación de los dirigentes de la empresa.
21. Del ejercicio anterior calcule la potencia de la prueba, suponga que  $\mu = 23.90$ , para la empresa BMB.

22. La cámara de comercio de una ciudad asegura que según sus estudios económicos, la cantidad promedio de dinero que gastan cada día las personas que asisten a convenciones, incluyendo comidas, alojamiento y entretenimiento, está entre \$950 y \$1 000. Para probar esta afirmación un supervisor de la cámara de comercio seleccionó a 16 personas que asisten a convenciones y les preguntó qué cantidad de dinero gastaban por día, de lo que se obtuvo la siguiente información (en pesos):

940 875 863 948 942 989 835 874 868 852 958 884 1034 1046 955 963

Suponga que la cantidad de dinero gastada en un día se distribuye de manera normal.

- Plantee el contraste de hipótesis apropiado para este problema.
  - Justifique a un nivel de significancia de 0.05, si estadísticamente es válida la afirmación de la cámara de comercio de la ciudad.
23. Del ejercicio anterior, calcule la potencia de la prueba, suponga que  $\mu = \$1\,050$ .
24. Del ejercicio anterior sobre la cámara de comercio:
- Suponga que la cámara de comercio asegura que los gastos ocurren con una desviación estándar menor a \$120. A 5% de significancia pruebe si es válida dicha suposición.
  - Si la verdadera desviación de los gastos es de \$140, encuentre la potencia de la prueba.
25. Los resultados de una investigación se muestran en la distribución de frecuencias en la tabla 4.4. Suponga que los datos se obtuvieron de una población aproximadamente normal. Con un nivel de significancia de 4%, pruebe el contraste de hipótesis  $H_0: \mu \in [4.3, 4.5]$  contra  $H_1: \mu \notin [4.3, 4.5]$ .

Tabla 4.4

Intervalos de clase (Problema 15)	Frecuencias
[0.15, 1.55)	2
[1.55, 2.95)	6
[2.95, 4.35)	11
[4.35, 5.75)	17
[5.75, 7.15)	10
[7.15, 8.55)	3
[8.55, 9.95]	2

26. Un geólogo que pretendía estudiar el movimiento de los cambios relativos en la corteza terrestre en un sitio particular, en un intento por determinar el ángulo medio de las fracturas, eligió  $n = 51$  de éstas y encontró que la media y la desviación estándar muestral eran  $41.8^\circ$  y  $13.2^\circ$ , respectivamente. Con los resultados de su estudio afirma que los ángulos medios de las fracturas de la corteza terrestre en el sitio estudiado son menores a  $43^\circ$ . Suponga que sus ángulos en la corteza terrestre de este sitio se distribuyen de manera normal.
- Plantee el contraste de hipótesis apropiado para este problema.
  - Justifique a un nivel de significancia de 0.10, si en estadística es válida la afirmación del geólogo.
27. Del ejercicio anterior calcule la potencia de la prueba, suponga que  $\mu = 39.5^\circ$ .
28. Del ejercicio anterior sobre el estudio del geólogo se afirma que la desviación estándar de las fracturas es menor a  $15^\circ$ . Pruebe a 10% de significancia si es válida la afirmación.
29. Del ejercicio anterior calcule la potencia de la prueba, si la verdadera desviación es  $12^\circ$ .

### 4.3 Pruebas de hipótesis para comparar dos poblaciones normales

El problema de la prueba de hipótesis para la diferencia de medias tiene las mismas ideas en su aplicación que los intervalos de confianza. Es decir, tendremos los mismos casos para comparar diferencia de medias:

- Cuando se conocen las varianzas poblacionales.
- Cuando se desconocen las varianzas poblacionales, pero se sabe que son iguales.
- Cuando se desconocen las varianzas poblacionales, pero se sabe que son diferentes.
- Muestras pareadas.

Comparaciones que podemos utilizar para hacer afirmaciones sobre la mejor calidad, mayor tiempo de vida, etc., de un producto sobre otro.

El problema se puede plantear de la siguiente forma: Sean dos poblaciones independientes con distribución normal,  $N(\mu_1, \sigma_1^2)$  y  $N(\mu_2, \sigma_2^2)$  que representan el comportamiento de dos fenómenos de interés que deseamos comparar, por ejemplo, el aprovechamiento de un grupo en dos materias diferentes, la producción de leche por vacas de dos establos diferentes, la duración de refrigeradores similares de dos marcas diferentes, etc., deseamos conocer si existe igualdad en sus medias. Al principio, el problema se planteó para la afirmación de que son iguales:

$$H_0: \mu_1 = \mu_2 \text{ contra } H_1: \mu_1 \neq \mu_2$$

Después, se generalizó para afirmaciones en las que una media es mejor que la otra:

$$H_0: \mu_1 \leq \mu_2 \text{ contra } H_1: \mu_1 > \mu_2 \text{ o } H_0: \mu_1 \geq \mu_2 \text{ contra } H_1: \mu_1 < \mu_2$$

Ahora, podemos generalizarlo aún más según sea la afirmación que se desee probar; recuerde que en la hipótesis nula van cerrados los intervalos; mientras que en la hipótesis alterna, sus extremos siempre van abiertos.

En las pruebas de hipótesis para comparar dos poblaciones podemos utilizar algunas frases como:

1. La media del proceso 1 es superior a la del proceso 2 en  $d_0$ ,  $\mu_1 = \mu_2 + d_0$

$$H_0: \mu_1 - \mu_2 = d_0 \text{ contra } H_1: \mu_1 - \mu_2 \neq d_0$$

2. La media del proceso 1 es superior a la del proceso 2 en más de  $d_0$ ,  $\mu_1 > \mu_2 + d_0$ .

$$H_0: \mu_0 - \mu_2 \leq d_0 \text{ contra } H_1: \mu_1 - \mu_2 > d_0$$

3. La media del proceso 1 es superior a la del proceso 2 en menos de  $d_0$ ,  $\mu_2 < \mu_1 < \mu_2 + d_0$

$$H_0: \mu_1 - \mu_2 \leq 0 \cup \mu_1 - \mu_2 \geq d_0 \text{ contra}$$

$$H_1: 0 < \mu_1 - \mu_2 < d_0$$

4. Etcétera.

5. La media del proceso 1 es inferior a la del proceso 2 en  $d_0$ ,  $\mu_1 = \mu_2 - d_0$ .

$$H_0: \mu_2 - \mu_1 = d_0 \text{ contra } H_1: \mu_2 - \mu_1 \neq d_0$$

6. La media del proceso 1 es inferior a la del proceso 2 en más de  $d_0$ ,  $\mu_1 < \mu_2 - d_0$ .

$$H_0: \mu_2 - \mu_1 \leq d_0 \text{ contra } H_1: \mu_2 - \mu_1 > d_0$$

7. La media del proceso 1 es inferior a la del proceso 2 en **menos** de  $d_0$ ,  $\mu_2 - d_0 < \mu_1 < \mu_2$

$$H_0: \mu_2 - \mu_1 \leq 0 \cup \mu_2 - \mu_1 \geq d_0 \text{ contra}$$

$$H_1: 0 < \mu_2 - \mu_1 < d_0$$

8. Etcétera.

Es común que estas pruebas sean usadas para evaluar los resultados de investigaciones en agricultura, medicina, educación, psicología, sociología y muchos otros campos de la ciencia.

### Pruebas de hipótesis para la diferencia de medias sobre poblaciones aproximadamente normales cuando se conocen

Este caso se deduce con facilidad de las distribuciones muestrales, puesto que la diferencia de las medias muestrales de poblaciones con distribución normal tienen una distribución normal con media  $\mu_1 - \mu_2$  y varianzas

$\frac{\sigma_{10}^2}{n_1} + \frac{\sigma_{20}^2}{n_2}$  conocida ( $\sigma_{10}^2$  y  $\sigma_{20}^2$  conocidas). Luego, la estadística de prueba estará dada por:

$$\bar{X} - \bar{Y} \text{ con la estadística } Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_{10}^2}{n_1} + \frac{\sigma_{20}^2}{n_2}}}$$

La formulación de los cuatro casos de pruebas de hipótesis para comparación de medias se presenta en el teorema 4.4.

#### Teorema 4.4

Sean  $X_1, X_2, \dots, X_{n_1}$  y  $Y_1, Y_2, \dots, Y_{n_2}$  dos muestras aleatorias independientes de  $N(\mu_1, \sigma_{10}^2)$  y  $N(\mu_2, \sigma_{20}^2)$ , entonces podemos tener alguno de los siguientes contrastes de hipótesis.

a)  $H_0: \mu_1 - \mu_2 \geq d_0$  contra  $H_1: \mu_1 - \mu_2 < d_0$ , luego la prueba UMP ( $\alpha$ ), para  $\alpha \in (0, 1)$  estará dada por la siguiente regla de decisión para una realización  $x_1, x_2, \dots, x_{n_1}$  y  $y_1, y_2, \dots, y_{n_2}$ :

$$\text{Rechazar } H_0: \mu_1 - \mu_2 \geq d_0, \text{ si } \bar{x} - \bar{y} < d_0 + \Phi^{-1}(\alpha) \sqrt{\frac{\sigma_{10}^2}{n_1} + \frac{\sigma_{20}^2}{n_2}} = d_0 - Z_{\alpha} \sqrt{\frac{\sigma_{10}^2}{n_1} + \frac{\sigma_{20}^2}{n_2}}$$

b)  $H_0: \mu_1 - \mu_2 \leq d_0$  contra  $H_1: \mu_1 - \mu_2 > d_0$ , luego la prueba UMP ( $\alpha$ ), para  $\alpha \in (0, 1)$  estará dada por la siguiente regla de decisión para una realización  $x_1, x_2, \dots, x_{n_1}$  y  $y_1, y_2, \dots, y_{n_2}$ :

$$\text{Rechazar } H_0: \mu_1 - \mu_2 \leq d_0, \text{ si } \bar{x} - \bar{y} > d_0 + \Phi^{-1}(1 - \alpha) \sqrt{\frac{\sigma_{10}^2}{n_1} + \frac{\sigma_{20}^2}{n_2}} = d_0 + Z_{\alpha} \sqrt{\frac{\sigma_{10}^2}{n_1} + \frac{\sigma_{20}^2}{n_2}}$$

c)  $H_0: \mu_1 - \mu_2 = d_0$  contra  $H_1: \mu_1 - \mu_2 \neq d_0$ , luego la prueba UMPI ( $\alpha$ ), para  $\alpha \in (0, 1)$  estará dada por la siguiente regla de decisión para una realización  $x_1, x_2, \dots, x_{n_1}$  y  $y_1, y_2, \dots, y_{n_2}$ :

$$\text{Rechazar } H_0: \mu_1 - \mu_2 = d_0, \text{ si } \bar{x} - \bar{y} < d_0 + \Phi^{-1}\left(\frac{\alpha}{2}\right) \sqrt{\frac{\sigma_{10}^2}{n_1} + \frac{\sigma_{20}^2}{n_2}} = d_0 - Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_{10}^2}{n_1} + \frac{\sigma_{20}^2}{n_2}}$$

$$\bar{x} - \bar{y} > d_0 + \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{\frac{\sigma_{10}^2}{n_1} + \frac{\sigma_{20}^2}{n_2}} = d_0 + Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_{10}^2}{n_1} + \frac{\sigma_{20}^2}{n_2}}$$

d)  $H_0: d_0 \leq \mu_1 - \mu_2 \leq d_1$  contra  $H_1: \mu_1 - \mu_2 < d_0$  o  $\mu_1 - \mu_2 > d_1$ , luego la prueba UMPI ( $\alpha$ ), para  $\alpha \in (0, 1)$  estará dada por la regla de decisión para una realización  $x_1, x_2, \dots, x_{n_1}$  y  $y_1, y_2, \dots, y_{n_2}$ :

$$\text{Rechazar } H_0: d_0 \leq \mu_1 - \mu_2 \leq d_1, \text{ si } \bar{x} - \bar{y} < d_0 + \Phi^{-1}\left(\frac{\alpha}{2}\right) \sqrt{\frac{\sigma_{10}^2}{n_1} + \frac{\sigma_{20}^2}{n_2}} = d_0 - Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_{10}^2}{n_1} + \frac{\sigma_{20}^2}{n_2}}$$

$$\bar{x} - \bar{y} > d_1 + \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{\frac{\sigma_{10}^2}{n_1} + \frac{\sigma_{20}^2}{n_2}} = d_1 + Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_{10}^2}{n_1} + \frac{\sigma_{20}^2}{n_2}}$$

Con  $d_0, d_1 \in \mathbb{R}$ ,  $\sigma_{10}^2$  y  $\sigma_{20}^2$  valores conocidos. En donde,  $\Phi^{-1}(\gamma)$  representa el cuantil  $\gamma$  de la distribución normal estándar  $\gamma \in (0, 1)$ , y  $Z_{\gamma}$  representa el valor de la variable normal estándar cuya área a la derecha es  $\gamma \in (0, 1)$ .

A continuación, se presentan dos ejemplos que muestran aplicaciones del teorema 4.4.

#### Ejemplo 4.19 Aplicaciones del teorema 4.4

Los fabricantes de tornillos tipo A y B aseguran que los primeros tienen en promedio una mayor resistencia a la tensión que los tornillos tipo B en más de 3 kg. Para comprobar su afirmación, los fabricantes prueban

de forma independiente 100 piezas de cada tipo de cuerda bajo condiciones similares, de lo que obtienen los siguientes resultados: el tipo  $A$ , tuvo una resistencia promedio de 88 kg, mientras que el tipo  $B$  una resistencia promedio de 83 kg. Suponga que la resistencia a la tensión de los tornillos tiene una distribución normal con  $X_A \sim N(\mu_A, 25)$  y  $X_B \sim N(\mu_B, 81)$ , realice una prueba estadística para verificar la afirmación de los fabricantes.

- Plantee el contraste de hipótesis apropiado para este problema.
- Justifique a un nivel de significancia de 0.05, si es válida la afirmación de los fabricantes.
- Calcule la potencia de la prueba para  $\mu_A - \mu_B = 4$  kg.

### Solución

- Comparamos medias en donde los fabricantes afirman que la resistencia promedio a la tensión de la cuerda de los dos tipos de tornillos es  $\mu_A > \mu_B + 3$ . Así,  $\mu_A - \mu_B > 3$  será la hipótesis alterna y la contrapuesta  $\mu_A - \mu_B \leq 3$  la hipótesis nula.
- Al seguir los pasos de la metodología para realizar la comprobación y determinar la prueba UMP(0.05), tendremos:
  - $H_0: \mu_A - \mu_B \leq 3$  contra  $H_1: \mu_A - \mu_B > 3$
  - Nivel de significancia  $\alpha = 0.05$ .
  - Estamos ante una situación similar a la del inciso b) del teorema 4.4. Requerimos calcular la CC:

$$d_0 + \Phi^{-1}(1 - \alpha) \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}} \text{ y comparar con el valor de la EP } \bar{x}_A - \bar{x}_B.$$

Si se calcula cada elemento,  $d_0 = 3$ ,  $n_A = n_B = 100$ ,  $\sigma_A^2 = 25$ ,  $\sigma_B^2 = 81$  y  $\alpha = 0.05$  con las tablas porcentuales para la distribución normal estándar,  $\Phi^{-1}(1 - 0.05) = \Phi^{-1}(95\%) = 1.6449$ . Por último, la regla de decisión:

$$\text{Rechazar: } H_0: \mu_A - \mu_B \leq 3, \text{ si } \bar{x}_A - \bar{x}_B > d_0 + \Phi^{-1}(1 - \alpha) \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}} = 3 + 1.6449 \sqrt{\frac{25}{100} + \frac{81}{100}} = 4.69$$

Es decir, rechazar  $H_0: \mu_A - \mu_B \leq 3$  si  $\bar{x}_A - \bar{x}_B > 4.69$ . En la figura 4.16 se muestran las regiones de la prueba.

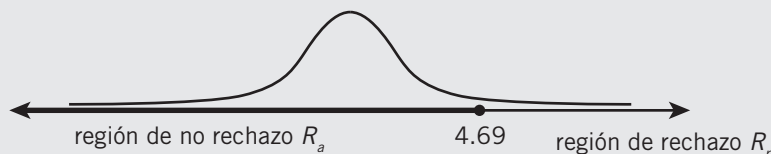


Figura 4.16 Regiones de la prueba del ejemplo 4.19.

- Para aplicar la regla de decisión,  $\bar{x}_A = 88$ ,  $\bar{x}_B = 83$  y  $\bar{x}_A - \bar{x}_B = 88 - 83 = 5 > 4.69$ . Luego, concluimos que  $H_0: \mu_A - \mu_B \leq 3$  se rechaza al 5% de significancia.

**Conclusión:** con 5% de significancia y la realización obtenida no existen evidencias para refutar la afirmación de los fabricantes de que  $\mu_A > \mu_B + 3$ , la cuerda de los tornillos tipo  $A$  tiene una resistencia promedio mayor a la del tipo  $B$  en más de 3 kg.

- Para calcular la potencia de la prueba utilizamos la región de rechazo:

Si la verdadera diferencia de medias fuera 4 kg, la prueba tendría una potencia baja.

$$\begin{aligned}
 P(\text{rechazar } H_0 \mid \mu_A - \mu_B > 3) &= P(\bar{X}_A - \bar{X}_B > 4.69 \mid \mu_A - \mu_B = 4) \\
 &= P\left(\frac{\bar{X}_A - \bar{X}_B - (\mu_A - \mu_B)}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} > \frac{4.69 - (\mu_A - \mu_B)}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} \mid \mu_A - \mu_B = 4\right) \\
 &= P\left(Z > \frac{4.69 - 4}{\sqrt{\frac{25}{100} + \frac{81}{100}}}\right) \\
 &= P(Z > 0.67) = 0.2514
 \end{aligned}$$

#### Ejemplo 4.20 Aplicaciones del teorema 4.4

En un experimento se compararon las economías en combustible de dos tipos de vehículos diésel equipados de manera similar. Se utilizaron de manera independiente 12 automóviles TY y 10 VG en pruebas de velocidad fija de 90 km/h. Si para los TY se obtuvo un promedio de 16 km/L y para los VG el promedio fue de 11 km/L, con los resultados obtenidos la persona que realiza el experimento afirma que los vehículos TY en promedio exceden a los VG entre 1 y 3 km/L. Suponga que el rendimiento por litro para cada modelo de vehículo se distribuye aproximadamente en forma normal con varianzas de 4.41 para TY y 2.25 para VG.

- Plantee el contraste de hipótesis apropiado para este problema.
- Justifique a un nivel de significancia de 0.04, si es válida la afirmación.
- Calcule la potencia de la prueba para una diferencia promedio de 5 km.

#### Solución

- Se pide una prueba de hipótesis para la diferencia de medias, en donde se tiene que comprobar si el rendimiento promedio por litro de los autos TY excede al de los autos VG entre [1, 3] km. Si se representa por  $\mu_1$  al rendimiento promedio por litro de los carros TY y por  $\mu_2$  a los VG, tenemos que la hipótesis nula será  $1 \leq \mu_1 - \mu_2 \leq 3$  y la contrapuesta  $\mu_1 - \mu_2 < 1$  o  $\mu_1 - \mu_2 > 3$  la hipótesis alterna.
- Al seguir la metodología para la comprobación y determinar la prueba UMP(0.04), tendremos:
  - $H_0: 1 \leq \mu_1 - \mu_2 \leq 3$  contra  $H_1: \mu_1 - \mu_2 < 1$  o  $\mu_1 - \mu_2 > 3$ .
  - Nivel de significancia  $\alpha = 0.04$ .
  - Se tiene una situación similar a la del inciso d) del teorema 4.4. Requerimos calcular la CC:

$$d_0 + \Phi^{-1}\left(\frac{\alpha}{2}\right)\sqrt{\frac{\sigma_{10}^2}{n_1} + \frac{\sigma_{20}^2}{n_2}} \text{ y } d_1 + \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\sqrt{\frac{\sigma_{10}^2}{n_1} + \frac{\sigma_{20}^2}{n_2}}; \text{ comparar con la EP } \bar{x}_1 - \bar{x}_2.$$

Se calcula cada elemento,  $d_0 = 1$  y  $d_1 = 3$ ,  $n_1 = 12$  y  $n_2 = 10$ ,  $\sigma_{10}^2 = 4.41$ ,  $\sigma_{20}^2 = 2.25$  y  $\alpha = 0.04$  con las tablas porcentuales para la distribución normal estándar,  $\Phi^{-1}(0.02) = -2.0537$   $\Phi^{-1}(1 - 0.02) = 2.0537$ . Por último, la regla de decisión.

Rechazar:  $H_0: 1 \leq \mu_1 - \mu_2 \leq 3$ , si



$$\bar{x}_1 - \bar{x}_2 < d_0 + \Phi^{-1}\left(\frac{\alpha}{2}\right)\sqrt{\frac{\sigma_{10}^2}{n_1} + \frac{\sigma_{20}^2}{n_2}} = 1 - 2.0537\sqrt{\frac{4.41}{12} + \frac{2.25}{10}} = -0.58, \text{ o}$$

$$\bar{x}_1 - \bar{x}_2 > d_1 + \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\sqrt{\frac{\sigma_{10}^2}{n_1} + \frac{\sigma_{20}^2}{n_2}} = 3 + 2.0537\sqrt{\frac{4.41}{12} + \frac{2.25}{10}} = 4.58$$

Es decir, rechazar  $H_0: 1 \leq \mu_1 - \mu_2 \leq 3$  si  $\bar{x}_1 - \bar{x}_2 < -0.58$  o  $\bar{x}_1 - \bar{x}_2 > 4.58$ .

iv) Por último, aplicamos la regla de decisión, para esto recordamos que  $\bar{x}_1 = 16$  y  $\bar{x}_2 = 11$ , luego  $\bar{x}_1 - \bar{x}_2 = 16 - 11 = 5$  y concluimos rechazar  $H_0: 1 \leq \mu_1 - \mu_2 \leq 3$  al 4% de significancia.

**Conclusión:** con 4% de significancia y la realización obtenida no existen evidencias para aceptar la afirmación de la persona que realiza el experimento de que  $1 \leq \mu_1 - \mu_2 \leq 3$ .

c) Para calcular la potencia de la prueba utilizamos la región de rechazo,

$$\begin{aligned} P(\text{rechazar } H_0 \mid \mu_1 - \mu_2 \notin [1, 3]) &= P((\bar{X}_1 - \bar{X}_2 < -0.58) \cup (\bar{X}_1 - \bar{X}_2 > 4.58) \mid \mu_1 - \mu_2 = 5) \\ &= P(\bar{X}_1 - \bar{X}_2 < -0.58 \mid \mu_1 - \mu_2) + P(\bar{X}_1 - \bar{X}_2 > 4.58 \mid \mu_1 - \mu_2 = 5) \\ &= P\left(\frac{\bar{X}_1 - \bar{X}_2 - 5}{\sqrt{\frac{4.41}{12} + \frac{2.25}{10}}} < \frac{-0.58 - 5}{\sqrt{\frac{4.41}{12} + \frac{2.25}{10}}}\right) + P\left(\frac{\bar{X}_1 - \bar{X}_2 - 5}{\sqrt{\frac{4.41}{12} + \frac{2.25}{10}}} > \frac{4.58 - 5}{\sqrt{\frac{4.41}{12} + \frac{2.25}{10}}}\right) \\ &= P(Z < -7.25) + P(Z > -0.55) \\ &= 0.7088 \end{aligned}$$

Si la verdadera diferencia de medias fuera de 5 km la prueba tendría una potencia considerable.

## Pruebas de hipótesis para la diferencia de medias sobre poblaciones aproximadamente normales cuando se desconocen $\sigma_1^2$ y $\sigma_2^2$ pero $\sigma_1^2 = \sigma_2^2$

El problema de la comparación de medias para el caso de igualdad de varianzas se le conoce como el **problema de Behrens-Fisher**, desarrollado entre 1935 y 1939. En este caso la estadística de prueba  $\bar{X} - \bar{Y}$  tiene una distribución t-Student con  $n_1 + n_2 - 2$  grados de libertad y está dada por:

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_p \sqrt{1/n_1 + 1/n_2}}, \text{ en donde } S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

A continuación, se formula el teorema que dará respuesta a los cuatro casos de pruebas de hipótesis para comparación de medias poblacionales para el problema de Behrens-Fisher.

### Teorema 4.5

Sean  $X_1, X_2, \dots, X_{n_1}$  y  $Y_1, Y_2, \dots, Y_{n_2}$  dos muestras aleatorias independientes de  $N(\mu_1, \sigma_1^2)$  y  $N(\mu_2, \sigma_2^2)$ , entonces podemos tener alguno de los siguientes contraste de hipótesis:

a)  $H_0: \mu_1 - \mu_2 \geq d_0$  contra  $H_1: \mu_1 - \mu_2 < d_0$ , luego la prueba  $UMP(\alpha)$ , para  $\alpha \in (0, 1)$  estará dada por la siguiente regla de decisión para una realización  $x_1, x_2, \dots, x_{n_1}$  y  $y_1, y_2, \dots, y_{n_2}$ :

$$\text{Rechazar } H_0: \mu_1 - \mu_2 \geq d_0, \text{ si } \bar{x} - \bar{y} < d_0 + F_{t(n_1+n_2-2)}^{-1}(\alpha) s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = d_0 - t_{\alpha} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

b)  $H_0: \mu_1 - \mu_2 \leq d_0$  contra  $H_1: \mu_1 - \mu_2 > d_0$ , luego la prueba UMP( $\alpha$ ), para  $\alpha \in (0, 1)$  estará dada por la siguiente regla de decisión para una realización  $x_1, x_2, \dots, x_{n_1}$  y  $y_1, y_2, \dots, y_{n_2}$ :

$$\text{Rechazar } H_0: \mu_1 - \mu_2 \leq d_0, \text{ si } \bar{x} - \bar{y} > d_0 + F_{t(n_1+n_2-2)}^{-1}(1-\alpha) s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = d_0 + t_{\alpha} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

c)  $H_0: \mu_1 - \mu_2 = d_0$  contra  $H_1: \mu_1 - \mu_2 \neq d_0$ , luego la prueba UMPI( $\alpha$ ), para  $\alpha \in (0, 1)$  estará dada por la siguiente regla de decisión para una realización  $x_1, x_2, \dots, x_{n_1}$  y  $y_1, y_2, \dots, y_{n_2}$ :

$$\text{Rechazar } H_0: \mu_1 - \mu_2 = d_0, \text{ si } \bar{x} - \bar{y} < d_0 + F_{t(n_1+n_2-2)}^{-1}\left(\frac{\alpha}{2}\right) s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = d_0 - t_{\frac{\alpha}{2}} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \text{ o}$$

$$\bar{x} - \bar{y} > d_0 + F_{t(n_1+n_2-2)}^{-1}\left(1 - \frac{\alpha}{2}\right) s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = d_0 + t_{\frac{\alpha}{2}} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

d)  $H_0: d_0 \leq \mu_1 - \mu_2 \leq d_1$  contra  $H_1: \mu_1 - \mu_2 \leq d_0$  o  $\mu_1 - \mu_2 \geq d_1$ , luego la prueba UMPI( $\alpha$ ), para  $\alpha \in (0, 1)$  estará dada por la siguiente regla de decisión para una realización  $x_1, x_2, \dots, x_{n_1}$  y  $y_1, y_2, \dots, y_{n_2}$ :

$$\text{Rechazar } H_0: d_0 \leq \mu_1 - \mu_2 \leq d_1, \text{ si } \bar{x} - \bar{y} < d_0 + F_{t(n_1+n_2-2)}^{-1}\left(\frac{\alpha}{2}\right) s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = d_0 - t_{\frac{\alpha}{2}} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \text{ o}$$

$$\bar{x} - \bar{y} > d_1 + F_{t(n_1+n_2-2)}^{-1}\left(1 - \frac{\alpha}{2}\right) s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = d_1 + t_{\frac{\alpha}{2}} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Con  $d_0, d_1 \in \mathbb{R}$  valores conocidos de antemano. En donde,  $F_{t(\nu)}^{-1}(\gamma)$  representa el cuantil  $\gamma$  de la distribución t-Student con  $\nu$  grados de libertad para  $\gamma \in (0, 1)$  y  $t_{\gamma}$ , el valor de la variable t-Student con  $\nu = n_1 + n_2 - 2$  grados de libertad, cuya área a la derecha es  $\gamma \in (0, 1)$ .

Los siguientes dos ejemplos muestran aplicaciones al problema de Behrens-Fisher.

#### Ejemplo 4.21 Problema de Behrens-Fisher

Se comparan dos tipos de rosca de tornillo para ver su resistencia a la tensión. Se prueban de manera independiente 12 piezas de cada tipo de cuerda bajo condiciones similares, con lo que se obtienen los resultados en kilogramos que se muestran en la tabla 4.5.

Tabla 4.5

Tipo de rosca	1	2	3	4	5	6	7	8	9	10	11	12
I	78	76	80	79	78	80	82	81	79	83	80	82
II	83	80	82	83	81	80	79	80	82	78	79	81

Suponga que la resistencia a la tensión de los tornillos tiene una distribución normal con varianzas desconocidas, pero iguales. Se desea saber si es posible concluir de manera estadística que la resistencia promedio a la tensión de los tornillos tipo I es menor que la de los tornillos tipo II.

- Plantee el contraste de hipótesis apropiado para este problema.
- Justifique a un nivel de significancia de 0.025 si la conclusión es válida.
- Calcule la potencia de la prueba para  $\mu_1 - \mu_2 = -2$  kg.

**Solución**

a) En este momento, comparamos medias en donde los fabricantes afirman que la resistencia promedio a la tensión de la cuerda de los dos tipos de tornillos es  $\mu_1 < \mu_2$ . Así,  $\mu_1 - \mu_2 < 0$  será la hipótesis alterna y la contrapuesta  $\mu_1 - \mu_2 \geq 0$  la hipótesis nula.

b) Si se sigue la metodología para la comprobación y se determina la prueba UMP(0.05):

i)  $H_0: \mu_1 - \mu_2 \geq 0$  contra  $H_1: \mu_1 - \mu_2 < 0$

ii) Nivel de significancia  $\alpha = 0.05$ .

iii) Estamos ante una situación similar a la del inciso a) del teorema 4.5. Por tanto, requerimos calcular la

$$\text{CC: } d_0 + F_{t(n_1+n_2-2)}^{-1}(\alpha) s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \text{ y comparar con la EP } \bar{x}_2 - \bar{x}_1.$$

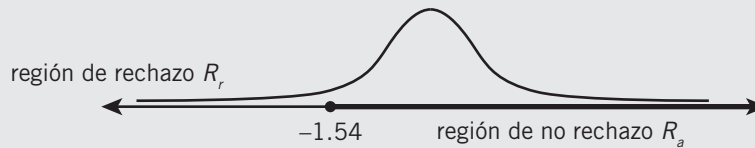
Si se calcula cada componente:  $d_0 = 0$  y  $\bar{x}_1 = 79.8333$ ,  $s_1^2 = 3.9697$ ;  $\bar{x}_2 = 80.6667$ ,  $s_2^2 = 2.6061$  y  $\alpha = 0.05$  con las tablas porcentuales para la distribución t-Student y  $n_1 + n_2 - 2 = 12 + 12 - 2 = 22$  grados de libertad,  $F_{t(22)}^{-1}(0.05) = -2.074$ . Por otro lado,

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(12 - 1)3.9697 + (12 - 1)2.6061}{12 + 12 - 2}} = 1.8133$$

Por último, la regla de decisión:

$$\text{Rechazar } H_0: \mu_1 - \mu_2 \geq 0, \text{ si } \bar{x}_1 - \bar{x}_2 < d_0 + F_{t(n_1+n_2-2)}^{-1}(\alpha) s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 0 - 2.074(1.8133) \sqrt{\frac{1}{12} + \frac{1}{12}} = -1.54$$

Es decir, rechazar  $H_0: \mu_1 - \mu_2 \geq 0$  si  $\bar{x}_1 - \bar{x}_2 < -1.54$ . La figura 4.17 muestra las regiones de la partición de la prueba.



**Figura 4.17** Regiones de la prueba del ejemplo 4.21.

iv) Al final, aplicamos la regla de decisión, para esto recordamos que  $\bar{x}_1 = 79.8333$  y  $\bar{x}_2 = 80.6667$ , luego  $\bar{x}_1 - \bar{x}_2 = 79.83 - 80.67 = -0.84$ , concluimos que  $H_0: \mu_1 - \mu_2 \geq 0$  no se rechaza al 2.5% de significancia.

**Conclusión:** el valor se encuentra en la región de no rechazo, por tal razón a partir de la realización dada no hay evidencias para rechazar  $H_0: \mu_1 - \mu_2 \geq 0$  a un nivel de significancia de 2.5%. Luego, la resistencia promedio a la tensión de los tornillos del tipo I no es menor a la de los tornillos tipo II.

c) Para calcular la potencia de la prueba utilizamos la región de rechazo:

$$\begin{aligned} P(\text{rechazar } H_0 \mid \mu_1 - \mu_2 < 0) &= P(\bar{X}_1 - \bar{X}_2 < -1.54 \mid \mu_1 - \mu_2 = -2) \\ &= P\left(\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < \frac{-1.54 - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \mid \mu_1 - \mu_2 = -2\right) = P\left(T_{22} < \frac{-1.54 - (-2)}{1.8133 \sqrt{\frac{1}{12} + \frac{1}{12}}}\right) \\ &= P(T_{22} < 0.621) \approx 0.7295 \end{aligned}$$

Si la verdadera diferencia de medias fuera  $-2$  kg la prueba tendría una potencia considerable.

**Ejemplo 4.22 Problema de Behrens-Fisher**

Un fabricante de soldadura creó un recubrimiento que afirma incrementa la resistencia a la tracción en más de 5 lb. Para probar su afirmación de manera estadística lleva a cabo pruebas de tracción en 10 puntos de soldadura en un dispositivo semiconductor (sin el recubrimiento), lo que produjo los siguientes resultados en libras requeridas para romper la soldadura:

15.8, 12.7, 13.2, 16.9, 10.6, 18.8, 11.1, 14.3, 17.0, 12.5

Después, tomó otro conjunto independiente del primero, de ocho puntos, que fueron probados después de recibir el recubrimiento para la resistencia a la tracción y se obtuvieron los siguientes resultados:

24.9, 23.6, 19.8, 22.1, 20.4, 21.6, 21.8, 22.5

Suponga normalidad con varianzas iguales en las pruebas de tracción, pruebe de manera estadística si es válida la afirmación del fabricante.

- Plantee el contraste de hipótesis apropiado para este problema.
- Justifique a un nivel de significancia de 0.05 si es válida la afirmación.
- Calcule la potencia de la prueba para  $\mu_1 - \mu_2 = 8$  lb, donde  $\mu_1$  representa la media en libras requeridas para romper la soldadura sin el recubrimiento, de manera similar  $\mu_2$ , pero con recubrimiento.

**Solución**

a) Primero, comparamos medias en donde el fabricante afirma que la resistencia promedio a la ruptura de la soldadura es  $\mu_2 > \mu_1 + 5$ . Así,  $\mu_2 - \mu_1 > 5$  será la hipótesis alterna y la contrapuesta  $\mu_2 - \mu_1 \leq 5$  la hipótesis nula.

b) Si seguimos la metodología para la comprobación y determinamos la prueba UMP(0.05), tenemos:

i)  $H_0: \mu_2 - \mu_1 \leq 5$  contra  $H_1: \mu_2 - \mu_1 > 5$

ii) Nivel de significancia  $\alpha = 0.05$ .

iii) Estamos ante una situación similar a la del inciso b) del teorema 4.5. Requerimos calcular la CC:

$$d_0 + F_{t(n_1+n_2-2)}^{-1}(1-\alpha)s_p\sqrt{1/n_1+1/n_2} \text{ y comparar con el valor de la EP } \bar{x}_2 - \bar{x}_1.$$

Si se calcula cada componente:  $d_0 = 5$ ,  $\bar{x}_1 = 14.29$ ,  $s_1^2 = 7.50$  y  $n_1 = 10$ ;  $\bar{x}_2 = 22.09$ ,  $s_2^2 = 2.68$ ,  $n_2 = 8$  y  $\alpha = 0.05$  con las tablas porcentuales para la distribución t-Student y  $n_1 + n_2 - 2 = 10 + 8 - 2 = 16$  grados de libertad,  $F_{t(16)}^{-1}(0.95) = 1.746$ . Por otro lado,

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} = \sqrt{\frac{(10-1)7.50 + (8-1)2.68}{10+8-2}} = 2.32$$

Por último, la regla de decisión:

$$\text{Rechazar: } H_0: \mu_2 - \mu_1 \leq 5, \text{ si } \bar{x}_2 - \bar{x}_1 > d_0 + F_{t(16)}^{-1}(1-\alpha)s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 5 + 1.746(2.32)\sqrt{\frac{1}{10} + \frac{1}{8}} = 6.92$$

Es decir, rechazar  $H_0: \mu_2 - \mu_1 \leq 5$  si  $\bar{x}_2 - \bar{x}_1 > 6.92$ . En la figura 4.18 se pueden apreciar las regiones de la partición para la prueba.



**Figura 4.18** Regiones de la prueba del ejemplo 4.22.

iv) Por último, aplicamos la regla de decisión, para esto recordamos que  $\bar{x}_1 = 14.29$  y  $\bar{x}_2 = 22.09$ ,  $\bar{x}_2 - \bar{x}_1 = 22.09 - 14.29 = 7.8$ . Así, concluimos que  $H_0: \mu_2 - \mu_1 \leq 5$  se rechaza 5% de significancia.

**Conclusión:** el recubrimiento sí aumenta la resistencia a la ruptura de la soldadura en más de 5 lb.

c) Para calcular la potencia de la prueba utilizamos la región de rechazo:

$$\begin{aligned} P(\text{rechazar } H_0 \mid \mu_2 - \mu_1 > 5) &= P(\bar{X}_2 - \bar{X}_1 > 6.92 \mid \mu_2 - \mu_1 = 8) = P\left(\frac{\bar{X}_2 - \bar{X}_1 - 8}{2.32\sqrt{\frac{1}{10} + \frac{1}{8}}} > \frac{6.92 - 8}{2.32\sqrt{\frac{1}{10} + \frac{1}{8}}}\right) \\ &= P(T_{16} > -0.9814) = 0.8295 \end{aligned}$$

Si la verdadera diferencia de medias fuera 8 lb la prueba tendría una potencia elevada.

## Pruebas de hipótesis para la diferencia de medias sobre poblaciones aproximadamente normales cuando se desconocen $\sigma_1^2$ y $\sigma_2^2$ pero $\sigma_1^2 \neq \sigma_2^2$

Cuando las distribuciones de la población son normales con varianzas desconocidas y diferentes, Esther Welch (1937) y Satterthwaite (1946) encontraron pruebas más potentes que la del teorema 4.5, pero en general la prueba de Welch es más recomendable y será la que explicaremos en esta sección, aunque aclaramos que en esta situación no tenemos la prueba UMP( $\alpha$ ). La prueba de Welch no es robusta en ausencia de la condición de normalidad (hecho demostrado por Yuen en 1974 y Cressie & Whitford, en 1986).

Veamos la formulación de la prueba de Welch-Aspin (Aspin, 1948). Sea  $X_1, X_2, X_3, \dots, X_{n_1}$  una muestra aleatoria de la población 1 que tiene una distribución normal con media  $\mu_1$  y varianza  $\sigma_1^2$  desconocida y  $Y_1, Y_2, \dots, Y_{n_2}$  una muestra aleatoria de la población 2 con una distribución normal con media  $\mu_2$  y varianza  $\sigma_2^2$  también desconocida. Además, suponga que las dos muestras son independientes y  $\sigma_1^2 \neq \sigma_2^2$ . Se desea contrastar alguno de los juegos de hipótesis dados en el teorema 4.5. En estas condiciones Welch-Aspin obtuvieron la estadística de prueba,  $\bar{X} - \bar{Y}$ , en una distribución t-Student con  $\nu$  grados de libertad, dados por:

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}, \text{ con grados de } \nu = \frac{(S_1^2/n_1 + S_2^2/n_2)^2}{\left(\frac{S_1^2}{n_1}\right)^2 \frac{1}{n_1 - 1} + \left(\frac{S_2^2}{n_2}\right)^2 \frac{1}{n_2 - 1}}$$

Los grados de libertad,  $\nu$ , son redondeados al entero más próximo. Es decir, si  $\nu = 17.3 \approx 17$ ,  $\nu = 17.8 \approx 18$  o  $\nu = 17.5 \approx 18$ . Además,  $(\bar{X}, S_1^2)$  y  $(\bar{Y}, S_2^2)$  son la media y varianza insesgadas de las muestras 1 y 2, respectivamente.

### Teorema 4.6

Sean  $X_1, X_2, \dots, X_{n_1}$  y  $Y_1, Y_2, \dots, Y_{n_2}$  dos muestras aleatorias independientes de  $N(\mu_1, \sigma_1^2)$  y  $N(\mu_2, \sigma_2^2)$ , entonces podemos tener alguno de los siguientes contrastes de hipótesis.

a)  $H_0: \mu_1 - \mu_2 \geq d_0$  contra  $H_1: \mu_1 - \mu_2 < d_0$ , luego la prueba de tamaño  $\alpha$ , para  $\alpha \in (0, 1)$  estará dada por la siguiente regla de decisión para una realización  $x_1, x_2, \dots, x_{n_1}$  y  $y_1, y_2, \dots, y_{n_2}$ :

$$\text{Rechazar } H_0: \mu_1 - \mu_2 \geq d_0, \text{ si } \bar{x} - \bar{y} < d_0 + F_{t(\nu)}^{-1}(\alpha) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = d_0 - t_{\alpha} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

b)  $H_0: \mu_1 - \mu_2 \leq d_0$  contra  $H_1: \mu_1 - \mu_2 > d_0$ , entonces la prueba de tamaño  $\alpha$ , para  $\alpha \in (0, 1)$  estará dada por la siguiente regla de decisión para una realización  $x_1, x_2, \dots, x_{n_1}$  y  $y_1, y_2, \dots, y_{n_2}$ :

$$\text{Rechazar } H_0: \mu_1 - \mu_2 \leq d_0, \text{ si } \bar{x} - \bar{y} > d_0 + F_{t(\nu)}^{-1}(1 - \alpha) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = d_0 + t_{\alpha} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

c)  $H_0: \mu_1 - \mu_2 = d_0$  contra  $H_1: \mu_1 - \mu_2 \neq d_0$ , entonces la prueba de tamaño  $\alpha$ , para  $\alpha \in (0, 1)$  estará dada por la siguiente regla de decisión para una realización  $x_1, x_2, \dots, x_{n_1}$  y  $y_1, y_2, \dots, y_{n_2}$ :

$$\text{Rechazar } H_0: \mu_1 - \mu_2 = d_0, \text{ si } \bar{x} - \bar{y} < d_0 + F_{t(\nu)}^{-1}\left(\frac{\alpha}{2}\right) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = d_0 - t_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$\bar{x} - \bar{y} > d_0 + F_{t(\nu)}^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = d_0 + t_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

d)  $H_0: d_0 \leq \mu_1 - \mu_2 \leq d_1$  contra  $H_1: \mu_1 - \mu_2 \leq d_0$  o  $\mu_1 - \mu_2 \geq d_1$ , entonces la prueba de tamaño  $\alpha$ , para  $\alpha \in (0, 1)$  estará dada por la siguiente regla de decisión para una realización  $x_1, x_2, \dots, x_{n_1}$  y  $y_1, y_2, \dots, y_{n_2}$ :

$$\text{Rechazar } H_0: d_0 \leq \mu_1 - \mu_2 \leq d_1, \text{ si } \bar{x} - \bar{y} < d_0 + F_{t(\nu)}^{-1}\left(\frac{\alpha}{2}\right) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = d_0 - t_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$\bar{x} - \bar{y} > d_1 + F_{t(\nu)}^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = d_1 + t_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Con  $d_0, d_1 \in \mathbb{R}$  valores conocidos de antemano. En donde,  $F_{t(\nu)}^{-1}(\gamma)$  representa el cuantil  $\gamma$  de la distribución t-Student con  $\nu$  grados de libertad para  $\gamma \in (0, 1)$  y  $t_{\gamma}$ , el valor de la variable t-Student con  $\nu$  grados de libertad cuya área a la derecha es  $\gamma \in (0, 1)$ .

En los dos ejemplos siguientes se muestran aplicaciones de la prueba de Welch-Aspin.

#### Ejemplo 4.23 Prueba Welch-Aspin

En un experimento se compararon las economías en combustible de dos tipos de vehículos diésel equipados de manera similar. Se utilizaron 12 automóviles TY y 10 VG en pruebas de velocidad fija de 90 km/h. Si para los autos TY se obtuvo un promedio de 16 km/L con una desviación estándar de 1.0 km/L, mientras que para los autos VG los resultados fueron de 11 km/L, con una desviación estándar de 1.8 km/L. Con los valores obtenidos la persona que realiza el experimento afirma que los vehículos TY en promedio exceden a los VG en 4 km/L. Suponga que el rendimiento por litro para cada modelo de vehículo se distribuye aproximadamente en forma normal con varianzas diferentes.

- Plantee el contraste de hipótesis apropiado para este problema.
- Justifique a un nivel de significancia de 0.10, si es válida la afirmación.
- Calcule la potencia de la prueba para una diferencia promedio de 5 kilómetros por litro.

#### Solución

- Se pide una prueba de hipótesis para la diferencia de medias, en donde se tiene que probar que el rendimiento promedio por litro de los autos TY excede al rendimiento de los autos VG en 4 km/L. Representando por  $\mu_1$  al rendimiento promedio por litro de los autos TY y por  $\mu_2$  a los VG,  $H_0: \mu_1 - \mu_2 = 4$  contra  $H_1: \mu_1 - \mu_2 \neq 4$ .
- Si sigue los pasos de la metodología para realizar la comprobación y determinar la prueba de tamaño 0.10. Utilizaremos la prueba de Welch-Aspin.

- i)  $H_0: \mu_1 - \mu_2 = 4$  contra  $H_1: \mu_1 - \mu_2 \neq 4$   
 ii) Nivel de significancia  $\alpha = 0.10$ .  
 iii) Estamos ante una situación similar a la del inciso c) del teorema 4.6. Requerimos calcular la CC dada por  $d_0 + F_{t(\nu)}^{-1}\left(\frac{\alpha}{2}\right)\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$  o  $d_0 + F_{t(\nu)}^{-1}\left(1 - \frac{\alpha}{2}\right)\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$  y comparar con el valor de la EP  $\bar{x}_2 - \bar{x}_1$ .

Si se calcula cada componente,  $d_0 = 4$ ,  $\bar{x}_1 = 16$ ,  $s_1^2 = 1$ ,  $n_1 = 12$ ,  $\bar{x}_2 = 11$ ,  $s_2^2 = 3.24$  y  $n_2 = 10$ , entonces los grados de libertad se obtienen con:

$$\nu = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]^2}{\left(\frac{s_1^2}{n_1}\right)^2 \left(\frac{1}{n_1 - 1}\right) + \left(\frac{s_2^2}{n_2}\right)^2 \left(\frac{1}{n_2 - 1}\right)} = \frac{\left[\frac{1}{12} + \frac{3.24}{10}\right]^2}{\left(\frac{1}{12}\right)^2 \left(\frac{1}{12 - 1}\right) + \left(\frac{3.24}{10}\right)^2 \left(\frac{1}{10 - 1}\right)} = 13.4946 \approx 13$$

Así, de las tablas porcentuales para la distribución t-Student con  $\nu = 13$  grados de libertad, y  $\alpha/2 = 0.05$ , resulta  $F_{t(13)}^{-1}(0.05) = -1.771$  y  $F_{t(13)}^{-1}(0.95) = 1.771$ . Por último, la regla de decisión:

$$\text{Rechazar } H_0: \mu_1 - \mu_2 = 4, \text{ si } \bar{x}_1 - \bar{x}_2 < d_0 + F_{t(\nu)}^{-1}\left(\frac{\alpha}{2}\right)\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 4 - 1.771\sqrt{\frac{1}{12} + \frac{3.24}{10}} = 2.87 \text{ o}$$

$$\bar{x}_1 - \bar{x}_2 > d_0 + F_{t(\nu)}^{-1}\left(1 - \frac{\alpha}{2}\right)\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 4 + 1.771\sqrt{\frac{1}{12} + \frac{3.24}{10}} = 5.13$$

Es decir, rechazar  $H_0: \mu_1 - \mu_2 = 4$  si  $\bar{x}_1 - \bar{x}_2 < 2.87$  o  $\bar{x}_1 - \bar{x}_2 > 5.13$ . En la figura 4.19 se muestran las regiones de la prueba.



Figura 4.19 Regiones de la prueba del ejemplo 4.23.

- iv) Por último, aplicamos la regla de decisión, para esto recordamos que  $\bar{x}_1 = 16$  y  $\bar{x}_2 = 11$ , entonces  $\bar{x}_1 - \bar{x}_2 = 16 - 11 = 5$  y concluimos que  $H_0: \mu_1 - \mu_2 = 4$  no se rechaza 10% de significancia.

**Conclusión:** el valor se encuentra en la región de no rechazo, entonces a partir de la realización dada no hay evidencias para rechazar  $H_0: \mu_1 - \mu_2 = 4$  a un nivel de significancia de 10%. Luego, el rendimiento promedio de los autos TY es superior al rendimiento promedio de los autos VG en 4 km/L.

- c) Para calcular la potencia de la prueba utilizamos la región de rechazo:

$$\begin{aligned} P(\text{rechazar } H_0 \mid \mu_1 - \mu_2 \neq 4) &= P((\bar{X}_1 - \bar{X}_2 < 2.87) \cup (\bar{X}_1 - \bar{X}_2 > 5.13) \mid \mu_1 - \mu_2 = 5) \\ &= P(\bar{X}_1 - \bar{X}_2 < 2.87 \mid \mu_1 - \mu_2 = 5) + P(\bar{X}_1 - \bar{X}_2 > 5.13 \mid \mu_1 - \mu_2 = 5) \\ &= P\left(\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} < \frac{2.87 - 5}{\sqrt{\frac{1}{12} + \frac{3.24}{10}}}\right) + P\left(\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} > \frac{5.13 - 5}{\sqrt{\frac{1}{12} + \frac{3.24}{10}}}\right) \end{aligned}$$



$$\begin{aligned}
 &= P(T_{13} < -3.34) + P(T_{13} > 0.20) \approx 0.0026 + 0.4223 \\
 &= 0.4249
 \end{aligned}$$

Si la verdadera diferencia de medias fuera 5 km/L, la prueba tendría una potencia media.

#### Ejemplo 4.24 Prueba Welch-Aspin

Resuelva el ejemplo del fabricante de soldadura que creó un recubrimiento que afirma incrementa la resistencia a la tracción en más de 5 lb. Ahora suponga que las varianzas poblacionales son diferentes y compare la respuesta con la obtenida en el ejemplo 4.22. En estas condiciones el inciso a) no se altera, es decir,  $\mu_2 - \mu_1 > 5$  será la hipótesis alterna y la contrapuesta  $\mu_2 - \mu_1 \leq 5$  la hipótesis nula.

#### Solución

a) Si se siguen los pasos de la metodología para realizar la comprobación y se determina la prueba de tamaño 0.05, tenemos:

i)  $H_0: \mu_2 - \mu_1 \leq 5$  contra  $H_1: \mu_2 - \mu_1 > 5$

ii) Nivel de significancia  $\alpha = 0.05$ .

iii) Estamos ante una situación similar a la del inciso b) del teorema 4.6. Luego, requerimos calcular la

$$\text{CC: } d_0 + F_{t(\nu)}^{-1}(1 - \alpha) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \text{ y compararla con el valor de la EP } \bar{x}_2 - \bar{x}_1.$$

Si calcula cada elemento,  $d_0 = 5$ ,  $\bar{x}_1 = 14.29$ ,  $s_1^2 = 7.50$ ,  $n_1 = 10$ ,  $\bar{x}_2 = 22.09$ ,  $s_2^2 = 2.68$  y  $n_2 = 8$  los g.l.

$$\nu = \frac{\left[ \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right]^2}{\left( \frac{s_1^2}{n_1} \right)^2 \left( \frac{1}{n_1 - 1} \right) + \left( \frac{s_2^2}{n_2} \right)^2 \left( \frac{1}{n_2 - 1} \right)} = \frac{\left[ \frac{7.50}{10} + \frac{2.68}{8} \right]^2}{\left( \frac{7.50}{10} \right)^2 \left( \frac{1}{10 - 1} \right) + \left( \frac{2.68}{8} \right)^2 \left( \frac{1}{8 - 1} \right)} = 14.99 \approx 15$$

Así, de las tablas porcentuales para la distribución t-Student con  $\nu = 15$  grados de libertad, y  $\alpha = 0.05$ , resulta  $F_{t(15)}^{-1}(0.95) = 1.753$ . Por último, la regla de decisión.

$$\text{Rechazar: } H_0: \mu_2 - \mu_1 \leq 5, \text{ si } \bar{x}_2 - \bar{x}_1 > d_0 + F_{t(15)}^{-1}(1 - \alpha) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 5 + 1.753 \sqrt{\frac{7.5}{10} + \frac{2.68}{8}} = 6.83$$

Es decir, rechazar  $H_0: \mu_2 - \mu_1 \leq 5$  si  $\bar{x}_2 - \bar{x}_1 > 6.83$ . En la figura 4.20 se pueden apreciar las regiones de la prueba.

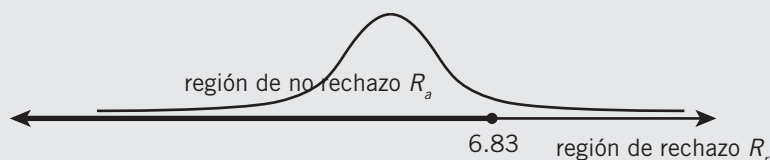


Figura 4.20 Regiones de la prueba del ejemplo 4.24.

iv) Por último, aplicamos la regla de decisión, para esto recordamos que  $\bar{x}_1 = 14.29$  y  $\bar{x}_2 = 22.09$ ,  $\bar{x}_2 - \bar{x}_1 = 22.09 - 14.29 = 7.8$ . Concluimos que  $H_0: \mu_2 - \mu_1 \leq 5$  se rechaza al 5% de significancia.

**Conclusión:** el recubrimiento si aumenta la resistencia a la ruptura de la soldadura en más de 5 lb.

b) Para calcular la potencia de la prueba utilizamos la región de rechazo:

$$\begin{aligned} P(\text{rechazar } H_0 \mid \mu_2 - \mu_1 > 5) &= P(\bar{X}_2 - \bar{X}_1 > 6.83 \mid \mu_2 - \mu_1 = 8) = P\left(\frac{\bar{X}_2 - \bar{X}_1 - 8}{\sqrt{\frac{7.5}{10} + \frac{2.68}{8}}} > \frac{6.83 - 8}{\sqrt{\frac{7.5}{10} + \frac{2.68}{8}}}\right) \\ &= P(T_{15} > -1.123) = 0.8605 \end{aligned}$$

Si la verdadera diferencia de medias fuera 8 lb la prueba tendría una potencia elevada.

Al compararla con la respuesta del ejemplo 4.22 vemos que la potencia aumentó un poco, por tal razón es lógico pensar que en este ejemplo la mejor suposición entre varianzas iguales (véase ejemplo 4.22) y la actual, varianzas diferentes, es más fuerte la última, hecho que confirmaremos en la sección de razón entre varianzas.

## Pruebas de hipótesis para la diferencia de medias de observaciones pareadas con diferencias normales

En la sección sobre intervalos de confianza de la unidad 3 tratamos con detenimiento la diferencia de medias de muestras dependientes con observaciones pareadas, por tanto, en esta parte de la prueba de hipótesis obviaremos esta discusión.

Sean  $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$  las parejas de variables aleatorias  $X$  y  $Y$  con  $\mu_X, \mu_Y$  y  $\sigma_X^2, \sigma_Y^2$ , respectivamente, denotemos por  $D$  a la variable aleatoria de la diferencia entre las variables  $X$  y  $Y$ , de manera que las  $D_i = X_i - Y_i$ ,  $i = 1, 2, \dots, n$ , representan la variable aleatoria resultante de la diferencia entre las variables  $X_i$  y  $Y_i$ . Suponga que las  $D_i$  tienen distribución normal con media  $\mu_D$  y varianza  $\sigma_D^2$  desconocida y son independientes (es decir, las variables aleatorias entre parejas diferentes son independientes, pero las variables dentro del mismo par son dependientes), se obtuvo  $\mu_D = E(X - Y) = \mu_X - \mu_Y$  y  $\sigma_D^2 = V(X - Y) = \sigma_X^2 + \sigma_Y^2 - \text{cov}(X, Y)$ . Parámetros que se estimaron con una realización  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  de las parejas  $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ , que denotamos por  $d_i = x_i - y_i$ , se tiene:

$$\bar{x}_d = \frac{1}{n} \sum_{i=1}^n d_i \text{ que estimará a } \mu_D \text{ y } s_d^2 = \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{x}_d)^2 \text{ estimará a } \sigma_D^2$$

Así, la estadística de prueba  $\frac{\bar{X}_d - \mu_d}{S_d / \sqrt{n}}$  tiene distribución t-Student con  $\nu = n - 1$  grados de libertad.

### Teorema 4.7

Si  $\bar{x}_d$  y  $s_d$  son la media y la desviación estándar muestrales de la diferencia de  $n$  pares independientes de realizaciones de muestras aleatorias pareadas, tomadas de mediciones de las que se desconoce  $\sigma_X^2$  y  $\sigma_Y^2$ , entonces podemos tener alguno de los siguientes contrastes de hipótesis.

a)  $H_0: \mu_d \geq \mu_0$  contra  $H_1: \mu_d < \mu_0$ , luego la prueba UMP( $\alpha$ ), para  $\alpha \in (0, 1)$  estará dada por la siguiente regla de decisión para una realización  $x_1, x_2, \dots, x_n$  y  $y_1, y_2, \dots, y_n$ :

$$\text{Rechazar } H_0: \mu_d \geq \mu_0, \text{ si } \bar{x}_d < \mu_0 + \frac{s_d}{\sqrt{n}} F_{t_{n-1}}^{-1}(\alpha) \text{ o } \bar{x}_d < \mu_0 - \frac{s_d}{\sqrt{n}} t_{\alpha}(n-1)$$

b)  $H_0: \mu_d \leq \mu_0$  contra  $H_1: \mu_d > \mu_0$ , luego la prueba UMP( $\alpha$ ), para  $\alpha \in (0, 1)$  estará dada por la siguiente regla de decisión para una realización  $x_1, x_2, \dots, x_n$  y  $y_1, y_2, \dots, y_n$ :

Rechazar  $H_0: \mu_d \leq \mu_0$ , si  $\bar{x}_d > \mu_0 + \frac{S_d}{\sqrt{n}} F_{t_{n-1}}^{-1}(1 - \alpha)$  o  $\bar{x}_d > \mu_0 + \frac{S_d}{\sqrt{n}} t_{\alpha}(n - 1)$

c)  $H_0: \mu_d = \mu_0$  contra  $H_1: \mu_d \neq \mu_0$ , luego la prueba UMPI( $\alpha$ ), para  $\alpha \in (0, 1)$  estará dada por la siguiente regla de decisión para una realización  $x_1, x_2, \dots, x_n$  y  $y_1, y_2, \dots, y_n$ :

Rechazar  $H_0: \mu_d = \mu_0$ , si  $\bar{x}_d < \mu_0 + \frac{S_d}{\sqrt{n}} F_{t_{n-1}}^{-1}(\alpha/2) = \mu_0 - \frac{S_d}{\sqrt{n}} t_{\alpha/2}(n - 1)$  o

$\bar{x}_d > \mu_0 + \frac{S_d}{\sqrt{n}} F_{t_{n-1}}^{-1}(1 - \alpha/2) = \mu_0 + \frac{S_d}{\sqrt{n}} t_{\alpha/2}(n - 1)$

d)  $H_0: \mu_0 \leq \mu_d \leq \mu_1$  contra  $H_1: \mu_d < \mu_0$  o  $\mu_d > \mu_1$ , luego la prueba UMPI( $\alpha$ ), para  $\alpha \in (0, 1)$  estará dada por la siguiente regla de decisión para una realización  $x_1, x_2, \dots, x_n$  y  $y_1, y_2, \dots, y_n$ :

Rechazar  $H_0: \mu_0 \leq \mu_d \leq \mu_1$ , si  $\bar{x}_d < \mu_0 + \frac{S_d}{\sqrt{n}} F_{t_{n-1}}^{-1}(\alpha/2) = \mu_0 - \frac{S_d}{\sqrt{n}} t_{\alpha/2}(n - 1)$  o

$\bar{x}_d > \mu_1 + \frac{S_d}{\sqrt{n}} F_{t_{n-1}}^{-1}(1 - \alpha/2) = \mu_1 + \frac{S_d}{\sqrt{n}} t_{\alpha/2}(n - 1)$

Con  $\mu_0, \mu_1 \in \mathbb{R}$  valores conocidos, donde,  $F_{t_{n-1}}^{-1}(\gamma)$  representa el cuantil  $\gamma$  de la distribución t-Student con  $n - 1$  grados de libertad para  $\gamma \in (0, 1)$ ,  $t_\gamma$  representa el valor de la variable t-Student con  $n - 1$  grados de libertad cuya área derecha es  $\gamma \in (0, 1)$ .

En los siguientes dos ejemplos se muestran aplicaciones del teorema 4.7.

#### Ejemplo 4.25 Aplicaciones del teorema 4.7

Un veterinario realizó un experimento con 10 animales que fueron sometidos a condiciones que simulaban una enfermedad. El veterinario registró el número de latidos del corazón, antes y después del experimento de lo que obtuvo los datos que se aprecian en la tabla 4.6.

Tabla 4.6

Antes	70	120	98	110	105	100	110	96	69	86
Después	105	130	112	120	138	118	124	118	92	104

El veterinario afirma que la condición experimental aumentó el número de latidos del corazón en más de 15. Suponga normalidad en la diferencia del número de latidos del corazón antes y después del experimento, se pretende dar una respuesta estadística a la afirmación del veterinario.

- Plantee el contraste de hipótesis apropiado para este problema.
- Justifique a un nivel de significancia de 0.05 si es válida la afirmación del veterinario.
- Calcule la potencia de la prueba para una diferencia promedio de 20 latidos.

#### Solución

- Note que entre las unidades experimentales hay dependencia debido a que cada pareja de observaciones fue tomada del mismo animal. Luego, estamos en una situación de muestras pareadas y queremos probar si el número de latidos del corazón aumenta después del experimento. A los latidos después del experimento se le restan los latidos previos, con lo que se debe probar que la media de las diferencias es mayor a 15. Esto último lo formulamos como la hipótesis alterna,  $\mu_d > 15$ , donde  $\mu_d$  representa la verdadera media de las diferencias. Así,  $H_0: \mu_d \leq 15$  y  $H_1: \mu_d > 15$ .

b) Para justificar si es válida la afirmación del veterinario, emplearemos la metodología del teorema 4.7, para muestras pareadas.

i)  $H_0: \mu_d \leq 15$  contra  $H_1: \mu_d > 15$

ii) Nivel de significancia  $\alpha = 0.05$ .

iii) Estamos ante una situación como la del inciso b) del teorema 4.7. Requerimos calcular la CC:

$$\mu_0 + \frac{s_d}{\sqrt{n}} F_{t_{n-1}}^{-1}(1 - \alpha) \text{ y compararla con el valor de la EP } \bar{x}_d.$$

Primero necesitamos calcular las diferencias por parejas.

Tabla 4.7

Después	105	130	112	120	138	118	124	118	92	104
Antes	70	120	98	110	105	100	110	96	69	86
Diferencia	35	10	14	10	33	18	14	22	23	18

Luego,  $d_0 = 15$ ,  $\bar{x}_d = 19.7$  y  $s_d = 8.7311$  y  $\alpha = 0.05$  con las tablas porcentuales para la distribución t-Student y  $n = 1 = 10 - 1 = 9$  grados de libertad,  $F_{t_9}^{-1}(1 - 0.05) = 1.833$ . Por último, la regla de decisión:

$$\text{Rechazar: } H_0: \mu_d \leq 15, \text{ si } \bar{x}_d > \mu_0 + \frac{s_d}{\sqrt{n}} F_{t_{n-1}}^{-1}(1 - \alpha) = 15 + \frac{8.7311}{\sqrt{10}} (1.833) = 20.0613$$

Es decir, rechazar  $H_0: \mu_d \leq 15$  si  $\bar{x}_d > 20.0613$ . En la figura 4.21 se muestran las regiones de la partición de la prueba.

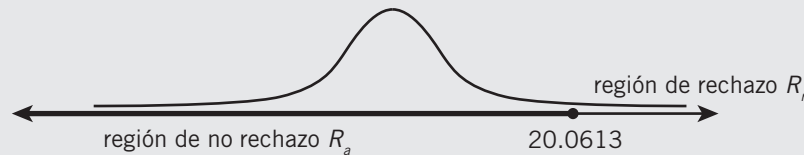


Figura 4.21 Regiones de la prueba del ejemplo 4.25.

iv) Por último, aplicamos la regla de decisión, para esto recordamos que  $\bar{x}_d = 19.7 < 20.0613$ . Luego, con 5% de significancia y la realización tomada no hay evidencias para rechazar  $H_0: \mu_d \leq 15$ .

**Conclusión:** con 5% de significancia y la realización obtenida no es válida la afirmación del veterinario de que  $\mu_d > 15$ .

c) Para calcular la potencia de la prueba utilizamos la región de rechazo,

$$P(\text{rechazar } H_0 | \mu_d > 15) = P(\bar{X}_d > 20.0613 | \mu_d = 20) = P\left(T_9 > \frac{20.0613 - 20}{8.7311/\sqrt{10}}\right) = P(T_9 > 0.0222) \approx 0.4914$$

La potencia de la prueba es pequeña cuando la verdadera media en los incrementos de los latidos del corazón sea 20.

#### Ejemplo 4.26 Aplicaciones del teorema 4.7

En las condiciones del problema anterior suponga que el veterinario afirma que la condición experimental aumenta el número de latidos del corazón en menos de 12.

- a) Plantee el contraste de hipótesis apropiado para este problema.  
 b) Justifique a un nivel de significancia de 0.05 si es válida la afirmación del veterinario.  
 c) Calcule la potencia de la prueba para una diferencia promedio de 20 latidos.

### Solución

- a) En este caso tenemos que  $0 < \mu_d < 12$ , debemos probar el contraste de hipótesis:

$$H_0: \mu_d \leq 0 \cup \mu_d \geq 12 \text{ contra } H_1: 0 < \mu_d < 12$$

- b) Para justificar si es válida la afirmación del veterinario, tenemos para muestras pareadas:

i)  $H_0: \mu_d \leq 0 \cup \mu_d \geq 12$  contra  $H_1: 0 < \mu_d < 12$

ii) Nivel de significancia  $\alpha = 0.05$ .

iii) Calcular  $\frac{s_d}{\sqrt{n}} F_{t_{n-1}}^{-1}(1 - \alpha/2)$  y  $\mu_0 + \frac{s_d}{\sqrt{n}} F_{t_{n-1}}^{-1}(1 - \alpha/2)$  y comparar con el valor de la EP  $\bar{x}_d$ .

Para esto se calculó  $d_0 = 15$ ,  $\bar{x}_d = 19.7$  y  $s_d = 8.7311$  y  $\alpha = 0.05$  con las tablas porcentuales para la distribución t-Student y  $n - 1 = 10 - 1 = 9$  grados de libertad,  $F_{t_9}^{-1}(1 - 0.025) = 2.262$  y  $F_{t_9}^{-1}(0.025) = -2.262$ . Al final, la regla de decisión:

Rechazar  $H_0: \mu_d \leq 0 \cup \mu_d \geq 12$ , si  $\frac{s_d}{\sqrt{n}} F_{t_{n-1}}^{-1}(\alpha/2) < \bar{x}_d < \mu_0 + \frac{s_d}{\sqrt{n}} F_{t_{n-1}}^{-1}(1 - \alpha/2)$ . Es decir:

$$-\frac{8.7311}{\sqrt{9}} 2.262 < \bar{x}_d < 12 + \frac{8.7311}{\sqrt{9}} 2.262$$

$$-6.2459 < \bar{x}_d < 18.2459$$

Es decir, rechazar  $H_0: \mu_d \leq 0 \cup \mu_d \geq 12$ , si  $\bar{x}_d \in (-6.2459, 18.2459)$ . En la figura 4.22 se muestran las regiones de la partición de la prueba.



Figura 4.22 Regiones de la prueba del ejemplo 4.26.

- iv) Por último, aplicamos la regla de decisión, para  $\bar{x}_d \in (-6.25, 18.25)$ . Luego, con 5% de significancia no hay evidencia para rechazar  $H_0: \mu_d \leq 0 \cup \mu_d \geq 12$ .

**Conclusión:** con 5% de significancia y la realización obtenida no es válida la afirmación de que la condición experimental aumenta el número de latidos del corazón en menos de 12.

- c) Para calcular la potencia de la prueba utilizamos la región de rechazo:

$$P(\text{rechazar } H_0 \mid \mu_d > 15) = P(-6.2459 < \bar{X}_d < 18.2459 \mid \mu_d = 20) = P\left(\frac{-6.2459 - 20}{8.73112/\sqrt{10}} < T_9 < \frac{18.2459 - 20}{8.7311/\sqrt{10}}\right)$$

$$= P(-9.5059 < T_9 < -0.6353) \approx 0.2705$$

La potencia de la prueba es pequeña para el caso en que la verdadera media en los incrementos de los latidos del corazón sea 20.

## Pruebas de hipótesis para la razón entre varianzas de poblaciones normales

Se ha visto hasta aquí que la forma de comparar dos varianzas de poblaciones normales es por medio de la estadística de prueba  $S_{n_1-1}^2/S_{n_2-1}^2$ , utilizando la distribución:

$$f = \frac{\chi_{n_1-1}^2}{\chi_{n_2-1}^2} = \left( \frac{S_{n_1-1}^2}{S_{n_2-1}^2} \right) \frac{1}{\sigma_1^2/\sigma_2^2}$$

llamada  $f$  de Snedecor con  $n_1 - 1$  y  $n_2 - 1$  g.l. en el numerador y denominador, respectivamente.

En el teorema 4.8 se formulan los resultados para los cuatro contrastes de hipótesis que tratamos en el texto, ahora en el caso de razón entre varianzas.

### Teorema 4.8

Sean  $X_1, X_2, \dots, X_{n_1}$  y  $Y_1, Y_2, \dots, Y_{n_2}$  muestras aleatorias independientes de  $N(\mu_1, \sigma_1^2)$  y  $N(\mu_2, \sigma_2^2)$ , respectivamente, entonces podemos tener alguno de los siguientes contrastes de hipótesis:

a)  $H_0 : \frac{\sigma_1^2}{\sigma_2^2} \geq r_0$  contra  $H_1 : \frac{\sigma_1^2}{\sigma_2^2} < r_0$ , entonces la prueba de tamaño  $\alpha$ , para  $\alpha \in (0, 1)$  está dada por la siguiente regla de decisión para una realización  $x_1, x_2, \dots, x_{n_1}$  y  $y_1, y_2, \dots, y_{n_2}$ :

$$\text{Rechazar } H_0 : \frac{\sigma_1^2}{\sigma_2^2} \geq r_0, \text{ si } \frac{s_{n_1-1}^2}{s_{n_2-1}^2} < r_0 F_{f(n_1-1, n_2-1)}^{-1}(\alpha) \text{ o } \frac{s_{n_1-1}^2}{s_{n_2-1}^2} < r_0 f_{1-\alpha}(n_1-1, n_2-1)$$

b)  $H_0 : \frac{\sigma_1^2}{\sigma_2^2} \leq r_0$  contra  $H_1 : \frac{\sigma_1^2}{\sigma_2^2} > r_0$ , entonces la prueba de tamaño  $\alpha$ , para  $\alpha \in (0, 1)$  estará dada por la siguiente regla de decisión para una realización  $x_1, x_2, \dots, x_{n_1}$  y  $y_1, y_2, \dots, y_{n_2}$ :

$$\text{Rechazar } H_0 : \frac{\sigma_1^2}{\sigma_2^2} \leq r_0, \text{ si } \frac{s_{n_1-1}^2}{s_{n_2-1}^2} > r_0 F_{f(n_1-1, n_2-1)}^{-1}(1-\alpha) \text{ o } \frac{s_{n_1-1}^2}{s_{n_2-1}^2} > r_0 f_{\alpha}(n_1-1, n_2-1)$$

c)  $H_0 : \frac{\sigma_1^2}{\sigma_2^2} = r_0$  contra  $H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq r_0$ , entonces la prueba de tamaño  $\alpha$ , para  $\alpha \in (0, 1)$  estará dada por la siguiente regla de decisión para una realización  $x_1, x_2, \dots, x_{n_1}$  y  $y_1, y_2, \dots, y_{n_2}$ :

$$\text{Rechazar } H_0 : \frac{\sigma_1^2}{\sigma_2^2} = r_0, \text{ si } \frac{s_{n_1-1}^2}{s_{n_2-1}^2} < r_0 F_{f(n_1-1, n_2-1)}^{-1}(\alpha/2) \text{ o } \frac{s_{n_1-1}^2}{s_{n_2-1}^2} > r_0 F_{f(n_1-1, n_2-1)}^{-1}(1-\alpha/2)$$

d)  $H_0 : r_0 \leq \frac{\sigma_1^2}{\sigma_2^2} \leq r_1$  contra  $H_1 : \frac{\sigma_1^2}{\sigma_2^2} < r_0$  o  $\frac{\sigma_1^2}{\sigma_2^2} > r_1$ , entonces la prueba de tamaño  $\alpha$ , para  $\alpha \in (0, 1)$  está dada por la siguiente regla de decisión para una realización  $x_1, x_2, \dots, x_{n_1}$  y  $y_1, y_2, \dots, y_{n_2}$ :

$$\text{Rechazar } H_0 : r_0 \leq \frac{\sigma_1^2}{\sigma_2^2} \leq r_1, \text{ si } \frac{s_{n_1-1}^2}{s_{n_2-1}^2} < r_0 F_{f(n_1-1, n_2-1)}^{-1}(\alpha/2) \text{ o } \frac{s_{n_1-1}^2}{s_{n_2-1}^2} > r_1 F_{f(n_1-1, n_2-1)}^{-1}(1-\alpha/2)$$

Con  $r_0, r_1 \in \mathbb{R}^+$  valores conocidos de antemano. Donde,  $F_{f(n,m)}^{-1}(\gamma)$  representa el cuantil  $\gamma$  de la distribución  $f$  con  $n$  y  $m$  grados de libertad en el numerador y denominador, respectivamente, para  $\gamma \in (0, 1)$ , o  $f_{\gamma}(n, m)$  representa el valor de la variable  $f$  con  $n$  y  $m$  grados de libertad en el numerador y denominador, respectivamente, para  $\gamma \in (0, 1)$ .

Note que con esta metodología se da respuesta a preguntas más generales sobre la relación entre dos varianzas debido a que utilizamos la razón  $r_0$  que puede ser igual, mayor o menor a 1.

En los siguientes dos ejemplos se muestran aplicaciones del teorema 4.8.

#### Ejemplo 4.27 Aplicaciones al teorema 4.8

Se comparan dos tipos de rosca de tornillo para ver su resistencia a la tensión. Se prueban 12 piezas de cada tipo de cuerda bajo condiciones similares, de lo que se obtienen los resultados en kilogramos que se aprecian en la tabla 4.8.

Tabla 4.8

Tipo de rosca	1	2	3	4	5	6	7	8	9	10	11	12
I	78	76	80	79	78	80	82	81	79	83	80	82
II	83	80	82	83	81	80	79	80	82	78	79	81

Se desea probar si es válida la suposición que se hizo en el ejemplo 4.21 sobre las varianzas  $\sigma_1^2 = \sigma_2^2$ .

- Plantee el contraste de hipótesis apropiado para este problema.
- Justifique a un nivel de significancia de 0.05, si es válida la suposición.
- Calcule la potencia de la prueba para  $\sigma_1^2/\sigma_2^2 = 2$ .

#### Solución

a) Comparamos varianzas por medio de la igualdad  $\sigma_1^2 = \sigma_2^2$ , pero no tenemos una metodología para la igualdad de varianzas. Por esta razón, pasamos a una razón entre varianzas, de lo que se obtienen las hipótesis  $H_0: \sigma_1^2/\sigma_2^2 = 1$  contra  $H_1: \sigma_1^2/\sigma_2^2 \neq 1$ .

b) Siguiendo los pasos de la metodología para realizar la comprobación y determinar la prueba:

i)  $H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$  contra  $H_1: \frac{\sigma_1^2}{\sigma_2^2} \neq 1$

ii) Nivel de significancia  $\alpha = 0.05$ .

iii) Tenemos que calcular  $r_0 F_{f(n_1-1, n_2-1)}^{-1}(\alpha/2)$  y  $r_0 F_{f(n_1-1, n_2-1)}^{-1}(1-\alpha/2)$  y compararla con la EP  $s_{n_1-1}^2/s_{n_2-1}^2$ . Si se calcula cada componente,  $r_0 = 1$ ,  $s_1^2 = 3.97$ ,  $s_2^2 = 2.61$ . De las tablas porcentuales para la distribución  $f$  de Snedecor con  $n_1 - 1 = n_2 - 1 = 12 - 1 = 11$  grados de libertad, del numerador y denominador;  $F_{f(11,11)}^{-1}(0.975) = 3.474$ , para encontrar  $F_{f(11,11)}^{-1}(0.025)$  recurrimos a la relación  $f_{1-\alpha}(m, n) = 1/f_\alpha(n, m)$ . Luego,

$$F_{f(11,11)}^{-1}(0.025) = \frac{1}{F_{f(11,11)}^{-1}(1-0.025)} = \frac{1}{3.474}$$

Al final, la regla de decisión con  $r_0 = 1$ :

$$\text{Rechazar } H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1, \text{ si } \frac{s_{n_1-1}^2}{s_{n_2-1}^2} < r_0 F_{f(n_1-1, n_2-1)}^{-1}(\alpha/2) = 1 \times \frac{1}{3.474} = 0.2879 \text{ o}$$

$$\frac{s_{n_1-1}^2}{s_{n_2-1}^2} > r_0 F_{f(n_1-1, n_2-1)}^{-1}(\alpha) = 1 \times 3.474 = 3.474$$



Es decir, rechazar  $H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$  si  $\frac{s_{n_1-1}^2}{s_{n_2-1}^2} < 0.2879$  o  $\frac{s_{n_1-1}^2}{s_{n_2-1}^2} > 3.474$ . En la figura 4.23 se muestra la partición de la prueba.

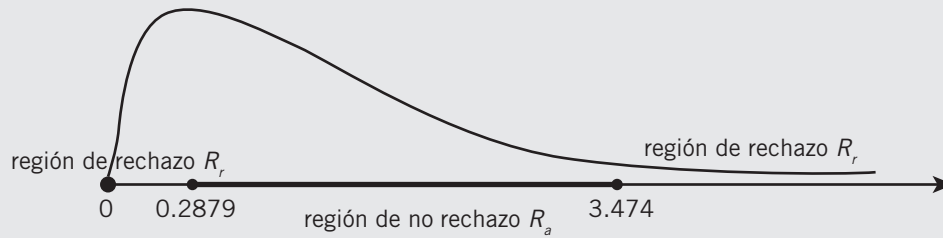


Figura 4.23 Regiones de la prueba del ejemplo 4.27.

iv) Por último aplicamos la regla de decisión, para esto recordamos que  $s_1^2 = 3.97$  y  $s_2^2 = 2.61$ , de donde  $\frac{s_{n_1-1}^2}{s_{n_2-1}^2} = \frac{3.97}{2.61} = 1.52 \in [0.2879, 3.474]$ ; luego, concluimos que no se rechaza  $H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$  a 5% de significancia.

**Conclusión:** con 5% de significancia se considera válida la afirmación de varianzas iguales.

c) Para calcular la potencia de la prueba utilizamos la región de rechazo:

$$\begin{aligned} P\left(\text{rechazar } H_0 \left| \frac{\sigma_1^2}{\sigma_2^2} \neq 1 \right. \right) &= P\left(\frac{s_{n_1-1}^2}{s_{n_2-1}^2} < 0.2879 \cup \frac{s_{n_1-1}^2}{s_{n_2-1}^2} > 3.474 \left| \frac{\sigma_1^2}{\sigma_2^2} = 2 \right. \right) \\ &= P\left(\frac{s_{n_1-1}^2}{s_{n_2-1}^2} \left( \frac{1}{\frac{\sigma_1^2}{\sigma_2^2}} \right) < 0.2879 \left( \frac{1}{\frac{\sigma_1^2}{\sigma_2^2}} \right) \left| \frac{\sigma_1^2}{\sigma_2^2} = 2 \right. \right) + P\left(\frac{s_{n_1-1}^2}{s_{n_2-1}^2} \left( \frac{1}{\frac{\sigma_1^2}{\sigma_2^2}} \right) > 3.474 \left( \frac{1}{\frac{\sigma_1^2}{\sigma_2^2}} \right) \left| \frac{\sigma_1^2}{\sigma_2^2} = 2 \right. \right) \\ &= P\left(f_{11,11}^2 < 0.2879 \left( \frac{1}{2} \right)\right) + P\left(f_{11,11}^2 > 3.474 \left( \frac{1}{2} \right)\right) \\ &= P\left(f_{11,11}^2 < 0.144\right) + P\left(f_{11,11}^2 > 1.737\right) = 0.0016 + 0.1868 = 0.1884 \end{aligned}$$

Si la verdadera razón entre varianzas fuera 2 la prueba tendría una potencia baja.

1. Para calcular el valor de la probabilidad con la distribución  $f$ , en el paquete Microsoft-Excel 2016, en la pestaña de función se escribe:  $= (1 - \text{DISTR.F}(0.144, 11, 11)) + \text{DISTR.F.CD}(1.737, 11, 11) = 0.1884586$ . 0.144 y 1.737 son los valores de los cuantiles con los que se van a calcular las probabilidades a la derecha, 11 y 11 son los grados de libertad del numerador y denominador, respectivamente. Se toma el complemento, debido a que Excel calcula las probabilidades a la derecha.
2. Donde 0.144 y 1.737 son los valores de los cuantiles con los cuales se calcularán las probabilidades a la izquierda y a la derecha, respectivamente, con 11 y 11 grados de libertad del numerador y denominador, de manera respectiva; en este caso se complica la búsqueda, porque las tablas estadísticas para la distribución  $f$  solo tienen una pequeña gama de valores. Para  $P(f_{11,11} > 1.737)$  sus valores más próximos son:  $P(f_{11,11} > 1.685) = 0.20$  y  $P(f_{11,11} > 2.227) = 0.10$ , interpolamos  $P(f_{11,11} > 1.737)$ , se tiene  $P(f_{11,11} > 1.737) \approx 0.1904$  próximo a 0.1868 valor obtenido con el paquete.

#### Ejemplo 4.28 Aplicaciones al teorema 4.8

En los ejemplos 4.22 y 4.24 trabajamos el problema de un fabricante de soldadura que creó un recubrimiento. Se hicieron las suposiciones de que las varianzas eran iguales (ejemplo 4.22) y diferentes (ejemplo 4.24). Lleva-

remos a cabo una prueba de hipótesis para las varianzas, en la que afirmaremos que la varianza de la población 1 es mayor que la varianza de la población 2. Para esto recordamos que  $s_1^2 = 7.50$ ,  $n_1 = 10$ ,  $s_2^2 = 2.68$ ,  $n_2 = 8$  y las poblaciones de soldaduras tienen distribución normal.

- a) Plantee el contraste de hipótesis apropiado para este problema.  
b) Justifique a un nivel de significancia de 0.10, si la afirmación es válida.

### Solución

- a) Se afirmó que  $\sigma_1^2 > \sigma_2^2$ . Así,  $\sigma_1^2 > \sigma_2^2$  será la hipótesis alterna y la contrapuesta  $\sigma_1^2 \leq \sigma_2^2$  la hipótesis nula.  
b) Si se siguen los pasos de la metodología para realizar la comprobación y determinar la prueba:

i)  $H_0: \sigma_1^2 \leq \sigma_2^2$  contra  $H_1: \sigma_1^2 > \sigma_2^2$  o  $H_0: \sigma_1^2/\sigma_2^2 \leq 1$  contra  $H_1: \sigma_1^2/\sigma_2^2 > 1$ .

ii) Nivel de significancia  $\alpha = 0.10$ .

iii) Estamos ante una situación similar a la del inciso b) del teorema 4.8. Requerimos calcular  $r_0 F_{f(n_1-1, n_2-1)}^{-1}(1-\alpha)$  y compararla con el valor de la EP  $s_{n_1-1}^2/s_{n_2-1}^2$ .

Si se calcula cada variable,  $r_0 = 1$ ,  $s_1^2 = 7.50$ ,  $s_2^2 = 2.68$ . Así, de las tablas porcentuales para la distribución  $f$  de Snedecor con  $n_1 - 1 = 10 - 1 = 9$  y  $n_2 - 1 = 8 - 1 = 7$  grados de libertad, del numerador y denominador;  $F_{f(9,7)}^{-1}(1 - 0.10) = F_{f(9,7)}^{-1}(0.90) = 2.725$ . Por último, la regla de decisión:

$$\text{Rechazar } H_0: \sigma_1^2/\sigma_2^2 \leq 1, \text{ si } \frac{s_{n_1-1}^2}{s_{n_2-1}^2} > r_0 F_{f(n_1-1, n_2-1)}^{-1}(1-\alpha) = 1 \times 2.725 = 2.725$$

Es decir,  $H_0: \sigma_1^2/\sigma_2^2 \leq 1$ , cuando  $s_{n_1-1}^2/s_{n_2-1}^2 > 2.725$ . La figura 4.24 muestra las regiones de la partición de la prueba.

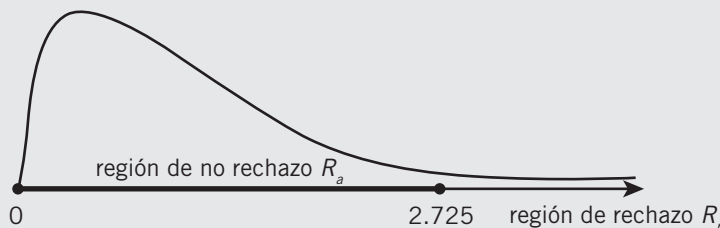


Figura 4.24 Regiones de la prueba del ejemplo 4.28.

- iv) Por último, aplicamos la regla de decisión, para esto recordamos que  $s_1^2 = 7.50$  y  $s_2^2 = 2.68$ , luego  $s_{n_1-1}^2/s_{n_2-1}^2 = 7.50/2.68 = 2.799$  pertenece a la región de rechazo.

**Conclusión:** con 10% de significancia se rechaza  $H_0: \sigma_1^2/\sigma_2^2 \leq 1$ .

## Ejercicios 4.3

- En un proceso químico se comparan dos catalizadores para verificar su efecto en el resultado de la reacción del proceso. Se preparó una muestra de 22 procesos al utilizar el catalizador 1 y una de 20 con el catalizador 2. En el primer caso se obtuvo un rendimiento promedio de 85, mientras que en la segunda fue de 81. Suponga que las poblaciones están distribuidas aproximadamente en forma normal con varianzas de 16 y 25. Un investigador A afirma que ambos catalizadores tienen un mismo efecto en promedio en la reacción del proceso. Para verificar la afirmación haga lo que se pide a continuación.
  - Plantee un contraste de hipótesis adecuado para este problema.
  - Realice la prueba con un nivel de significancia de 0.05.

2. Del ejercicio anterior calcule la potencia de la prueba si  $\mu_1 - \mu_2 = \bar{x}_1 - \bar{x}_2$ .
3. Del ejercicio del proceso químico un investigador *B* afirma que el catalizador 1 tiene un efecto promedio en la reacción del proceso en más de una unidad. Para verificar la afirmación:
  - a) Plantee un contraste de hipótesis adecuado para este problema.
  - b) Realice la prueba con un nivel de significancia de 0.05.
4. Del ejercicio anterior calcule la potencia de la prueba si  $\mu_1 - \mu_2 = \bar{x}_1 - \bar{x}_2$ .
5. Para comparar la resistencia a la tensión de dos tipos de rosca de tornillos se lleva a cabo un experimento; se prueban 31 piezas del tipo de cuerda 1 y 41 para el tipo de cuerda 2, de lo que se obtienen los siguientes resultados: El tipo 1 tuvo una resistencia promedio de 83 kg con una desviación estándar de 6 kg, mientras que el tipo 2 tuvo una resistencia promedio a la tensión de 88 kilogramos con una desviación estándar de 9. Suponga normalidad en las tensiones de las roscas de los tornillos, los fabricantes afirman que en promedio la resistencia a la tensión del tipo 2 es mayor a la 1 entre 1 y 2 kg.
  - a) Plantee un juego de hipótesis adecuado al problema.
  - b) Pruebe el contraste anterior a 0.10 de significancia, suponga varianzas iguales, pero desconocidas.
  - c) Calcule la potencia de la prueba para  $\mu_2 - \mu_1 = 7$  kg.
6. Resuelva el ejercicio anterior, suponga varianzas diferentes y desconocidas. Además calcule la potencia de la prueba para  $\mu_2 - \mu_1 = 7$  kg.
7. Resuelva el ejercicio anterior de la resistencia a la tensión de los dos tipos de rosca por aproximación. Además, calcule la potencia de la prueba para  $\mu_2 - \mu_1 = 7$  kg.
8. En el ejercicio anterior de la resistencia a la tensión de los dos tipos de rosca:
  - a) Plantee el contraste de hipótesis para conocer qué supuestos fueron los correctos, varianzas iguales o diferentes.
  - b) Pruebe el contraste de hipótesis al 0.10 de significancia, ¿qué supuesto fue el correcto?
9. En el ejercicio anterior:
  - a) Pruebe a 0.10 de significancia la afirmación de que la varianza 2 es mayor a la varianza 1.
  - b) Calcule la potencia de la prueba para una razón de varianzas de 3.
10. Un centro de investigación en medicina del deporte afirmó en 2010 que hay diferencias en las tasas medias de consumo de oxígeno para varones universitarios entrenados con dos métodos diferentes e independientes. La afirmación se basó en dos experimentos que realizó con los deportistas. Uno utilizó entrenamiento continuo y otro, entrenamiento intermitente con la misma duración. En la tabla 4.9 se registran los tamaños de muestra, medias y desviaciones estándar respectivas, expresados en ml por kg/min. Suponga que las poblaciones tienen distribución normal. Con base en esta información los responsables del centro afirman que  $\sigma_c^2 \neq \sigma_i^2$ . Con un nivel de significancia de 0.10 verifique si es válida la afirmación.

Tabla 4.9

Entrenamiento continuo	Entrenamiento intermitente
$n_c = 16$	$n_i = 15$
$\bar{x}_c = 43.71$	$\bar{x}_i = 39.63$
$s_c = 4.87$	$s_i = 9.68$

11. Con base en el resultado del ejercicio anterior, pruebe a 0.10 de significancia si con las dos realizaciones es válida la afirmación del consumo medio de oxígeno de los deportistas.
12. Los promotores de una nueva dieta afirman que reduce el peso de una persona en promedio más de 2 kg en un periodo de dos semanas. Los pesos de siete mujeres que siguieron esta dieta fueron anotados antes y después del periodo de estudio.

Tabla 4.10

Mujer	1	2	3	4	5	6	7	8	9
Peso anterior	58.5	60.3	61.7	69.0	64.0	62.6	56.7	70.5	68.4
Peso posterior	60.0	54.9	58.1	62.1	58.5	59.9	54.4	62.4	63.7

Suponga que la distribución de las diferencias por persona es aproximadamente normal.

- a) Es adecuado suponer muestras pareadas. Explique su respuesta.
  - b) Plantee un contraste de hipótesis adecuado para verificar si estadísticamente es válida la afirmación de los promotores de la dieta y lleve a cabo la prueba con un nivel de significancia de 10%.
13. Del ejercicio anterior, formule las pruebas de hipótesis adecuadas con 10% de significancia y resuelva:
- a) ¿Qué pasa si después de ver el promedio de las diferencias los promotores afirman que la dieta disminuye el peso en más de 3.5 kg? Explique ambos resultados.
  - b) Con base en los resultados anteriores y con conocimientos de estadística, la competencia afirma que esta nueva dieta sí disminuye el peso, pero en promedio menos de 2.4 kg.
14. Se aplicó un examen de matemáticas financieras (1) a un grupo de alumnos y se obtuvieron las siguientes calificaciones: 3.0, 3.5, 4.0, 8.1, 7.2, 8.9, 8.2, 10, 10, 9. A otro grupo de álgebra lineal (2), independiente del anterior, se aplicó otro examen, quienes obtuvieron las calificaciones siguientes: 2.0, 3.0, 3.7, 8.0, 5.0, 4.0, 3.0, 8.0, 9.0, 10, 7.7, 6.0. El director afirma que el promedio en la materia de matemáticas financieras es mayor al de álgebra lineal en 1 punto. Suponga normalidad en las calificaciones.
- a) Es correcto aplicar muestras pareadas. Justifique su respuesta.
  - b) Plantee el contraste de hipótesis apropiado para la afirmación del director.
15. Del ejercicio anterior, justifique con 5% de significancia, la información de ambos exámenes la igualdad o diferencia de las varianzas poblacionales.
16. Con base en el resultado del ejercicio anterior:
- a) Pruebe con 5% de significancia el contraste de hipótesis con respecto a la afirmación del director sobre ambas materias. ¿Resulta válida la afirmación?
  - b) Calcule la potencia de la prueba para  $\mu_M - \mu_A = 3$ .
17. Los empleados de una empresa realizan dos actividades; el supervisor afirma que las personas son en promedio igual de hábiles en ambas, pues tardan el mismo tiempo. El supervisor dice que realizará la prueba estadística de su afirmación. Como primer paso selecciona una muestra aleatoria de 11 trabajadores y anota sus tiempos en ambas actividades.

Tabla 4.11 Número de trabajo

	1	2	3	4	5	6	7	8	9	10	11
Actividad 1	0.50	1.40	0.95	0.45	0.25	1.20	1.60	2.6	1.30	0.35	0.80
Actividad 2	1.46	1.52	0.09	0.33	0.71	1.31	1.49	2.9	1.41	0.83	0.74

El supervisor sabe que existe normalidad en las diferencias de los tiempos, pero desconoce qué hacer y le pide ayuda a usted para realizar la prueba. ¿Es válida la afirmación del supervisor con 8% de significancia?

18. Las pruebas de tracción en 10 puntos de soldadura en un dispositivo semiconductor (1) produjeron los siguientes resultados en libras requeridas para romper la soldadura:

14.6    13.2    11.5    15.9    12.6    17.1    11.1    12.9    15.4    15.2

Otro conjunto de 10 puntos fue probado después que el dispositivo recibió un recubrimiento (2). Para determinar si la resistencia a la tracción se incrementa con éste, se obtuvieron los siguientes resultados.

18.9    21.1    18.8    20.1    17.4    16.9    19.5    22.5    19.3    20.8

El fabricante afirma que el recubrimiento es eficaz (un recubrimiento en el dispositivo, para aumentar la resistencia a la tensión, se considera así cuando la resistencia es superior en más de 3.8 lb con respecto al dispositivo sin recubrimiento). Suponga normalidad en las pruebas de tracción. Pruebe con 0.05 de significancia si con las dos realizaciones es válido suponer que las varianzas poblacionales son iguales.

19. Del ejercicio anterior.

- Plantee el contraste de hipótesis apropiado para la afirmación del fabricante sobre el recubrimiento.
- Con base en el resultado del ejercicio anterior pruebe de manera estadística, a un nivel de significancia de 0.05, si es válida la afirmación del fabricante.

Tabla 4.12

BM	GMD	BM	GMD
25.02	27.79	24.54	27.66
24.84	27.73	24.12	27.80
25.19	27.70	24.09	27.78
24.80	27.06	24.19	28.10
24.83	27.17	23.85	28.25
24.60	27.06	23.52	28.30
24.44	27.14	23.51	28.00
24.30	27.32	23.35	28.04
24.31	27.41		

20. En el ejercicio anterior calcule la potencia de la prueba para  $\mu_2 - \mu_1 = 6$  lb.

21. El IPC de las empresas BM (B) y GMD (G) se muestran en la tabla 4.12. Suponga que el IPC anual de las empresas tiene una distribución normal y que las muestras son independientes. Los gerentes suponen que las varianzas anuales son diferentes. Pruebe con 0.05 de significancia si con las dos realizaciones es válida la suposición sobre las varianzas poblacionales.

22. Del ejercicio anterior:

- El gerente de GMD afirma que su IPC en promedio es mayor al de BM en más de dos unidades. Plantee el contraste de hipótesis apropiado para esta afirmación.
- En el nivel de significancia de 0.05, puede concluirse que es válida la afirmación del gerente de GMD.

23. Del ejercicio anterior:

- Calcule la potencia de la prueba para  $\mu_G - \mu_B = 2.5$ .
- ¿Qué conclusiones encontró con respecto a la prueba y su potencia?

24. En un proceso químico se comparan dos catalizadores para verificar su efecto en el resultado de la reacción del proceso. Se preparó una muestra de 12 procesos con el catalizador marca  $L$  y también 12 catalizadores de la marca  $M$ , independientes de la marca  $L$ . En la tabla 4.13 se muestran los datos con los rendimientos.

Tabla 4.13

$L$	0.89	0.92	0.68	0.76	0.78	0.87	0.76	0.67	0.65	0.61	0.68	0.97
$M$	0.95	0.79	0.62	0.64	0.68	0.70	0.56	0.62	0.56	0.72	0.56	0.78

Suponga normalidad en los efectos de los catalizadores  $L$  y  $M$ . Los responsables del proceso afirman que las varianzas poblacionales de ambos catalizadores son iguales. Pruebe con 0.05 de significancia si con las 12 realizaciones es válida la afirmación sobre las varianzas.

25. En el ejercicio anterior:

- Los responsables del proceso afirman que los efectos de ambos catalizadores son iguales. Plantee un contraste de hipótesis adecuado para esta afirmación.
- Al nivel de significancia de 5% pruebe si es válida la afirmación de los responsables.

26. Repita el ejercicio anterior para un nivel de significancia de 10%.

27. En los dos ejercicios anteriores.

- Calcule la potencia de la prueba con  $\mu_L - \mu_M = 0.10$  para ambos niveles de significancia.
- ¿Qué conclusiones haría de ambas pruebas y cuál elegiría?

28. En las ciudades de Guadalajara y Monterrey se llevó a cabo una investigación sobre el costo de la vida, para estimar el costo promedio en alimentación en familias con cuatro integrantes. De cada una de estas ciudades se seleccionaron muestras aleatorias independientes de 21 familias, cuyos resultados son:

$$\sum_{i=1}^{21} x_i = 139,150, \quad \sum_{i=1}^{21} x_i^2 = 1,103,192,500, \quad \sum_{i=1}^{21} y_i = 139,720 \quad \text{y} \quad \sum_{i=1}^{21} y_i^2 = 1,114,254,400$$

Si supone que la distribución sobre el costo de vida en ambas ciudades es normal. Pruebe al 0.10 de significancia si es válido suponer varianzas poblacionales iguales.

29. En el ejercicio anterior, el investigador afirma que el costo de vida en ambas ciudades es igual.

- Plantee un contraste de hipótesis adecuado para esta afirmación.
- Realice una prueba con 10% de significancia si la afirmación del investigador es válida. Utilice el resultado del ejercicio anterior.

30. Del ejercicio anterior calcule la potencia de la prueba para  $\mu_M - \mu_G = 700$ .

## 4.4 Pruebas para poblaciones tipo Bernoulli, proporciones

En la presente sección cambiaremos de distribución y trabajaremos con poblaciones que tienen una distribución tipo Bernoulli. Es decir, revisaremos problemas donde las variables aleatorias aparecen en situaciones en las que el decisor solo tiene dos opciones.

- Al entrevistar a una persona para ver si apoya o no la política de un gobernante.
- Si una ama de casa compra o no un producto determinado.
- Si una persona consume o no una pasta de dientes marca  $A$ .
- Si un estudiante consumió droga en alguna ocasión.

Por ejemplo, en la situación política muy controvertida de 2013 sobre el plantón de maestros en la plancha del Zócalo de la Ciudad de México se pudo entrevistar a un grupo de capitalinos y preguntarles si estaban a favor o en contra, de manera que si  $X$  representa a la variable aleatoria, “la persona está a favor del plantón”, el éxito sería que la persona conteste que sí está a favor y el valor de la variable es 1, en caso contrario 0. Este tipo de variables aleatorias llamadas dicotómicas son muy comunes en las investigaciones, pues son variables aleatorias muy propicias para realizar estudios sobre preferencias, de hecho en una muestra aleatoria lo que nos interesa son la **suma** o **promedio**, este último lleva el nombre particular de **proporción**.

En la unidad 2 revisamos que si tenemos una muestra aleatoria  $X_1, X_2, \dots, X_n$  de variables tipo Bernoulli con parámetro  $p$ , entonces su suma  $T = \sum_{i=1}^n X_i$  tiene una distribución tipo binomial con parámetros  $n$  y  $p$ . Por otro lado,

cuando trabajamos con proporciones y las muestras estudiadas son grandes se acostumbra utilizar una aproximación con el teorema central del límite. Entonces, necesitamos el valor esperado y la varianza de la suma,  $E(T) = np$  y  $V(T) = npq$ , o en el caso de proporciones  $\hat{P} = \bar{X} = \frac{1}{n}T$ ,  $E(\hat{P}) = p$  y  $V(\hat{P}) = pq/n$ . Por último, el estadístico de prueba tendrá una distribución asintótica normal estándar dada por:

$$Z = \frac{T - np}{\sqrt{npq}} \text{ (sumas) o si se divide entre } n \quad Z = \frac{\hat{P} - p}{\sqrt{pq/n}} \text{ (proporciones).}$$

En el siguiente teorema se formulan los resultados para las cuatro pruebas de hipótesis en el caso de proporciones o sumas de variables de Bernoulli con muestras grandes.

### Teorema 4.9

Sea  $X_1, X_2, \dots, X_n$  una muestra aleatoria de distribuciones tipo Bernoulli ( $p$ ), entonces podemos tener alguno de los siguientes contraste de hipótesis condicionados para  $n$  grande.

a)  $H_0: p \geq p_0$  contra  $H_1: p < p_0$ , luego la prueba de tamaño  $\alpha$ , para  $\alpha \in (0, 1)$  estará dada por la siguiente regla de decisión para una realización  $x_1, x_2, \dots, x_n$ :

$$\text{Rechazar } H_0: p \geq p_0, \text{ si } \hat{p} < p_0 + \Phi^{-1}(\alpha)\sqrt{\frac{p_0q_0}{n}} \text{ o } \hat{p} < p_0 - Z_\alpha\sqrt{\frac{p_0q_0}{n}}$$

b)  $H_0: p \leq p_0$  contra  $H_1: p > p_0$ , luego la prueba de tamaño  $\alpha$  para  $\alpha \in (0, 1)$  estará dada por la siguiente regla de decisión para una realización  $x_1, x_2, \dots, x_n$ :

$$\text{Rechazar } H_0: p \leq p_0 \text{ si } \hat{p} > p_0 + \Phi^{-1}(1 - \alpha)\sqrt{\frac{p_0q_0}{n}} \text{ o } \hat{p} > p_0 + Z_\alpha\sqrt{\frac{p_0q_0}{n}}$$

c)  $H_0: p = p_0$  contra  $H_1: p \neq p_0$ , luego la prueba de tamaño  $\alpha$ , para  $\alpha \in (0, 1)$  estará dada por la siguiente regla de decisión para una realización  $x_1, x_2, \dots, x_n$ :

$$\text{Rechazar } H_0: p = p_0 \text{ si } \hat{p} < p_0 + \Phi^{-1}(\alpha/2)\sqrt{\frac{p_0q_0}{n}} \text{ o } \hat{p} > p_0 + \Phi^{-1}(1 - \alpha/2)\sqrt{\frac{p_0q_0}{n}}$$

d)  $H_0: p_0 \leq p \leq p_1$  contra  $H_1: p < p_0$  o  $p > p_1$ , luego la prueba de tamaño  $\alpha$ , para  $\alpha \in (0, 1)$  estará dada por la siguiente regla de decisión para una realización  $x_1, x_2, \dots, x_n$ :

$$\text{Rechazar } H_0: p_0 \leq p \leq p_1 \text{ si } \hat{p} < p_0 + \Phi^{-1}(\alpha/2)\sqrt{\frac{p_0q_0}{n}} \text{ o } \hat{p} > p_1 + \Phi^{-1}(1 - \alpha/2)\sqrt{\frac{p_0q_0}{n}}$$

Con  $p_0, p_1 \in [0, 1]$  y  $q_0 = 1 - p_0$  valores conocidos de antemano;  $\hat{t} = \sum_{i=1}^n x_i$  y  $\hat{p} = \bar{x} = \frac{1}{n}\hat{t}$ ,  $\Phi^{-1}(\gamma)$  representa el cuantil  $\gamma$  de la distribución normal estándar para  $\gamma \in (0, 1)$  y  $Z_\gamma$  representa el valor de la variable normal estándar cuya área a la derecha es  $\gamma \in (0, 1)$ .

En el caso de querer trabajar con la suma de variables y no con las proporciones se tiene que multiplicar la constante crítica por el tamaño de la muestra,

$$\hat{p} < p_0 + \Phi^{-1}(\alpha)\sqrt{\frac{p_0q_0}{n}} \Rightarrow \hat{t} < np_0 + \Phi^{-1}(\alpha)\sqrt{np_0q_0}$$



A continuación, se muestran dos ejemplos con aplicaciones de los resultados del teorema 4.9.

### Ejemplo 4.29 Aplicaciones del teorema 4.9

El director general de un canal de televisión asegura que la proporción de audiencia que ve cierto programa el sábado por la noche es mayor a 40%. Se eligió una muestra de 100 televidentes a quienes se entrevistó, cuyos resultados determinan que 45 de ellos ven el programa.

- Plantee un contraste de hipótesis adecuado para el problema.
- Al nivel de significancia de 2.5% pruebe si la afirmación es válida. Realice la comprobación con la suma y con la proporción y verifique que se obtiene el mismo resultado.
- Calcule la potencia de la prueba, suponga que  $p = 0.60$ .

#### Solución

a) Se pide una prueba de hipótesis para la proporción de televidentes que ven un programa el sábado por la noche, en cuyo caso el director del canal afirma que es mayor a 0.40,  $p > 0.40$ . Así, la suposición del director será la hipótesis alterna  $H_1: p > 0.40$  y su opuesta  $H_0: p \leq 0.40$ .

b) Si se siguen los pasos de la metodología resulta:

i)  $H_0: p \leq 0.40$  contra  $H_1: p > 0.40$ .

ii) Nivel de significancia  $\alpha = 0.025$ .

iii) Estamos ante una situación similar a la del inciso b) del teorema 4.9. Requerimos calcular para la suma  $np_0 + \Phi^{-1}(1 - \alpha)\sqrt{np_0q_0}$  o la proporción  $p_0 + \Phi^{-1}(1 - \alpha)\sqrt{\frac{p_0q_0}{n}}$  y comparar con el valor de la EP suma  $\hat{t}$  o la  $\hat{p} = \bar{x} = \frac{1}{n}\hat{t}$ .

Los valores calculados son los componentes  $p_0 = 0.40$ ,  $n = 100$ ,  $\hat{t} = 45$  para  $\alpha = 0.025$  y de las tablas porcentuales de la distribución normal estándar,  $\Phi^{-1}(0.975) = 1.96$ . Por último, la regla de decisión:

Rechazar  $H_0: p \leq 0.40$ , si:

$$\text{Suma: } \hat{t} > np_0 + \Phi^{-1}(1 - \alpha)\sqrt{np_0q_0} = 100(0.4) + 1.96\sqrt{100(0.4)(0.6)} = 49.60$$

$$\text{Proporciones: } \hat{p} > p_0 + \Phi^{-1}(1 - \alpha)\sqrt{\frac{p_0q_0}{n}} = 0.40 + 1.96\sqrt{\frac{0.4(0.6)}{100}} = 0.4960, (49.6/100)$$

Es decir, rechazar  $H_0: p \leq 0.40$  si  $\hat{t} > 49.6$  o las proporciones  $\hat{p} > 0.496$ , solo cambia su escala (49.6 o 0.469). De manera gráfica tenemos las regiones de la partición de la prueba en la figura 4.25.

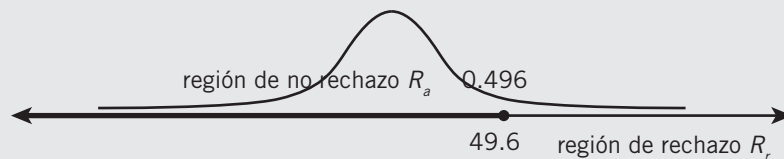


Figura 4.25 Regiones de la prueba del ejemplo 4.29.

- Por último, aplicamos la regla de decisión, para esto recordamos que  $\hat{t} = 45$  ( $\hat{p} = 45/100 = 0.45$ ); por tanto, concluimos que con la realización tomada no hay evidencias para rechazar  $H_0: p \leq 0.40$  con 2.5% de significancia.



**Conclusión:** con 2.5% de significancia y la realización obtenida no existen evidencias para validar la afirmación del director de la televisora  $p > 0.40$ .

c) Para calcular la potencia de la prueba utilizamos la región de rechazo:

$$P(\text{rechazar } H_0 \mid p > 0.40) = P(T > 49.6 \mid p = 0.60) = P\left(Z > \frac{49.6 - 100(0.6)}{\sqrt{100(0.6)(0.4)}}\right) = P(Z > -2.12) = 0.983$$

Si la verdadera proporción es 0.6 la prueba es buena porque tiene una potencia muy elevada.

### Ejemplo 4.30 Aplicaciones del teorema 4.9

Una persona quiere probar estadísticamente si una moneda está cargada hacia un lado, para lo cual la lanza 200 veces, de lo que obtiene 80 águilas. ¿Se puede concluir que la moneda está cargada?

- Plantee un contraste de hipótesis adecuado para el problema.
- Al nivel de significancia de 5% pruebe si la moneda está cargada.
- Calcule la potencia de la prueba, suponga que  $p = 0.40$ .

#### Solución

- Se pide una prueba de hipótesis para la proporción de resultados de la moneda, por ejemplo para las águilas, para esto suponemos que la moneda no está cargada. Luego, la proporción es 0.5, contra la alternativa de que la proporción sea diferente de 0.5. Así,  $H_0: p = 0.5$  y la alterna  $H_1: p \neq 0.5$ .
- Si se siguen los pasos de la metodología resulta:
  - $H_0: p = 0.5$  contra  $H_1: p \neq 0.5$
  - Nivel de significancia  $\alpha = 0.05$ .
  - Requerimos calcular la CC:  $p_0 + \Phi^{-1}(\alpha/2)\sqrt{p_0q_0/n}$  y  $p_0 + \Phi^{-1}(1 - \alpha/2)\sqrt{p_0q_0/n}$  y comparar con el valor de la EP  $\hat{p}$ .

Calculando  $p_0 = 0.5$ ,  $n = 200$ ,  $t = 80$  para  $\alpha = 0.05$  y de las tablas porcentuales de la distribución normal estándar,  $\Phi^{-1}(0.025) = -1.96$  y  $\Phi^{-1}(0.975) = 1.96$ . Al final, la regla de decisión:

Rechazar  $H_0: p = 0.5$ , si:

$$\hat{p} < p_0 + \Phi^{-1}(\alpha/2)\sqrt{p_0q_0/n} = 0.5 - 1.96\sqrt{(0.5)(0.5)/200} = 0.4307$$

$$\hat{p} > p_0 + \Phi^{-1}(1 - \alpha/2)\sqrt{p_0q_0/n} = 0.5 + 1.96\sqrt{(0.5)(0.5)/200} = 0.5693$$

Es decir, rechazar  $H_0: p = 0.5$  si  $\hat{p} < 0.4307$  o  $\hat{p} > 0.5693$ . En la figura 4.26 podemos apreciar la partición de la prueba.



Figura 4.26 Regiones de la prueba del ejemplo 4.30.

- Por último, aplicamos la regla de decisión, para lo cual recordamos que  $\hat{p} = 0.40$  y concluimos que con la realización tomada rechazamos  $H_0: p = 0.5$  con 5% de significancia.

**Conclusión:** con 5% de significancia y la realización obtenida existen evidencias para decidir que la moneda está cargada.

c) Para calcular la potencia de la prueba utilizamos la región de rechazo:

$$\begin{aligned} P(\text{rechazar } H_0 \mid p \neq 0.40) &= P(p < 0.4307 \cup p > 0.5693 \mid p = 0.40) \\ &= P\left(Z < \frac{0.4307 - 0.40}{\sqrt{(0.4)(0.6)/200}}\right) + P\left(Z > \frac{0.5693 - 0.40}{\sqrt{(0.4)(0.6)/200}}\right) \\ &= P(Z < 0.89) + P(Z > 4.89) = 0.8133 \end{aligned}$$

Si la verdadera proporción de águilas fuera 0.4 la prueba tendría una potencia elevada.

Ahora bien, ¿cómo podemos comparar proporciones de dos poblaciones? Con frecuencia, tenemos problemas en los que se desea comparar qué producto es más aceptado por los consumidores, el producto A o el B. Por este motivo, los investigadores o gerentes de negocios requieren resultados estadísticos que justifiquen sus aseveraciones.

Sean dos muestras aleatorias  $X_1, X_2, \dots, X_{n_1}$  y  $Y_1, Y_2, \dots, Y_{n_2}$  independientes de distribuciones tipo Bernoulli con parámetros  $p_1$  y  $p_2$ , respectivamente. La prueba de hipótesis para  $p_1 - p_2$  con muestras grandes y estimador  $\hat{p}_1 - \hat{p}_2 = \bar{X} - \bar{Y}$ , se realiza con la aproximación del TCL en donde  $E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$  y  $V(\hat{p}_1 - \hat{p}_2) = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$ .

Por último, el estadístico:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$$

El resultado son los dos parámetros en el denominador, que al buscar la mejor prueba no podemos eliminarlos. Por tal razón, usamos una estimación puntual para  $p_1$  y  $p_2$  en el denominador, que se sustituyen por  $\hat{p}_1$  y  $\hat{p}_2$ , valores de  $\hat{p}_1$  y  $\hat{p}_2$ , respectivamente, obtenidos de una realización previa de la muestra, cuyo resultado es la estadística de prueba que aproxima:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}}$$

Note que en esta situación solo consideramos las proporciones, puesto que con las sumas su estadística de prueba es:

$$Z = \frac{T_1 - T_2 - (n_1 p_1 - n_2 p_2)}{\sqrt{n_1 \hat{p}_1 \hat{q}_1 + n_2 \hat{p}_2 \hat{q}_2}}$$

la cual solo podemos utilizar cuando  $n_1 = n_2$ .

#### Teorema 4.10

Sean  $X_1, X_2, \dots, X_{n_1}$  y  $Y_1, Y_2, \dots, Y_{n_2}$  dos muestras aleatorias independientes de Bernoulli ( $p_1$ ) y Bernoulli ( $p_2$ ), entonces podemos tener alguno de los siguientes contrastes de hipótesis con tamaños de muestras grandes.

a)  $H_0: p_1 - p_2 \geq p_0$  contra  $H_1: p_1 - p_2 < p_0$ , luego la prueba de tamaño  $\alpha$ , para  $\alpha \in (0, 1)$  estará dada por la siguiente regla de decisión para una realización  $x_1, x_2, \dots, x_{n_1}$  y  $y_1, y_2, \dots, y_{n_2}$ :

$$\text{Rechazar } H_0: p_1 - p_2 \geq p_0, \text{ si } \hat{p}_1 - \hat{p}_2 < p_0 + \Phi^{-1}(\alpha) \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

b)  $H_0: p_1 - p_2 \leq p_0$  contra  $H_1: p_1 - p_2 > p_0$ , luego la prueba de tamaño  $\alpha$ , para  $\alpha \in (0, 1)$  estará dada por la siguiente regla de decisión para una realización  $x_1, x_2, \dots, x_{n_1}$  y  $y_1, y_2, \dots, y_{n_2}$ :

$$\text{Rechazar } H_0: p_1 - p_2 \leq p_0, \text{ si } \hat{p}_1 - \hat{p}_2 > p_0 + \Phi^{-1}(1 - \alpha) \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

c)  $H_0: p_1 - p_2 = p_0$  contra  $H_1: p_1 - p_2 \neq p_0$ , luego la prueba de tamaño  $\alpha$ , para  $\alpha \in (0, 1)$  estará dada por la siguiente regla de decisión para una realización  $x_1, x_2, \dots, x_{n_1}$  y  $y_1, y_2, \dots, y_{n_2}$ :

$$\text{Rechazar } H_0: p_1 - p_2 = p_0, \text{ si } \hat{p}_1 - \hat{p}_2 < p_0 + \Phi^{-1}(\alpha/2) \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}; \text{ o}$$

$$\hat{p}_1 - \hat{p}_2 > p_0 + \Phi^{-1}(1 - \alpha/2) \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

d)  $H_0: p_0 \leq p_1 - p_2 \leq p_{01}$  contra  $H_1: p_1 - p_2 < p_0$  o  $p_1 - p_2 > p_{01}$ , luego la prueba de tamaño  $\alpha$ , para  $\alpha \in (0, 1)$  estará dada por la siguiente regla de decisión para una realización  $x_1, x_2, \dots, x_{n_1}$  y  $y_1, y_2, \dots, y_{n_2}$ :

$$\text{Rechazar } H_0: p_0 \leq p_1 - p_2 \leq p_{01}, \text{ si } \hat{p}_1 - \hat{p}_2 < p_0 + \Phi^{-1}(\alpha/2) \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}; \text{ o}$$

$$\hat{p}_1 - \hat{p}_2 > p_{01} + \Phi^{-1}(1 - \alpha/2) \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

Con  $p_0, p_{01} \in [0, 1]$  valores conocidos de antemano,  $q_k = 1 - p_k$  y  $\hat{p}_k = \bar{x}_k$ ,  $\Phi^{-1}(\gamma)$  representa el cuantil  $\gamma$  de la distribución normal estándar para  $\gamma \in (0, 1)$ , y  $Z_\gamma$  representa el valor de la variable normal estándar cuya área a la derecha es  $\gamma \in (0, 1)$ .

Existen otras formas para obtener una estimación puntual de la varianza:

$$V(\hat{p}_1 - \hat{p}_2) = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$$

- Si se sustituyen  $p_1$  y  $p_2$  por un promedio de sus estimadores; es decir,  $\bar{p} = \frac{\hat{p}_1 + \hat{p}_2}{2}$ , tenemos:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

- Si se sustituye  $p_1$  y  $p_2$  por un promedio ponderado de sus estimadores, es decir  $\tilde{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2} = \frac{\hat{t}_1 + \hat{t}_2}{n_1 + n_2}$ , tenemos:

$$Z = \frac{(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)}{\sqrt{\tilde{p}\tilde{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Por último, podemos notar que en muchos textos prefieren utilizar el promedio ponderado de  $\hat{p}_1$  y  $\hat{p}_2$ . Si alguien desea hacerlo, solo necesita cambiar en el teorema anterior:

$$\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \text{ por } \sqrt{\tilde{p}\tilde{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \text{ o en su caso por } \sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

A continuación se muestran dos ejemplos con aplicaciones del teorema 4.10.

### Ejemplo 4.31 Aplicaciones del teorema 4.10

Una firma manufacturera de cigarros distribuye las marcas 1 y 2. El gerente de ventas desea saber si alguna de las marcas aventaja a la otra, para lo cual realiza dos encuestas independientes, cuyo resultado es que 56 de 200 fumadores prefieren la marca 1, mientras que 29 de 150 fumadores prefieren la marca 2. ¿Puede el gerente concluir que la marca 1 aventaja en ventas a la marca 2?

- Plantee un contraste de hipótesis adecuado para el problema.
- Al nivel de significancia de 6% pruebe si es válida la conclusión.
- Realice la prueba anterior, utilice la media ponderada de las proporciones.
- Calcule la potencia de la prueba, suponga que  $p_1 - p_2 = 0.08$ .

#### Solución

- Se pide una prueba de hipótesis para una diferencia de proporciones de consumidores de cigarros de las marcas 1 y 2. La suposición que se hace es que  $p_1 > p_2$ , así la opuesta será  $p_1 \leq p_2$  (el signo de igualdad se encuentra en la opuesta). Por tal razón,  $H_0: p_1 \leq p_2$  y  $H_1: p_1 > p_2$ .
- Si se siguen los pasos de la metodología, al convertir las hipótesis a una diferencia de proporciones:

i)  $H_0: p_1 - p_2 \leq 0$  contra  $H_1: p_1 - p_2 > 0$

ii) Nivel de significancia  $\alpha = 0.06$ .

iii) Estamos ante una situación similar a la del inciso b) del teorema 4.10, por lo que requerimos calcular

$$\text{la CC: } p_0 + \Phi^{-1}(1 - \alpha) \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \text{ y comparar con el valor de la EP } \hat{p}_1 - \hat{p}_2.$$

Si se calculan los componentes  $p_0 = 0$ ,  $n_1 = 200$ ,  $t_1 = 56 \Rightarrow \hat{p}_1 = 56/200 = 0.280$  y  $n_2 = 150$ ,  $t_2 = 29 \Rightarrow \hat{p}_2 = 29/150 = 0.193$  para  $\alpha = 0.06$  y de las tablas porcentuales de la distribución normal estándar,  $\Phi^{-1}(0.94) = 1.5548$ . Por último, la regla de decisión:

Rechazar  $H_0: p_1 - p_2 \leq 0$ , si:

$$\hat{p}_1 - \hat{p}_2 > p_0 + \Phi^{-1}(1 - \alpha) \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} = 0 + 1.5548 \sqrt{\frac{0.28(0.72)}{200} + \frac{0.193(0.807)}{150}} = 0.070$$

Es decir, rechazar  $H_0: p_1 - p_2 \leq 0$  si  $\hat{p}_1 - \hat{p}_2 > 0.070$ . En la figura 4.27 se muestra la partición de la prueba.

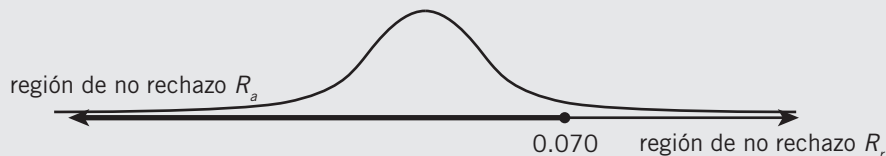


Figura 4.27 Regiones de la prueba del ejemplo 4.31b.

- Por último, aplicamos la regla de decisión. Para esto recordamos que  $\hat{p}_1 = 0.280$  y  $\hat{p}_2 = 0.193$ , de tal forma que  $\hat{p}_1 - \hat{p}_2 = 0.280 - 0.193 = 0.087 > 0.070$ , con lo cual concluimos que con la realización tomada rechazamos  $H_0: p_1 - p_2 \leq 0$  al 6% de significancia.

**Conclusión:** con 6% de significancia y la realización obtenida existen evidencias para validar la afirmación del gerente,  $p_1 > p_2$ .

c) Realizamos solo los cálculos necesarios con  $\tilde{p} = \frac{t_1 + t_2}{n_1 + n_2} = \frac{56 + 29}{200 + 150} = 0.2429$

Rechazar  $H_0: p_1 - p_2 \leq 0$ , si:

$$\hat{p}_1 - \hat{p}_2 > p_0 + \Phi^{-1}(1 - \alpha) \sqrt{\tilde{p}\tilde{q} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} = 0 + 1.5548 \sqrt{0.2429(0.7571) \left( \frac{1}{200} + \frac{1}{150} \right)} = 0.072$$

Es decir, rechazar  $H_0: p_1 - p_2 \leq 0$  si  $\hat{p}_1 - \hat{p}_2 > 0.072$ . Como  $\hat{p}_1 - \hat{p}_2 = 0.087 > 0.072$  rechazamos  $H_0: p_1 - p_2 \leq 0$  con 6% de significancia. Así obtuvimos la misma conclusión que en el inciso b). En la figura 4.28 se muestra la partición de la prueba.

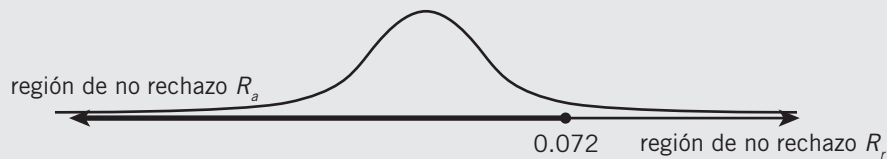


Figura 4.28 Regiones de la prueba del ejemplo 4.31c.

d) Para calcular la potencia de la prueba utilizamos la región de rechazo:

$$\begin{aligned} P(\text{rechazar } H_0 \mid p_1 - p_2 > 0) &= P(\hat{p}_1 - \hat{p}_2 > 0.070 \mid p_1 - p_2 = 0.08) = P\left(Z > \frac{0.070 - 0.080}{\sqrt{\frac{0.28(0.72)}{200} + \frac{0.193(0.807)}{150}}}\right) \\ &= P(Z > -0.22) = 0.5871 \end{aligned}$$

Si se calcula la potencia con el promedio ponderado, resulta:

$$\begin{aligned} P(\text{rechazar } H_0 \mid p_1 - p_2 > 0) &= P(\hat{p}_1 - \hat{p}_2 > 0.072 \mid p_1 - p_2 = 0.08) = P\left(Z > \frac{0.072 - 0.080}{\sqrt{0.2429(0.7571) \left( \frac{1}{200} + \frac{1}{150} \right)}}\right) \\ &= P(Z > -0.173) = 0.5687 \end{aligned}$$

Valor muy próximo al obtenido de 0.5871 con la estimación previa, y que es un poco mayor, esto significa que es mejor la primera prueba que con el promedio ponderado.

#### Ejemplo 4.32 Aplicaciones del teorema 4.10

Dos empresas televisivas compiten por la audiencia a una determinada hora, el director de la empresa 1 afirma que la proporción de televidentes que ve su programa a esa hora excede la proporción de televidentes de la empresa 2 exactamente en 0.20. Para probar la afirmación se realizan dos encuestas independientes, cuyo resultado es que 650 de 1 000 televidentes prefieren la programación de la televisora 1 y que 380 de 800 prefieren la programación de la televisora 2.

- Plantee un contraste de hipótesis adecuado para el problema.
- Al nivel de significancia de 5% pruebe si es válida la afirmación.

c) Calcule la potencia de la prueba, suponga que  $p_1 - p_2 = 0.15$ .

### Solución

a) Se pide una prueba de hipótesis para una diferencia de proporciones de audiencia televisiva de las televisoras 1 y 2. La afirmación que se hace es  $p_1 - p_2 = 0.20$ , así la opuesta será  $p_1 - p_2 \neq 0.20$ . Luego, las hipótesis  $H_0: p_1 - p_2 = 0.20$  y  $H_1: p_1 - p_2 \neq 0.20$ .

b) Si sigue los pasos de la metodología:

i)  $H_0: p_1 - p_2 = 0.20$  contra  $H_1: p_1 - p_2 \neq 0.20$

ii) Nivel de significancia  $\alpha = 0.05$ .

iii) Estamos ante una situación similar a la del inciso c) del teorema 4.10 y requerimos calcular

$$p_0 + \Phi^{-1}(\alpha/2) \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}, p_0 + \Phi^{-1}(1 - \alpha/2) \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \text{ y comparar con la EP } \hat{p}_1 - \hat{p}_2.$$

Si se calculan los componentes  $p_0 = 0.20$ ,  $n_1 = 1000$ ,  $t_1 = 650 \Rightarrow \hat{p}_1 = \frac{650}{1000} = 0.650$  y  $n_2 = 800$ ,  $t_2 = 380$   $\hat{p}_2 = \frac{380}{800} = 0.475$  para  $\alpha = 0.05$  y las tablas porcentuales de la distribución normal estándar,  $\Phi^{-1}(0.025) = -1.96$  y  $\Phi^{-1}(0.975) = 1.96$ . Al final, la regla de decisión.

Rechazar  $H_0: p_1 - p_2 = 0.20$ , si:

$$\hat{p}_1 - \hat{p}_2 < p_0 + \Phi^{-1}(\alpha/2) \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} = 0.20 - 1.96 \sqrt{\frac{0.65(0.35)}{1000} + \frac{0.475(0.525)}{800}} = 0.1545 \text{ o}$$

$$\hat{p}_1 - \hat{p}_2 > p_0 + \Phi^{-1}(1 - \alpha/2) \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} = 0.20 + 1.96 \sqrt{\frac{0.65(0.35)}{1000} + \frac{0.475(0.525)}{800}} = 0.2455$$

Es decir, rechazar  $H_0: p_1 - p_2 = 0.20$  si  $\hat{p}_1 - \hat{p}_2 < 0.1545$  o  $\hat{p}_1 - \hat{p}_2 > 0.2455$ . En la figura 4.29 se muestra la partición de la prueba.



Figura 4.29 Regiones de la prueba del ejemplo 4.32.

iv) Por último, aplicamos la regla de decisión con  $\hat{p}_1 = 0.650$  y  $\hat{p}_2 = 0.475$ , de tal forma que  $\hat{p}_1 - \hat{p}_2 = 0.650 - 0.475 = 0.175 \in [0.1545, 0.2455]$ . Concluimos que con la realización tomada no hay evidencias para rechazar  $H_0: p_1 - p_2 = 0.20$  con 5% de significancia.

**Conclusión:** con 5% de significancia y la realización obtenida no existen evidencias para rechazar la afirmación del director del programa televisivo 1 que  $p_1 - p_2 = 0.20$ .

c) Para calcular la potencia de la prueba utilizamos la región de rechazo:

$$\begin{aligned} P(\text{rechazar } H_0 \mid p_1 - p_2 > 0) &= P\left(\left(\hat{p}_1 - \hat{p}_2 < 0.1545\right) \cup \left(\hat{p}_1 - \hat{p}_2 > 0.2455\right) \mid p_1 - p_2 = 0.15\right) \\ &= P\left(Z < \frac{0.1545 - 0.15}{\sqrt{\frac{0.65(0.35)}{1000} + \frac{0.475(0.525)}{800}}}\right) + P\left(Z > \frac{0.2455 - 0.15}{\sqrt{\frac{0.65(0.35)}{1000} + \frac{0.475(0.525)}{800}}}\right) \\ &= P(Z < 0.19) + P(Z > 4.11) = 0.5754 \end{aligned}$$

## Ejercicios 4.4

---

1. Un fabricante, afirma que más de 30% de los consumidores prefiere su producto. Para realizar una prueba estadística, selecciona una muestra aleatoria de 60 personas y pregunta si lo prefieren o no, de los cuales resulta que 28 entrevistados contestaron que sí. Con esta información.
  - a) Plantee el contraste de hipótesis adecuado al problema.
  - b) A un nivel de significancia de 5% pruebe si se justifica la afirmación del fabricante.
2. Del ejercicio anterior calcule la potencia de la prueba para  $p = \hat{p}$ .
3. El gobierno de la Ciudad de México afirma que la proporción de la población que sufrió algún tipo de robo es menor a 20%. Para probar de manera estadística si es válida la afirmación se seleccionó una muestra aleatoria de 500 ciudadanos, de los cuales 90 dijeron haber sufrido algún tipo de robo.
  - a) Plantee el contraste de hipótesis adecuado al problema.
  - b) A un nivel de significancia de 10% pruebe si se justifica la afirmación del gobierno de la Ciudad de México.
4. Del ejercicio anterior calcule la potencia de la prueba para  $p = \hat{p}$ .
5. La administración del ISSSTE quiere justificar que el servicio ofrecido a personas embarazadas es alto. Para esto afirma que entre 60 y 90% de mujeres en relación de afiliación tuvieron algún cuidado prenatal al entrar en su tercer trimestre de embarazo. Entonces decide comprobar esta afirmación de manera estadística, selecciona una muestra aleatoria de 200 mujeres de esta población y encuentra que 60% tuvo algún cuidado prenatal. Con una prueba de hipótesis adecuada, pruebe 5% de significancia si es válida la afirmación de la administración del ISSSTE.
6. Del ejercicio anterior:
  - a) Una mujer que se encuentra en esta situación y tiene conocimientos estadísticos, afirma que aunque la prueba anterior es válida es mejor hacer la siguiente afirmación: La proporción de mujeres que tuvieron algún cuidado prenatal al entrar en su tercer trimestre de embarazo es superior a 50%. Pruebe con 5% de significancia que también es válida esta afirmación.
  - b) La justificación que ofreció la mujer es que se fijó en las potencias para  $p = \hat{p}$ . Compruébelo.
7. El gerente de una industria de bombillas de luz afirma que la proporción de bombillas que duran menos de 780 horas es de 10%. Para lo cual, el supervisor de control de calidad toma una muestra de 250 bombillas y los prende hasta que dejen de funcionar, con lo que obtiene que 18 bombillas duraron menos de 780 horas.
  - a) Plantee el contraste de hipótesis adecuado al problema.
  - b) A un nivel de significancia de 5% pruebe si se justifica la afirmación del gerente.
8. Del ejercicio anterior calcule la potencia de la prueba para  $p = 0.06$ .
9. Según especialistas que investigan las causas del desempleo en México aseguran que entre 45 y 55% de los trabajadores de cierta zona de la Ciudad de México cambiaron al menos una vez su empleo en los últimos cinco años. Se seleccionó una muestra aleatoria de 200 trabajadores de esta zona y resultó que 90 habían cambiado de trabajo al menos una vez en los últimos cinco años.
  - a) Plantee el contraste de hipótesis adecuado al problema.
  - b) A un nivel de significancia de 10% pruebe si se justifica la afirmación de los investigadores.

### Diferencia de proporciones

10. Según un estudio para conocer la proporción de amas de casa que tienen lavadora en la ciudad, se asegura que ésta excede a la de zonas rurales en más de 0.40. Para probar la afirmación se encuesta a 300 amas de casa de la ciudad, con lo que obtiene que 240 tienen lavadora en su casa, mientras que en la zona rural se entrevistó a 180, de las cuales solo 58 cuentan con una.

- a) Plantee un contraste de hipótesis adecuado para el problema.
- b) Al nivel de significancia de 10% pruebe si es válida la afirmación que se hace del estudio.
11. Del ejercicio anterior, calcule la potencia de la prueba, suponga que  $p_1 - p_2 = 0.50$ .
12. Se afirma que los medicamentos genéricos ( $g$ ) son igual de efectivos que los otros medicamentos ( $m$ ) para cierta enfermedad. En una muestra de 200 personas con este padecimiento, la proporción de personas que consumió el medicamento genérico y para la que resultó efectivo es de 60%, mientras que en otra muestra independiente de 150 personas con este padecimiento se les aplicó el otro medicamento y resultó ser efectivo en 70%.
- a) Plantee un contraste de hipótesis adecuado para el problema.
- b) Al nivel de significancia de 5% pruebe si es válida la afirmación sobre los medicamentos.
13. Del ejercicio anterior calcule la potencia de la prueba, suponga que  $p_g - p_m = 0.08$ .
14. Suponga que en un estudio acerca del uso de internet se observa que en una muestra de 120 alumnos de escuelas particulares, 85% tiene que usar la red para sus trabajos al menos dos veces por semana, mientras que en una muestra de 150 alumnos de las escuelas oficiales 40% usan la red con la misma periodicidad. Con base en esta información se puede decir que la proporción de alumnos de escuelas particulares aventaja en el uso de internet a los alumnos de las escuelas oficiales en más de 30%. Formule un contraste de hipótesis adecuado y pruebe con 5% de significancia.
15. Un sociólogo desea verificar la hipótesis nula de que la proporción de parejas casadas participantes en actividades informales de grupo es la misma en dos comunidades. Dos muestras aleatorias independientes de parejas de las dos comunidades arrojan los resultados de la tabla 4.14.

Tabla 4.14

Comunidad	Tamaño de muestra	Número de parejas participantes en actividades informales de grupo
A	175	88
B	225	101

Utilice un nivel de significancia de 10%.

16. De acuerdo con un estudio que se realizó con 230 familias del sur de la Ciudad de México se obtuvo que 16% de los hogares tienen ingresos totales que se clasifican entre los de nivel económico alto, mientras que en una muestra de 280 familias del norte de la ciudad el porcentaje es de 11%. ¿Con esta información se puede asegurar que la proporción de familias con ingresos de nivel económico alto es mayor en el sur que en el norte de la ciudad? Realice la prueba con 10% de significancia con los tres casos de estimación de la varianza.
17. Del ejercicio anterior calcule la potencia de la prueba para una diferencia de 0.06 (sur menos norte). Compare resultados e indique una conclusión.
-



## Ejercicios de repaso

## Preguntas de autoevaluación

- 4.1 ¿Qué se entiende por prueba de hipótesis?
- 4.2 ¿Qué es la región de rechazo y la región de no rechazo?
- 4.3 ¿Qué significa la prueba más potente de tamaño 0.05?
- 4.4 ¿Qué significa la prueba uniformemente más potente de tamaño 0.10?
- 4.5 ¿Cuál es la filosofía para determinar un contraste de hipótesis?
- 4.6 ¿Cómo definir al contraste de hipótesis?
- 4.7 ¿Qué es un contraste de hipótesis unilateral?
- 4.8 ¿Qué es un contraste de hipótesis bilateral?
- 4.9 ¿Es cierto que las fórmulas que se determinaron para las pruebas de hipótesis de la media se pueden aplicar a cualquier distribución?
- 4.10 ¿En qué situación de las pruebas de hipótesis para la diferencia de medias se pide que las observaciones sean dependientes?
- 4.11 ¿Para qué es necesario que las muestras sean independientes en las pruebas de hipótesis de la razón de varianzas?
- 4.12 ¿Cuál es la prueba de Behrens-Fisher?
- 4.13 ¿Cuál es la prueba de Welch-Aspin?

## Ejercicios complementarios con grado de dificultad uno

- 4.14 ¿Qué es el error tipo I?, ¿cómo se calcula la probabilidad de este tipo de error?
- 4.15 ¿Qué es el error tipo II?, ¿cómo se calcula la probabilidad de este tipo de error?
- 4.16 ¿Qué significa el nivel de significancia?
- 4.17 ¿Qué es la potencia de la prueba?
- 4.18 ¿Qué es el estadístico de prueba?
- 4.19 ¿Qué es la constante crítica?
- 4.20 ¿Cómo se puede disminuir el valor de  $\alpha$  y  $\beta$ , respectivamente?
- 4.21 ¿Cuáles son los pasos que se deben seguir en la metodología para probar una hipótesis estadística?
- 4.22 Estadísticamente, ¿es válido decir que se acepta la hipótesis nula?

## Ejercicios complementarios con grado de dificultad dos

- 4.23 Suponga que la compresión del cemento es una variable aleatoria con distribución  $N(\mu, 14400)$ . Se desarrolló una nueva preparación para cierto tipo de cemento, el comprador afirma que la compresión para la nueva pre-

paración es menor a 5000 kg/cm<sup>2</sup> y para tal efecto establece el siguiente contraste de hipótesis:

$$H_0: \mu \geq 5000$$

$$H_1: \mu < 5000$$

Para verificar su conjetura, el comprador revisa una muestra de tamaño 50 y decide que rechazará la hipótesis nula cuando  $\bar{X} < 4970$ . Suponga que por experiencia se conoce que la desviación estándar poblacional de la compresión del cemento no cambia con respecto a una nueva preparación.

- a) Encuentre la expresión para la función de potencia de la prueba y con la ayuda de algún paquete matemático trace su gráfica y anote sus conclusiones.
- b) Calcule el nivel de significancia de la prueba.

4.24 En el ejercicio anterior:

- a) Encuentre la probabilidad de cometer el error tipo I, cuando la verdadera media a la compresión sea  $\mu = 5010$ .
- b) Evalúe  $\beta$  para la alternativa  $\mu = 4960$ .

4.25 Del ejercicio anterior sobre la compresión del cemento, calcule la potencia de la prueba para el caso en que la verdadera media sea  $\mu = 4950$  e interprete el resultado.

4.26 El encargado de una lavandería afirma que un nuevo quitamanchas es efectivo en menos de 70% de los casos en que se utiliza. Para comprobar esta afirmación se aplicará este producto en 12 manchas elegidas al azar. Si menos de 10 manchas son eliminadas, no se rechazará la hipótesis nula de que  $p \geq 0.7$ ; de otra forma, se concluirá que la afirmación de la lavandería,  $p < 0.7$ , es cierta.

- a) Encuentre la expresión para la función de potencia de la prueba y, con la ayuda de algún paquete matemático, trace su gráfica y escriba sus conclusiones.
- b) Calcule el nivel de significancia de la prueba.

4.27 Del ejercicio anterior:

- a) Evalúe  $\alpha$ , suponga que  $p = 0.8$ .
- b) Obtenga una expresión para la probabilidad del error tipo II.

4.28 Del ejercicio anterior sobre el nuevo quitamanchas:

- a) Evalúe  $\beta$  para la alternativa  $p = 0.65$ .
- b) Calcule la potencia de la prueba para el caso en que la verdadera proporción de artículos defectuosos sea  $p = 0.6$  e interprete el resultado.

4.29 Del ejercicio anterior sobre el nuevo quitamanchas:

- a) Con los resultados encontrados y una realización de 12 manchas a las cuales se les aplica el nuevo producto, con 0 para el caso de no quitar la mancha y 1 cuando si la quita:

1, 1, 0, 1, 1, 1, 0, 0, 1, 1, 0, 1

decida si es válida la afirmación de los dueños de la lavandería.

- b) Considerando el valor de la realización para la estadística de prueba, como el valor del parámetro, calcule la potencia de la prueba.

- 4.30** Un estudiante que realiza su tesis sobre accidentes de trabajo encontró que los accidentes en las fábricas tienen una distribución normal. Sobre los parámetros no conoce los resultados, pero según su estudio afirma que la media de accidentes es menor a 13. Para explicar de manera estadística su afirmación sobre la cantidad promedio de accidentes realiza un muestreo aleatorio de 30 fábricas, de lo que obtiene  $\bar{x} = 10$  y  $s_{n-1} = 9.25$ .
- a) Plantee el contraste de hipótesis apropiado para este problema.
- b) Justifique a un nivel de significancia de 0.05 si de manera estadística es válida la afirmación del estudio sobre la cantidad promedio de accidentes en las fábricas. Suponga que  $\sigma = 8.5$ .
- 4.31** Del ejercicio anterior calcule la potencia de la prueba en el inciso b), suponga que la media poblacional es  $\mu = \bar{x}$ .
- 4.32** Del ejercicio anterior sobre los accidentes de trabajo, repita el inciso b) y la potencia de la prueba sin la suposición de que  $\sigma = 8.5$ .
- 4.33** Del ejercicio anterior sobre los accidentes de trabajo:
- a) Verifique si estadísticamente es correcta la suposición del estudiante de que  $\sigma = 8.5$ , plantee el juego de hipótesis adecuado y utilice  $\alpha = 0.05$ .
- b) Calcule la potencia de la prueba anterior, suponga que  $\sigma = 12$ .
- 4.34** Una muestra aleatoria de los gastos operacionales mensuales de una compañía en  $n = 26$  meses, arrojó un promedio de \$5774 USD con una desviación estándar de \$800 USD, es razonable suponer que la distribución poblacional es normal con una desviación estándar de \$750 USD. Si los administradores de la compañía aseguran que los gastos operacionales medios están por arriba de \$5400 USD.
- a) Plantee el contraste de hipótesis apropiado para este problema.
- b) Justifique a un nivel de significancia de 0.05 si estadísticamente es válida la afirmación de los administradores. Suponga correcta la suposición que la desviación estándar es de \$750 USD.
- 4.35** Del ejercicio anterior calcule la potencia de la prueba, suponga que  $\mu = 5700$ .
- 4.36** Del ejercicio anterior sobre los gastos operacionales:
- a) Verifique si de manera estadística es correcta la suposición  $\sigma = 750$ , plantee el juego de hipótesis adecuado y utilice  $\alpha = 0.10$ .
- b) En caso de que no resulte válido el supuesto, repita el inciso b) con  $\sigma$  desconocida.
- 4.37** Del ejercicio anterior calcule la potencia de la prueba anterior, suponga que  $\sigma = 1000$ .
- 4.38** Los consumidores de cierta marca de cigarrillos aseguran que el producto tiene un contenido promedio de nicotina mayor a 1.2 mg. Para probar esta afirmación se tomó una muestra aleatoria de 41 cigarrillos de dicha marca y se obtuvo un contenido promedio de nicotina de 1.3 mg con una desviación estándar de 0.17 mg. Los fabricantes afirman que la cantidad de nicotina de estos cigarrillos tiene una distribución normal con una varianza de 0.018.
- a) Plantee el contraste de hipótesis apropiado para validar a 0.06 de significancia la suposición de la varianza por parte de los fabricantes.
- b) Calcule la potencia de la prueba en el inciso a), si  $\sigma^2 = 0.03$ .
- 4.39** Del ejercicio anterior:
- a) Plantee el contraste de hipótesis apropiado para la aseveración del contenido promedio de los consumidores.
- b) Justifique a un nivel de significancia de 0.06, si es válida la afirmación de los consumidores. Utilice el resultado del inciso a) del ejercicio anterior.
- c) Calcule la potencia de la prueba en el inciso b) para  $\mu = \bar{x}$ .
- 4.40** Un fabricante de cascos de seguridad para trabajadores de la construcción pretende vender su producto. El fabricante afirma que la fuerza media transmitida por sus cascos es mayor a 1100 lb por arriba del límite oficial que es de 1000, si supone normalidad en la fuerza transmitida por los cascos. El fabricante quiere probar su afirmación, para lo cual tomó una muestra aleatoria de  $n = 4$  cascos y encontró que  $\bar{x} = 1120$  lb y  $s^2 = 2350$  lb<sup>2</sup>.
- a) Plantee el contraste de hipótesis apropiado para este problema.
- b) Justifique a un nivel de significancia de 0.02 si es válida la afirmación del fabricante.
- c. Calcule la potencia de la prueba para  $\mu = \bar{x}$  lb.
- 4.41** Del ejercicio anterior el fabricante de cascos hace otra afirmación con respecto a la varianza de las fuerzas que transmiten los cascos, en la que asegura que es menor a 3500 lb<sup>2</sup>.
- a) Plantee el contraste de hipótesis apropiado para este problema.
- b) Justifique a un nivel de significancia de 0.02 si es válida la afirmación del fabricante.
- 4.42** Del ejercicio anterior, si la verdadera varianza vale  $\sigma^2 = 2350$  calcule la potencia de la prueba
- 4.43** El IPC de la empresa WM se muestra en la tabla 4.15 y tiene una distribución normal durante el año. Suponga que la desviación estándar del IPC durante todo el año es igual a 0.90. Formule el juego de hipótesis adecuado y pruebe dicha suposición a 6% de significancia.
- 4.44** Con base en el resultado del ejercicio anterior, resuelva:
- a) Sus dirigentes aseguran que el IPC en promedio este año será mayor a 35.5. Plantee el contraste de hipótesis apropiado para este problema.

- b) Justifique a un nivel de significancia de 0.06 si de manera estadística es válida la afirmación de los directivos de la empresa.

Tabla 4.15

Fecha	WM
09/06/2013	37.10
09/03/2013	36.99
09/02/2013	37.83
09/01/2013	36.36
08/31/2013	36.17
08/30/2013	35.98
08/27/2013	35.87
08/26/2013	35.68
08/25/2013	35.92
08/24/2013	35.91
08/23/2013	35.29
08/20/2013	34.86
08/19/2013	34.83

- 4.45** Del ejercicio anterior calcule la potencia de la prueba, suponga que  $\mu = 35.4$  para la empresa WM.
- 4.46** Un fabricante de pilas asegura que las que él produce duran en promedio más de 39.5 horas. Para probar la afirmación del fabricante se tomó una muestra aleatoria de nueve pilas, de las que se obtuvieron las duraciones: 39, 41.1, 43, 38, 39.5, 37, 42, 41, 40.1 horas. Suponga que la duración de las pilas se distribuye aproximadamente en forma normal con una desviación estándar de dos horas.
- a) Plantee el contraste de hipótesis apropiado para este problema.
- b) Justifique a un nivel de significancia de 0.05 si es válida la afirmación del fabricante.
- 4.47** Del ejercicio anterior calcule la potencia de la prueba para  $\mu = 39.7$  horas.
- 4.48** Del ejercicio anterior de las pilas, pruebe a 5% de significancia si es válida la suposición del fabricante que  $\sigma = 2$ .
- 4.49** Los resultados del IPC de la empresa CM para una muestra de 71 cotizaciones se agruparon por medio de clases de frecuencia (véase tabla 4.16). Suponga que el IPC de la CM para ese año tiene una distribución aproximadamente normal, con un nivel de significancia de 4%. Pruebe el contraste de hipótesis  $H_0: \mu \geq 12.5$  contra  $H_1: \mu < 12.5$ .

Tabla 4.16

Intervalos de clase	Frecuencias
[11.19, 11.58)	4
[11.58, 11.97)	15
[11.97, 12.36)	20
[12.36, 12.75)	18
[12.75, 13.14)	9
[13.14, 13.53]	5

- 4.50** Los datos que se muestran en la tabla 4.17 son los grados de dureza Brinells obtenidos para muestras independientes de dos aleaciones de magnesio.

Tabla 4.17

Aleación 1	66.3	63.5	64.9	61.8	64.3	64.7	65.1	64.5	68.4	63.2
Aleación 2	71.3	60.4	62.6	63.9	68.8	70.1	64.8	68.9	65.8	66.9

Suponga que provienen de poblaciones aproximadamente normales. Los responsables del proceso afirman que las varianzas poblacionales de ambas aleaciones son iguales. Pruebe a 0.10 de significancia si con las 10 realizaciones es válida la afirmación sobre las varianzas.

- 4.51** Del ejercicio anterior:
- a) Los responsables del proceso afirman que ambas aleaciones son iguales. Plantee un contraste de hipótesis adecuado para esta afirmación.
- b) Al nivel de significancia de 10% pruebe si es válida la afirmación de los responsables.
- 4.52** Repita el ejercicio anterior, considere varianzas desconocidas, pero iguales.
- 4.53** De los dos ejercicios anteriores:
- a) Calcule la potencia de la prueba con  $\mu_1 - \mu_2 = -3$  para ambas pruebas.
- b) ¿Qué conclusiones haría de ambas pruebas y cuál elegiría?
- 4.54** Cierta metal se produce, por lo regular, mediante un proceso estándar. Se desarrolla un nuevo proceso en el que se añade una aleación a la producción del metal. Para cada metal se seleccionan al azar 10 muestras y cada una de éstas se somete a una tensión hasta que se rompe. La tabla 4.18 muestra las tensiones de ruptura de los especímenes en kg/cm<sup>2</sup>.

Tabla 4.18

Proceso nuevo (e)	438	462	478	442	453	445	433	480	448	462
Proceso estándar (n)	457	473	443	372	458	474	419	479	474	449

Si se supone que el muestreo se llevó a cabo en dos distribuciones normales e independientes. Pruebe a 0.10 de significancia si las varianzas poblacionales son iguales.

- 4.55** Del ejercicio anterior, los fabricantes afirman que con la aleación añadida se aumenta en promedio la tensión de ruptura de los metales producidos por los dos procesos más de 2 kg/cm<sup>2</sup>. Plantee el contraste de hipótesis apropiado para probar la afirmación del fabricante a 10% de significancia. Utilice el resultado del ejercicio anterior.
- 4.56** Repita el ejercicio anterior si los fabricantes afirman que en promedio tienen la misma resistencia a la tensión.
- 4.57** De los dos ejercicios anteriores calcule la potencia de cada una de las pruebas si  $\mu_n - \mu_e = \bar{x}_n - \bar{x}_e$  kg/cm<sup>2</sup>. ¿Qué conclusión puede dar de las dos pruebas realizadas?
- 4.58** Después de varios años de analizar los resultados de cálculo diferencial y cálculo integral, se probó que éstos tienen cierta dependencia. De manera que el director de la facultad afirma que las calificaciones de los alumnos en cálculo diferencial son en promedio superiores a las de cálculo integral entre 1 y 2 puntos. Para probar esta afirmación se elige una muestra aleatoria de 10 alumnos y se anotan sus calificaciones (véase tabla 4.19).

Tabla 4.19

Estudiante	1	2	3	4	5	6	7	8	9	10
Cálculo diferencial	7	8	6	5	7	9	8	6	8	7
Cálculo integral	5	6	3	2	6	10	7	2	6	6

Suponga normalidad en las diferencias de las calificaciones y pruebe de manera estadística si es válida la afirmación de la dirección.

- a) Plantee el contraste de hipótesis apropiado para este problema.
- b) Justifique a un nivel de significancia de 0.10 si es válida la afirmación.
- 4.59** Del ejercicio anterior calcule la potencia de la prueba para  $\mu_d = 0.4$  donde  $\mu_d$  representa la media de las calificaciones de cálculo diferencial menos cálculo integral.
- 4.60** En una tienda de cinescopios, los vendedores aseguran que la vida media de los importados ( $I$  o 1) y nacionales ( $N$  o 2) son iguales. Para lo cual toman una muestra aleatoria de 26 cinescopios importados, de la que se obtiene una vida media de 6.5 años; en tanto que la vida media de 20 cinescopios nacionales fue de 6.0 años. Considere que se trata de poblaciones distribuidas de manera normal con desviaciones estándar de 0.4 años para  $I$  y 0.9 años para  $N$ .
- a) Plantee el contraste de hipótesis apropiado para probar la afirmación de los vendedores.
- b) Al nivel de significancia de 0.06 se justifica la afirmación de los vendedores.
- 4.61** Del ejercicio anterior, calcule la potencia de la prueba para  $\mu_I - \mu_N = 0.6$  años.
- 4.62** En el problema anterior de los cinescopios un cliente que dice saber de estadística hace la observación de que las

desviaciones estándar poblacionales tienen una gran diferencia, por tanto, la vida media de los cinescopios de importación debe ser mayor a la de los nacionales, pero no en mucho porque los promedios de vidas medias muestrales de ambos cinescopios no varían de manera considerable. Entonces, afirma que la vida media de los cinescopios importados sí es superior a la de los nacionales, pero en menos de un 0.4 años.

- a) Plantee el contraste de hipótesis apropiado para probar la afirmación de los vendedores.
- b) ¿Al nivel de significancia de 0.06 se justifica la afirmación de los vendedores?
- 4.63** Del ejercicio anterior calcule la potencia de la prueba para  $\mu_I - \mu_N = 0.6$  años.
- 4.64** Los empleados  $A$  ( $x$ ) y  $B$  ( $y$ ) completan el trámite de las cuentas corrientes personales para nuevos clientes. Se asignaron al azar 10 clientes a cada empleado y registraron los tiempos de servicio para cada cliente, de lo que se obtuvo los siguientes resultados:

$$\sum_{i=1}^{10} x_i = 218, \quad \sum_{i=1}^{10} x_i^2 = 4824; \quad \sum_{i=1}^{10} y_i = 246, \quad \sum_{i=1}^{10} y_i^2 = 6140$$

Si se supone que los tiempos de servicio por empleado tienen una distribución normal, pruebe a 0.05 de significancia si es válido suponer varianzas poblacionales iguales.

- 4.65** Del ejercicio anterior, el empleado  $B$  afirma que en promedio es más lento que  $A$  entre 1 y 2 minutos. Plantee el contraste de hipótesis apropiado para la afirmación del empleado  $A$ . Verifique al nivel de significancia de 0.05, si es válida la afirmación del empleado  $A$ . Utilice el resultado del ejercicio anterior.
- 4.66** Del ejercicio anterior calcule la potencia de la prueba para  $\mu_A - \mu_B = -3.5$ .
- 4.67** En la Ciudad de México y el área metropolitana se introdujo un nuevo refresco de cola. Los fabricantes de la bebida afirman que la proporción de consumidores que prefieren su bebida está entre 0.15 y 0.30. Para probar esta afirmación, se seleccionó una muestra aleatoria de 350 ciudadanos, de los cuales 45 contestaron que sí consumen el producto.
- a) Plantee el contraste de hipótesis adecuado al problema.
- b) A un nivel de significancia de 5% pruebe si se justifica la afirmación del fabricante.
- 4.68** Del ejercicio anterior calcule la potencia de la prueba si  $p = 0.10$ .
- 4.69** Debido al alto índice delictivo en la Ciudad de México se elaboró un estudio referente a las tiendas que podrían aceptar tarjetas de débito en los pagos de los consumidores. Se encontró que de una muestra de 150 tiendas, 52% podría hacerlo. Con una prueba de hipótesis adecuada, pruebe con 10% de significancia si es válido suponer que la proporción de tiendas que aceptan las tarjetas de débito es superior a 50%.
- 4.70** Los productores de un nuevo medicamento para tratar cierta enfermedad aseguran que su producto es efectivo

en 60% de los casos. Para evaluar la afirmación de los productores del nuevo medicamento, se administró a 200 pacientes que la padecían. Al término de tres días se habían recuperado 105.

- Plantee el contraste de hipótesis adecuado al problema.
- A un nivel de significancia de 2% pruebe si se justifica la afirmación de los productores.

**4.71** Del ejercicio anterior:

- Si ahora afirman que la medicina es efectiva en más de 50%, compruebe con  $\alpha = 0.02$ .
- Compare ambas respuestas y explique a qué se debe esta diferencia que parece ilógica.

**4.72** Los propietarios de una empresa que fabrican baterías para linterna afirman que la proporción de artículos defectuosos que producen es menor a 5%. Para probar esto se toma una muestra aleatoria de 250 baterías, de lo que se obtienen ocho baterías defectuosas.

- Plantee el contraste de hipótesis adecuado al problema.
- A un nivel de significancia de 10% pruebe si se justifica la afirmación de los propietarios.

**4.73** Un fabricante de rodamientos realizó un muestreo de 640 artículos y encontró que 15% de los artículos producidos por la máquina *A* presentan un defecto menor, mientras que en otra muestra independiente de 760 rodamientos por la máquina *B* solo 8% presentó este tipo de defecto. Los fabricantes afirman que la proporción de artículos defectuosos producidos por la máquina *A* es mayor a la producida por la máquina *B*.

- Plantee un contraste de hipótesis adecuado para el problema.
- Al nivel de significancia de 5% pruebe si es válida la afirmación de los fabricantes.

**4.74** Del ejercicio anterior calcule la potencia de la prueba, suponga que  $p_A - p_B = 0.05$ .

**4.75** La industria cervecera está interesada en comparar dos marcas de cerveza (*A* y *B*). Debido a que sospechan preferencia de la marca *B* sobre la marca *A*, se eligen dos muestras independientes de 200 personas entrevistadas; de las cuales, 116 prefieren la marca *B*; y de otras 150 personas, 78 prefieren la marca *A*. Formule un contraste de hipótesis adecuado y pruebe con 10% de significancia si es válida la sospecha de la industria cervecera.

**4.76** De la fracción de productos defectuosos provenientes de dos líneas de producción se analiza una muestra aleatoria de 1000 unidades provenientes de la línea 1, que tiene 12 defectuosos; mientras que una muestra aleatoria de 1500 unidades provenientes de la línea 2 tiene 20 defectuosos. Con base en esta información se puede decir que la proporción de productos defectuosos en ambas líneas son iguales. Formule un contraste de hipótesis adecuado y pruebe con 2% de significancia.

**4.77** Del ejercicio anterior se puede decir que la diferencia de proporciones de productos defectuosos de la línea 2 me-

nos la línea 1 es menor a 0.02. Formule un contraste de hipótesis adecuado y pruebe con 2% de significancia.

## Ejercicios complementarios con grado de dificultad tres

**4.78** El director de la SCT afirma que el número de accidentes que ocurren en un crucero determinado de la Ciudad de México es mayor a dos por semana, para lo cual pide revisar de manera aleatoria 10 reportes semanales y escriba las hipótesis  $H_0: \lambda_T \leq 21$  y  $H_1: \lambda_T > 21$ , en donde  $\lambda_T$  es el número de accidentes en el crucero en 10 semanas y  $T$  la variable aleatoria que representa el total de accidentes durante las 10 semanas analizadas. Para tomar decisiones, el director establece la región crítica para  $T > 23$  y define las variables aleatorias  $X_i$  cantidad de accidentes que ocurrieron en el crucero en la semana  $i$  para  $i = 1, 2, \dots, 10$ , por último, suponga que las variables  $X_i$  tienen una distribución de Poisson con  $\lambda = 2$  accidentes por semana.

- Evalúe  $\alpha$  suponiendo  $\lambda_T = 21$ .
- Evalúe  $\beta$  para la alternativa  $\lambda_T = 25$ .

**4.79** Del ejercicio anterior calcule la potencia de la prueba si  $\lambda_T = 28$ .

**4.80** Suponga que tenemos una población con distribución normal de la que conocemos su varianza igual a  $\sigma_0^2$  y el contraste de hipótesis  $H_0: \mu \geq 54$  contra  $H_1: \mu < 54$ . Además, se estableció la región de rechazo para  $\bar{x} < a$ . ¿Cuál debe ser el valor crítico de la prueba,  $a$ , si la probabilidad de cometer un error tipo I, cuando  $\mu = 60$ , es de 0.05 y una probabilidad de error tipo II de 0.01 cuando  $\mu = 50$ ?

**4.81** Suponga que tenemos una población con distribución normal de la que conocemos su varianza igual a  $\sigma_0^2$  y el contraste de hipótesis  $H_0: \mu \leq \mu_0$  contra  $H_1: \mu > \mu_0$ . Además, se estableció la región de rechazo  $\bar{x} < a$ . Deduzca una expresión para determinar el valor crítico  $a$  y el tamaño de la muestra, cuando se conoce el nivel de significancia  $\alpha$  y la probabilidad de cometer un error tipo II  $\beta$  cuando  $\mu = \mu_1$ , con  $\mu > \mu_1$ .

**4.82** Sea una observación simple de una distribución beta,  $Beta(\theta, 1)$ , donde  $\theta > 0$ . Suponga el contraste de hipótesis:

$$H_0: \theta \geq 4$$

$$H_1: \theta < 4$$

Debido a que se trata de una observación y  $f(x; \theta) = \theta x^{\theta-1} I_{(0,1)}(x)$  podemos considerar a la estadística de prueba  $T(X) = X$ . La región de rechazo se establece para  $x < 0.5$ .

- Encuentre la expresión para la función de potencia de la prueba y con la ayuda de algún paquete matemático trace su gráfica y escriba sus conclusiones.
- Calcule el nivel de significancia de la prueba.

**4.83** Del ejercicio anterior:

- Obtenga una expresión para la probabilidad del error tipo II.



- b) Evalúe  $\beta$  para la alternativa  $\theta = 3.5$ .
- c) Calcule la potencia de la prueba para el caso en que  $\theta = 3$ .

**4.84** Un geólogo que pretendía estudiar el movimiento de los cambios relativos en la corteza terrestre en un sitio particular, en un intento por determinar el ángulo medio de las fracturas, eligió  $n = 51$  fracturas y encontró que la media y la desviación estándar muestral eran  $41.8^\circ$  y  $13.2^\circ$ , respectivamente. Con los resultados de su estudio afirma que los ángulos medios de las fracturas de la corteza terrestre en el sitio estudiado son menores a  $43^\circ$ . Suponga que sus ángulos se distribuyen de manera normal. Plantee el contraste de hipótesis apropiado para este problema y aproxime el valor mínimo de  $\alpha$  que se requiere para que sea válida la afirmación.

**4.85** En una empresa, el supervisor afirma que el tiempo promedio de armado de un producto es menor en los hombres ( $h$ ) que en las mujeres ( $m$ ), pero la variancia de los tiempos para las mujeres es menor que la de los hombres. Además, se sabe que la distribución de tiempo, tanto para hombres como para mujeres es aproximadamente normal. Para probar de manera estadística la afirmación del supervisor se extraen muestras aleatorias de tiempos para 11 hombres y 14 mujeres, de lo que se obtienen los siguientes valores  $\bar{x}_h = 35$ ,  $s_h = 6.1$ ,  $\bar{x}_m = 40$  y  $s_m = 5.3$ . Con respecto al material de esta sección, qué pruebas de hipótesis tendrá que formular el supervisor si desea verificar con 5% de significancia si son válidas sus afirmaciones. Formule las hipótesis convenientes y resuelva.

## Proyectos de la unidad 4

- I. Para averiguar si un nuevo suero detendrá o no la leucemia, se seleccionan 18 ratones que alcanzaron un estado avanzado de la enfermedad, nueve de los cuales reciben el tratamiento. Los tiempos de sobrevivencia en años desde que se inició el experimento se aprecian en la tabla 4.20.

Tabla 4.20

Con tratamiento (C)	0.8	3.2	2.7	5.2	3.7	4.4	5.3	3.4	2.6
Sin tratamiento (S)	1.9	2.1	2.6	4.5	2.2	2.1	1.2	2.8	0.8

Este es un problema típico que requiere realizar una buena prueba, por lo cual se harán varias para que decida cuál le conviene elegir y de esta forma realizar una mejor aseveración en el artículo escrito sobre la enfermedad. Suponga que los tiempos de sobrevivencia siguen una distribución normal.

- a) Primero realice una prueba de igualdad de varianzas con 5% de significancia, después con el resultado obtenido realice las siguientes pruebas.
- b) Se afirma que el tiempo de vida media de sobrevivencia con y sin tratamiento es igual. Realice su comprobación con 5% de significancia.
- c) Se afirma que el tiempo de vida media de sobrevivencia con tratamiento es mayor que sin éste. Realice su comprobación con 5% de significancia.

Después de analizar la media y varianza muestrales las afirmaciones cambian a:

- a) Se afirma que el tiempo de vida media de sobrevivencia con tratamiento es mayor que sin tratamiento entre 0.5 y 1 años. Realice su comprobación con 5% de significancia.
- b) Se afirma que el tiempo de vida media de sobrevivencia con tratamiento es mayor que sin éste en más de 0.2 años. Realice su comprobación con 5% de significancia.
- c) Después de haber tomado varias muestras como las del experimento, los investigadores consideran que  $\mu_C - \mu_S = 1.5$  años. Calcule la potencia en cada prueba si esto fuera cierto.
- d) ¿Cuál de todas las pruebas elegiría? Justifique su respuesta.

- II. Dos empresas televisivas compiten por la audiencia a una determinada hora. Para llevar a cabo un estudio estadístico sobre cuál tienen mayor audiencia se realizan dos encuestas independientes, cuyos resultados arrojan que 620 de 900 televidentes prefieren la programación de la televisora 1 y que 480 de 1 100 prefieren la programación de la televisora 2. Se realizan las siguientes afirmaciones y resulta una situación que el director de la empresa 1 no entiende.

*Afirmación 1:* El director de la empresa 1 asegura que la proporción de televidentes que ve su programa a esa hora excede la proporción de televidentes de la empresa 2 de 0.1 a 1.

- a) Plantee un contraste de hipótesis adecuado para el problema.
- b) Al nivel de significancia de 5% pruebe si es válida la afirmación.

*Afirmación 2:* El director de la empresa 1 asegura que la proporción de televidentes que ve su programa a esa hora excede la proporción de televidentes de la empresa 2 en más de 0.25.

- a) Plantee un contraste de hipótesis adecuado para el problema.
- b) Al nivel de significancia de 5% pruebe si es válida la afirmación.
- c) Después de resolver ambos incisos puede ayudar a explicar qué pasó.

# Pruebas de bondad de ajuste

UNIDAD  
**5**



## Competencias específicas a desarrollar

- Identificar y aplicar los conceptos de las pruebas de bondad de ajuste.
- Establecer cuál es la metodología aplicable a una prueba de bondad de ajuste.
- Identificar y aplicar los conceptos de una prueba no paramétrica.

## ¿Qué sabes?

- ¿Sabes cómo determinar el tipo de distribución del que proviene un conjunto de datos?
- ¿Qué es una prueba de bondad de ajuste?
- ¿Conoces la técnica Q-Q para normalidad?
- ¿Qué es una prueba no paramétrica?



## Introducción

En las unidades 3 y 4 revisamos algunos métodos de estimación de parámetros por medio de estimadores puntuales, intervalos de confianza y pruebas de hipótesis. En los métodos revisados utilizamos el hecho de que los datos provenían de una distribución normal, ahora surgen las preguntas: ¿cómo saber cuándo un conjunto de datos proviene de una? En general, ¿cómo determinar el tipo de distribución del que proviene un conjunto de datos?

El problema para establecer la procedencia de los datos,  $x_1, x_2, \dots, x_n$ , se puede resolver formulando una hipótesis estadística, donde:

$$\begin{aligned} H_0 & \text{ los datos } x_1, x_2, \dots, x_n \text{ tienen una distribución } F. \\ H_1 & \text{ los datos } x_1, x_2, \dots, x_n \text{ no tienen una distribución } F. \end{aligned} \quad (5.1)$$

Podemos observar que en este contraste de hipótesis no tenemos puntos de comparación determinados, como en el caso de la metodología estudiada para los parámetros de la distribución normal, en la cual, la hipótesis alterna era el conjunto de valores reales que se contraponen a los valores del parámetro de la hipótesis nula.

En el caso de las hipótesis formuladas en (5.1), podemos observar que la hipótesis alterna es mucho más general que en el caso de los parámetros, ya que no se indica contra qué distribución se contraponen la comparación. Este hecho dificulta de manera considerable la prueba para el contraste de hipótesis (5.1), resultando un problema mucho más complejo que las hipótesis revisadas en la unidad previa.

Por estas razones existe una gran gama de pruebas de hipótesis para comprobar (5.1), a éstas se les llama de bondad de ajuste. En esta unidad revisaremos cuatro. Dos de ellas serán explicadas a detalle, una de forma gráfica llamada prueba Q-Q, y la otra mediante las clases de frecuencia, *ji* cuadrada. En las dos pruebas restantes, Kolmogorov-Smirnov y Anderson-Darling, vamos a bosquejar su estadístico de prueba con base en la función de distribución acumulada, pero su uso lo ilustraremos por medio del paquete Minitab.

## 5.1 Pruebas de bondad de ajuste de forma gráfica

En la estadística inferencial, como en cualquier ciencia de tipo cuantitativo, en la solución de problemas reales contamos con fórmulas para llevar a cabo el cálculo de las variables o parámetros desconocidos. Pero antes de aplicarlas, el primer problema que el investigador debe considerar se refiere a las condiciones que deben satisfacer los datos para que la aplicación de las fórmulas sea correcta. Así, en general, un estudioso de la estadística o de sus aplicaciones que se encuentre ante un conjunto de datos tomados de una muestra afronta su primer problema en determinar a qué distribución corresponden los datos.

El problema general para la determinación de la distribución de procedencia de los datos es tan antiguo como complejo y en estadística se le suele llamar: *Prueba de bondad de ajuste*. El planteamiento general está relacionado con las llamadas pruebas de hipótesis, tema que se trató en la unidad previa.

### Cuantiles

En muchas aplicaciones, al tener un conjunto de datos requerimos conocer el valor por debajo del cual está una fracción de los datos. Por ejemplo, al realizar un examen a un grupo de 30 personas, suponga que queremos conocer la calificación debajo de la que se encuentra 40% de los alumnos. El valor que buscamos se define a continuación.

Dado un conjunto de datos, se llama **C cuantil** a la  $C_p$ , que representa el número para el cual la fracción  $C \in [0, 1]$  de los valores son menores o iguales que éste.

A continuación, se proporciona una regla para buscar al  $C$  cuantil.

Para calcular el  $C$  cuantil de un conjunto de datos  $x_1, x_2, \dots, x_n$  ( $n$  datos no agrupados) podemos seguir los siguientes pasos:

1. Primero ordenamos los datos en forma no decreciente  $\tilde{x}_1 \leq \tilde{x}_2 \leq \dots \leq \tilde{x}_n$ .
2. En segundo lugar, determinamos el valor de la fracción  $C$  de los  $n$  datos; es decir, calcular  $\tilde{c} = nC$ .
3. Por último, dependiendo del valor de  $\tilde{c}$ , resulta:

a) Si la cantidad anterior es entera, entonces  $C_{\text{cuantil}} = \frac{\tilde{x}_{\tilde{c}} + \tilde{x}_{\tilde{c}+1}}{2}$ .

b) Si  $\tilde{c}$  no es entero, entonces  $C_{\text{cuantil}} = \tilde{x}_{[\tilde{c}]+1}$ .

Donde,  $[\tilde{c}]$  representa a la parte entera de  $\tilde{c}$ . Por ejemplo, si  $\tilde{c} = 24.7$ ,  $[\tilde{c}] = 24$ ;  $\tilde{c} = 24.2$ ,  $[\tilde{c}] = 24$ .

Para entender mejor la búsqueda de los cuantiles, revisemos el ejemplo 5.1.

### Ejemplo 5.1 Cuantiles

Sean las calificaciones de 20 estudiantes 45, 69, 79, 83, 38, 27, 98, 100, 84, 79, 67, 84, 92, 35, 56, 69, 47, 95, 100, 86, calcule:

- a) El cuantil 0.65 de la distribución de las calificaciones.
- b) El cuantil 0.42 de la distribución de las calificaciones.

#### Solución

- a) Los datos originales se muestran en la tabla 5.1.

Tabla 5.1

$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$	$x_8$	$x_9$	$x_{10}$
45	69	79	83	38	27	98	100	84	79
$x_{11}$	$x_{12}$	$x_{13}$	$x_{14}$	$x_{15}$	$x_{16}$	$x_{17}$	$x_{18}$	$x_{19}$	$x_{20}$
67	84	92	35	56	69	47	95	100	86

Primero, ordenamos los datos en forma no decreciente y resulta:

27, 35, 38, 45, 47, 56, 67, 69, 69, 79, 79, 83, 84, 84, 86, 92, 95, 98, 100, 100

Tabla 5.2

$\tilde{x}_1$	$\tilde{x}_2$	$\tilde{x}_3$	$\tilde{x}_4$	$\tilde{x}_5$	$\tilde{x}_6$	$\tilde{x}_7$	$\tilde{x}_8$	$\tilde{x}_9$	$\tilde{x}_{10}$
27	35	38	45	47	56	67	69	69	79
$\tilde{x}_{11}$	$\tilde{x}_{12}$	$\tilde{x}_{13}$	$\tilde{x}_{14}$	$\tilde{x}_{15}$	$\tilde{x}_{16}$	$\tilde{x}_{17}$	$\tilde{x}_{18}$	$\tilde{x}_{19}$	$\tilde{x}_{20}$
79	83	84	84	86	92	95	98	100	100

Ahora, se calcula  $\tilde{c} = nC = 20 \times 0.65 = 13$ .

Del inciso anterior  $\tilde{c}$  resultó entero, luego:

$$C_{0.65} = \frac{\tilde{x}_{\tilde{c}} + \tilde{x}_{\tilde{c}+1}}{2} = \frac{\tilde{x}_{13} + \tilde{x}_{13+1}}{2} = \frac{\tilde{x}_{13} + \tilde{x}_{14}}{2} = \frac{84 + 84}{2} = 84$$

Esto significa que 65% de las calificaciones de los 20 estudiantes están por debajo de 84.

- b) De los resultados anteriores, falta calcular  $\tilde{c} = nC = 20 \times 0.42 = 8.4$ .

Del inciso anterior  $\tilde{c}$  no es entero, luego  $C_{0.42} = \tilde{x}_{[\tilde{c}]+1} = \tilde{x}_{8+1} = \tilde{x}_9 = 69$ .

Esto significa que 42% de las calificaciones de los 20 estudiantes están por debajo de 69.

## Técnica gráfica Q-Q para una prueba de ajuste de distribuciones

Sea  $x_1, x_2, \dots, x_n$  un conjunto de datos provenientes de una muestra tomada de una población sobre la que se desconoce su distribución. Por medio de los conceptos estudiados durante la unidad 1, podemos llevar a cabo una prueba gráfica del ajuste de curvas de la siguiente forma.

1. En caso de tener una gran cantidad de datos, primero se debe descomponer en clases de frecuencia y después se traza un histograma de las clases con la finalidad de identificar y proponer una posible distribución de los datos, según las distribuciones teóricas vistas en un curso de probabilidad.
2. Se ordenan los datos  $x_1, x_2, \dots, x_n$  en forma no decreciente, denotamos por  $y_1 \leq y_2 \leq \dots \leq y_n$  al  $y_i$  cuantil  $i/n$  de la muestra, cuya fracción corresponde a la probabilidad estimada para la variable  $X$ , que representa a los datos y cuya distribución desconocemos.
3. Calculamos los cuantiles teóricos (según la distribución de la que se cree provienen los datos), que denotamos por  $q_1, q_2, \dots, q_n$ , correspondientes a los cuantiles muestrales  $y_1, y_2, \dots, y_n$ . Es decir,  $q_1 \leq q_2 \leq \dots \leq q_n$  son tales que  $P(X \leq q_i) = (i - 0.5)/n$ , para  $i = 1, 2, \dots, n$ .

En la práctica se acostumbra tomar las fracciones  $(i - 0.5)/n$  para definir los cuantiles teóricos (de esta forma se asegura que no resulte 0 o 1, solo una aproximación).

4. Trazamos la gráfica entre los cuantiles teóricos ( $q_1, q_2, \dots, q_n$ ) y muestrales ( $y_1, y_2, \dots, y_n$ ). Ahora, solo falta concluir si la distribución propuesta para  $X$ , con la que se calcularon los cuantiles teóricos, es válida. La conclusión se basa en la gráfica cuantil (teórico eje de las abscisas) contra cuantil (muestral eje de las ordenadas); si la gráfica se asemeja a una línea recta, entonces se dice que los datos sí provienen de la distribución teórica propuesta.

Podemos observar que con la construcción de la gráfica cuantil contra cuantil se deriva el nombre de la técnica, Q-Q (quantile-quantile).

## Ejemplo de la técnica gráfica Q-Q para una prueba de normalidad

Sugerimos revisar las pruebas de Shapiro Wilk, ji cuadrada, Kolmogorov Smirnov, etcétera. En general, existen más de 60 pruebas diferentes reconocidas para probar la normalidad, aunque quizá la más socorrida por los usuarios es la prueba de Shapiro Wilk.

La mayoría de los métodos estadísticos clásicos están diseñados para datos que tienen un comportamiento normal. Por este motivo, se tiene una gran cantidad de pruebas de bondad de ajuste para verificar si los datos muestrales cumplen la normalidad. En este caso, emplearemos la técnica Q-Q, explicada en la subsección anterior. Pero antes de aplicarla, se puede apreciar que en el caso de la distribución normal se tiene un mayor conocimiento del resultado posible.

La técnica Q-Q para normalidad consiste en los cuatro pasos mencionados arriba, no obstante, según el resultado de la gráfica cuantil contra cuantil del paso 4, se pueden hacer conclusiones más completas. Por ejemplo, en el paso 4, puede resultar alguno de los siguientes casos mostrados en la figura 5.1.

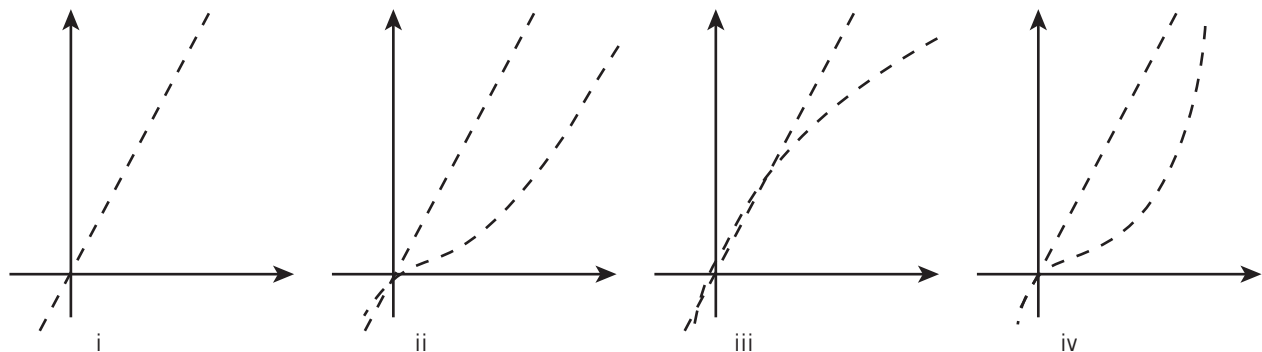


Figura 5.1 Posibles resultados de la técnica Q-Q para la prueba de normalidad.

Para emplear el método Q-Q se propone el siguiente ejemplo.

## Ejemplos 5.2 Método Q-Q

Con los datos de la tabla 5.3 y la técnica gráfica Q-Q, determine si existen evidencias de normalidad en su distribución.

Tabla 5.3

10.10	4.60	3.23	8.75	10.33	5.75	7.38	2.65	7.08	9.14	3.84	6.98
6.55	8.43	-0.71	4.81	2.11	6.52	9.81	9.85	12.12	4.50	5.31	13.15
6.29	9.03	7.83	5.94	11.11	9.02	5.56	6.00	6.04	4.95	7.93	13.32
12.63	7.15	8.41	11.77	3.32	11.78	11.28	10.48	8.93	8.19	10.84	10.18
10.42	14.62	8.53	11.14	6.27	11.94	8.81	5.93	5.19	12.39	13.43	10.04

**Solución**

En este caso seguimos los pasos propuestos.

1. Distribuimos los 60 datos en seis clases de frecuencia y resulta:

Clase 1,  $[-0.71, 1.845]$  con frecuencia 1 y marca de clase 0.5675.

Clase 2,  $(1.845, 4.400]$  con frecuencia 5 y marca de clase 3.1225

Clase 3,  $(4.400, 6.955]$  con frecuencia 16 y marca de clase 5.6775

Clase 4,  $(6.955, 9.510]$  con frecuencia 16 y marca de clase 8.2325

Clase 5,  $(9.510, 12.060]$  con frecuencia 15 y marca de clase 10.7875

Clase 6,  $(12.060, 14.620]$  con frecuencia 7 y marca de clase 13.3425

De las frecuencias con sus marcas de clase y el histograma correspondiente podemos apreciar que sí existe cierto comportamiento normal, con media y desviación estándar que se aproximan a las muestrales 8.15 y 3.19, respectivamente (véase figura 5.2).

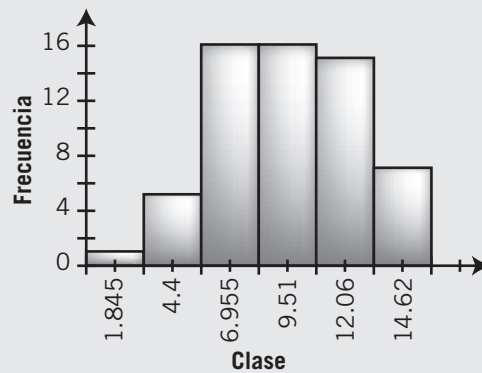


Figura 5.2 Distribución por clases de frecuencia de los 60 datos del ejemplo 5.1.

2. Ordenamos los datos en forma no decreciente  $y_1 \leq y_2 \leq \dots \leq y_n$  (véase tabla 5.4).

Tabla 5.4

-0.71	2.11	2.65	3.23	3.32	3.84	4.5	4.6	4.81	4.95	5.19	5.31
5.56	5.75	5.93	5.94	6.00	6.04	6.27	6.29	6.52	6.55	6.98	7.08
7.15	7.38	7.83	7.93	8.19	8.41	8.43	8.53	8.75	8.81	8.93	9.02
9.03	9.14	9.81	9.85	10.04	10.1	10.18	10.33	10.42	10.48	10.84	11.11
11.14	11.28	11.77	11.78	11.94	12.12	12.39	12.63	13.15	13.32	13.43	14.62

Calculamos las fracciones  $(i - 0.5)/n$  con  $i = 1, 2, \dots, 60$  y  $n = 60$ , para sus correspondientes cuantiles teóricos (véase tabla 5.5).

**Tabla 5.5**

0.008	0.025	0.042	0.058	0.075	0.092	0.108	0.125	0.142	0.158	0.175	0.192
0.208	0.225	0.242	0.258	0.275	0.292	0.308	0.325	0.342	0.358	0.375	0.392
0.408	0.425	0.442	0.458	0.475	0.492	0.508	0.525	0.542	0.558	0.575	0.592
0.608	0.625	0.642	0.658	0.675	0.692	0.708	0.725	0.742	0.758	0.775	0.792
0.808	0.825	0.842	0.858	0.875	0.892	0.908	0.925	0.942	0.958	0.975	0.992

3. Calculamos los cuantiles teóricos  $q_1, q_2, \dots, q_n$  correspondientes a las fracciones anteriores; si se supone una distribución normal con media  $\mu \approx 8.15$  y desviación estándar  $\sigma \approx 3.19$ . Es decir:

$$F(q_i) = \frac{1}{3.19\sqrt{2\pi}} \int_{-\infty}^{q_i} e^{-\frac{(w-8.15)^2}{2(3.19)^2}} dw = \frac{i - 0.5}{n}$$

Estandarizando, tenemos:

$$F(q_i) = \Phi\left(\frac{q_i - 8.15}{3.19}\right) = \frac{i - 0.5}{n}$$

Al despejar a  $q_i$  se tiene:

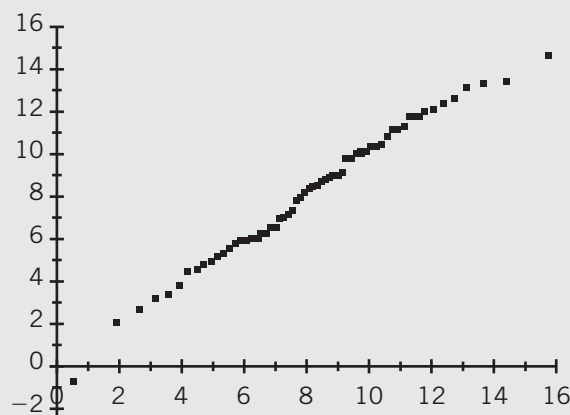
$$q_i = 8.15 + 3.19 \times \Phi^{-1}\left(\frac{i - 0.5}{n}\right)$$

Así, resulta que los cuantiles teóricos  $q_i$  serían (véase tabla 5.6):

**Tabla 5.6**

0.513	1.898	2.625	3.145	3.556	3.904	4.210	4.482	4.727	4.957	5.167	5.368
5.560	5.742	5.914	6.083	6.242	6.399	6.552	6.702	6.848	6.992	7.132	7.273
7.410	7.547	7.681	7.815	7.949	8.083	8.217	8.351	8.485	8.619	8.753	8.890
9.027	9.168	9.308	9.452	9.598	9.748	9.901	10.058	10.217	10.386	10.558	10.740
10.932	11.133	11.343	11.573	11.819	12.090	12.396	12.744	13.155	13.675	14.402	15.787

4. Por último, construimos la gráfica Q-Q, con las parejas de cuantiles  $(q_i, y_i)$ ,  $i = 1, 2, \dots, n$  (véase figura 5.3).



**Figura 5.3** Gráfica Q-Q para los datos muestrales.

Como podemos apreciar, la gráfica Q-Q sí se asemeja a una línea recta. Entonces, se dice que los datos originales sí proporcionan evidencias de normalidad.

## Técnica analítica Q-Q para una prueba de normalidad

La prueba de normalidad para datos es muy importante debido a que muchas de las fórmulas de la metodología que revisamos en el texto están basadas justo para datos normales o aproximadamente normales. Por esta razón, estudiamos otra técnica para probar normalidad de los datos que también se basa en los cuantiles, pero que no recurre a las gráficas y se le da el nombre de técnica analítica Q-Q.

Esta técnica consiste en calcular los cuantiles muestrales,  $y_1 \leq y_2 \leq \dots \leq y_n$ , y sus correspondientes cuantiles teóricos,  $q_1 \leq q_2 \leq \dots \leq q_n$ , tal y como se hizo en la subsección anterior. Luego, con dichos cuantiles se calcula su coeficiente de correlación

$$r_Q = \frac{\sum_{i=1}^n (y_i - \bar{y})(q_i - \bar{q})}{\sqrt{\left(\sum_{i=1}^n (y_i - \bar{y})^2\right)\left(\sum_{i=1}^n (q_i - \bar{q})^2\right)}}$$

Enseguida, el valor obtenido se compara con los valores de la tabla 5.7, llamados críticos para la prueba de normalidad con base en el coeficiente de correlación de los cuantiles de la gráfica Q-Q.

**Tabla 5.7** Valores de  $r^*$  que se emplean en la técnica analítica Q-Q para probar normalidad de los datos

Tamaño de la muestra	Nivel de significación de la prueba		
	0.01	0.05	0.10
10	0.880	0.918	0.935
15	0.911	0.938	0.951
20	0.929	0.950	0.960
25	0.941	0.958	0.966
30	0.949	0.964	0.971
35	0.955	0.968	0.974
40	0.960	0.972	0.977
50	0.966	0.976	0.981
60	0.971	0.980	0.984
75	0.976	0.984	0.987
100	0.981	0.986	0.989
150	0.987	0.991	0.992
200	0.990	0.993	0.994

Por último, la conclusión se lleva a cabo con base en la regla:

Los datos muestrales  $x_1, x_2, \dots, x_n$  dan evidencia de normalidad; por el método analítico Q-Q, al nivel de significancia de la prueba elegida (0.01, 0.05 o 0.10), si  $r_Q > r^*$ , en donde  $r^*$  es el valor de la tabla 5.7, para el tamaño de la muestra y nivel de significancia de la prueba. En caso contrario, se dice que los datos no dan evidencias para considerar que provienen de una distribución normal.

### Ejemplos 5.3 Prueba de normalidad

Con los datos del ejemplo anterior y la técnica analítica Q-Q determine si existen evidencias de normalidad en dicho conjunto de datos.

**Solución**

En el ejemplo anterior se realizaron varios cálculos que también debemos llevar a cabo aquí.

1. Calculamos los cuantiles muestrales y teóricos, lo cual se realizó en el ejemplo anterior.
2. Calculamos el coeficiente de correlación de dichas parejas de cuantiles. En este ejemplo, se obtiene al realizar los cálculos  $r_Q = 0.9943$ .
3. Por último, concluimos con base en la regla anterior.

En este caso, para un tamaño de muestra  $n = 60$  se tienen los valores de  $r^* = 0.971$ ,  $r^* = 0.980$  y  $r^* = 0.984$  para los niveles de significancia de la prueba 0.01, 0.05 y 0.10, respectivamente. Por tanto, en cualquiera de los casos  $r_Q > r^*$  y se concluye que los datos, según el método analítico Q-Q sí presentan evidencias de que provienen de una población normal, a los niveles de significancia de la prueba 0.01, 0.05 y 0.10.

Por último, cuando el tamaño de la muestra no está en la tabla 5.7 se puede interpolar su valor.

## 5.2 Prueba de bondad de ajuste ji cuadrada

Se ha comentado que el problema de las pruebas de bondad de ajuste es tan importante como complejo; en su solución se pueden aplicar diferentes tipos de pruebas que, de acuerdo con las observaciones, una prueba puede ser preferida sobre otra, ya que presenta mayor potencia. No existe una prueba de bondad de ajuste que sea mejor a todas las demás para cualquier tipo de observaciones, por esta razón es indispensable que se consideren otro tipo de pruebas, en esta sección revisaremos una de tipo no paramétrico.

Otra prueba de bondad de ajuste diferente a la Q-Q que puede ser aplicada a distribuciones, tanto discretas como continuas, es la prueba ji cuadrada, la cual se basa en una comparación de las funciones de densidad de las observaciones con la densidad teórica propuesta.

### Metodología de la prueba ji cuadrada

Suponga que se desea conocer de qué distribución provienen las  $n$  observaciones  $x_1, x_2, \dots, x_n$  que se obtuvieron de un muestreo aleatorio.

**Paso 1.** Descomponer las  $n$  observaciones en  $s$  clases de frecuencia con extremos derechos  $t_i$  para  $i = 1, 2, \dots, s$  y se traza el histograma por clases de frecuencia  $[t_0, t_1], (t_1, t_2], \dots, (t_{s-1}, t_s]$ . Sean las frecuencias observadas por clase  $O_1, O_2, \dots, O_s$ .

**Paso 2.** Con base en el histograma de frecuencias y el conocimiento de las funciones de densidad, se propone una función de densidad que se asemeje al histograma de frecuencias. Sea la función de  $f(x)$ , con esta función se calculan las frecuencias esperadas por clase

$$n_i = n \int_{t_{i-1}}^{t_i} f(x) dx \quad \text{para } i = 1, 2, \dots, s$$

**Paso 3.** En caso de resultar frecuencias esperadas  $n_i < 5$ , deben agruparse las frecuencias esperadas contiguas de manera que la suma de ellas no sea menor a 5. Denotemos por  $m$  a la cantidad de clases, con frecuencias esperadas mayores o iguales a 5. Entonces tenemos las frecuencias esperadas  $n_i^* \geq 5$  para  $i = 1, 2, \dots, m$  ( $m \leq s$ ) y sus correspondientes frecuencias observadas  $O_1^*, O_2^*, \dots, O_m^*$ , calculamos el estadístico de prueba:

$$\chi_{cal}^2 = \sum_{i=1}^m \frac{(O_i^* - n_i^*)^2}{n_i^*}$$

**Paso 4.** Con el nivel de significancia deseado,  $\alpha$ , para la prueba calculamos el valor de la distribución ji cuadrada con  $m - k - 1$  grados de libertad cuya área derecha sea igual a  $\alpha$ ,  $k$  representa la cantidad de paráme-



tros de la función de densidad propuesta. Sea este valor  $\chi_{m-k-1, \alpha}^2$  y lo comparamos con  $\chi_{cal}^2$ , aplicando la regla de decisión:

Rechazar  $H_0$ :  $x$ 's tienen un comportamiento  $f$ , al nivel de significancia, si  $\chi_{cal}^2 > \chi_{m-k-1}^2$ .

En caso contrario, se concluye que, con la muestra presentada y un nivel de significancia  $\alpha$ , no existen evidencias para rechazar  $H_0$  (véase figura 5.4).

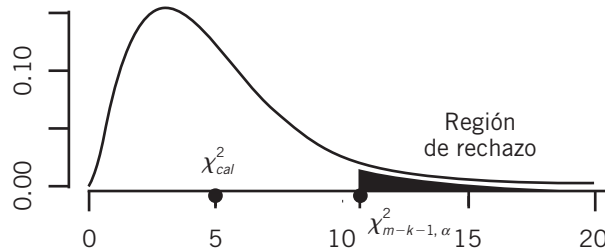


Figura 5.4 Valores calculados y teóricos de la ji cuadrada, caso de no rechazo.

### Ejemplo 5.4 Prueba ji cuadrada

Con los datos del ejemplo 5.2 y la prueba ji cuadrada, determine a un nivel de significancia de 5% si existen evidencias de normalidad en su distribución.

#### Solución

**Paso 1.** En este caso primero identificamos las variables con sus valores del método. La cantidad de datos es  $n = 60$  que se descompusieron en  $s = 6$  clases de frecuencia.

Clase 1,  $[-0.71, 1.845]$  con frecuencia 1

Clase 2,  $(1.845, 4.400]$  con frecuencia 5

Clase 3,  $(4.400, 6.955]$  con frecuencia 16

Clase 4,  $(6.955, 9.510]$  con frecuencia 16

Clase 5,  $(9.510, 12.060]$  con frecuencia 15

Clase 6,  $(12.060, 14.620]$  con frecuencia 7

**Paso 2.** Con base en el histograma de frecuencias (véase figura 5.2), decidimos que la función de densidad que los puede representar es la normal. Los parámetros de esta función de densidad son,  $\mu$  y  $\sigma^2$ , que es posible estimar con  $\bar{x}$  y  $s_{n-1}^2$ , respectivamente. En este caso  $\bar{x} = 8.15$  y  $s_{n-1}^2 = 10.18$  a partir de estos valores calculamos las frecuencias esperadas, considerando que en ambos extremos, inferior y superior, los límites de la distribución esperada son,  $-\infty$  y  $\infty$  para abarcar todo el soporte de la distribución teórica.

$$n_1 = 60 \int_{-0.71}^{1.845} \frac{1}{3.19\sqrt{2\pi}} \exp\left(-\frac{(x-8.15)^2}{2 \times 10.18}\right) dx \approx 60 \int_{-\infty}^{1.845} \frac{1}{3.19\sqrt{2\pi}} \exp\left(-\frac{(x-8.15)^2}{2 \times 10.18}\right) dx = 1.44$$

$$n_2 = 60 \int_{1.845}^{4.400} \frac{1}{3.19\sqrt{2\pi}} \exp\left(-\frac{(x-8.15)^2}{2 \times 10.18}\right) dx = 60 \left[ \Phi\left(\frac{4.400-8.15}{3.19}\right) - \Phi\left(\frac{1.845-8.15}{3.19}\right) \right] \approx 5.75$$

$$n_3 = 60 \left[ \Phi\left(\frac{6.955-8.15}{3.19}\right) - \Phi\left(\frac{4.400-8.15}{3.19}\right) \right] \approx 14.05$$

$$n_4 = 60 \left[ \Phi\left(\frac{9.510-8.15}{3.19}\right) - \Phi\left(\frac{6.955-8.15}{3.19}\right) \right] \approx 18.67$$

$$n_5 = 60 \left[ \Phi\left(\frac{12.060 - 8.15}{3.19}\right) - \Phi\left(\frac{9.510 - 8.15}{3.19}\right) \right] \approx 13.49$$

$$n_6 = 60 \left[ \Phi(\infty) - \Phi\left(\frac{12.060 - 8.15}{3.19}\right) \right] \approx 6.61$$

**Paso 3.** Resultaron frecuencias esperadas menores a 5, que debemos agrupar:

$$\begin{cases} n_1^* = n_1 + n_2 = 7.19 > 5 \\ n_2^* = n_3 = 14.05 > 5 \\ n_3^* = n_4 = 18.67 > 5 \\ n_4^* = n_5 = 13.49 > 5 \\ n_5^* = n_6 = 6.61 > 5 \end{cases} \Rightarrow \begin{cases} O_1^* = O_1 + O_2 = 6 \\ O_2^* = O_3 = 16 \\ O_3^* = O_4 = 16 \\ O_4^* = O_5 = 15 \\ O_5^* = O_6 = 7 \end{cases}$$

Además,  $m = 5$  y  $k = 2$  parámetros, la estadística de prueba vale:

$$\begin{aligned} \chi_{cal}^2 &= \sum_{i=1}^m \frac{(O_i^* - n_i^*)^2}{n_i^*} = \frac{(6 - 7.19)^2}{7.19} + \frac{(16 - 14.05)^2}{14.05} + \frac{(16 - 18.67)^2}{18.67} + \frac{(15 - 13.49)^2}{13.49} + \frac{(7 - 6.61)^2}{6.61} \\ &= 1.041 \end{aligned}$$

**Paso 4.** Con el nivel de significancia,  $\alpha = 0.05$  y  $m - k - 1 = 5 - 2 - 1 = 2$  grados de libertad se tiene de tablas  $\chi_{2, 0.05}^2 = 5.99$ , comparando  $\chi_{cal}^2 = 1.041$ , y la regla de decisión, concluimos que a 5% de significancia no existen evidencias para rechazar que los datos tienen un comportamiento normal.

## Valor- $p$ en una prueba de hipótesis

Cuando se utilizan los paquetes estadísticos para llevar a cabo una prueba de hipótesis, en general utilizan el llamado valor- $p$ , el cual representa el área derecha que le corresponde al valor del estadístico de prueba calculado. La prueba de decisión se establece de la siguiente forma:

Rechazar  $H_0$ :  $x$ 's tiene un comportamiento  $f$ , al nivel de significancia  $\alpha$ , si  $p < \alpha$ .

En caso contrario, se concluye que con la muestra obtenida no existen evidencias para rechazar la hipótesis nula.

En el ejemplo 5.4 resultó  $\chi_{cal}^2 = 1.041$ , entonces por tablas de la distribución  $ji$  cuadrada el valor- $p$  está dado por:

$$p = P(\chi_2^2 > \chi_{cal}^2) = P(\chi_2^2 > 1.041) = 0.594 > 0.05$$

## 5.3 Uso de las pruebas de bondad de ajuste K-S y A-D

En esta sección continuaremos con el estudio de dos pruebas de bondad de ajuste no paramétricas que tienen la particularidad de utilizar la función de distribución acumulada en lugar de la función de densidad, como fue el caso de la prueba  $ji$  cuadrada.

La metodología en ambas pruebas es la misma, lo que las diferencia es su estadística de prueba, así como la tabla de valores comparativos para la estadística de tablas. Por esta razón vamos a ejemplificar ambas pruebas de forma paralela y los cálculos se realizarán con la ayuda del paquete estadístico Minitab v17.

### Prueba de bondad de ajuste Kolmogorov-Smirnov

Sean las  $n$  observaciones  $x_1, x_2, \dots, x_n$  y suponga que se desea probar el contraste de hipótesis (5.1), mediante la prueba Kolmogorov-Smirnov, que denotaremos por K-S. Es decir, tenemos que probar si las observaciones provienen de una función de distribución acumulada  $F_0$ , esto se puede realizar mediante la siguiente metodología.

**Paso 1.** Ordenar las  $n$  observaciones en forma no decreciente, denotadas por  $x_1, x_2, \dots, x_n$ .

**Paso 2.** Calcular los valores de la distribución teórica propuesta  $F_0$ , para los valores  $\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n$ .

**Paso 3.** Calcular el estadístico de prueba K-S dado por:

$$D = \max\{D^+, D^-\}$$

donde:

$$D^+ = \max_{1 \leq i \leq n} \left\{ \frac{i}{n} - F_0(\tilde{x}_i) \right\}; \quad D^- = \max_{1 \leq i \leq n} \left\{ F_0(\tilde{x}_i) - \frac{i-1}{n} \right\}$$

**Paso 4.** Con el valor del tamaño de muestra y nivel de significancia dado, determinar el valor del estadístico de tablas  $D_T(n, \alpha)$  (véase tabla 5.9). Por último, aplicar la regla de decisión.

Rechazar  $H_0$ :  $x$ 's tiene un comportamiento  $F$ , al nivel de significancia  $\alpha$ , si  $D > D_T(n, \alpha)$ .

En caso contrario, se concluye que, con la muestra presentada y un nivel de significancia  $\alpha$ , no existen evidencias para rechazar  $H_0$ .

### Ejemplo 5.5 Prueba K-S

Con los datos del ejemplo 5.2 y la prueba K-S, determine a un nivel de significancia de 5% si existen evidencias de normalidad en su distribución.

#### Solución

En la primera columna de la tabla 5.8 se muestran las 60 observaciones.

**Paso 1.** Ordenar las 60 observaciones en forma no decreciente, denotemos por  $\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_{60}$  (véase la columna 3 de la tabla 5.8).

**Paso 2.** Calcular los valores de la distribución teórica propuesta  $F_0$ , para los valores  $\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n$  (véase la columna 4 de la tabla 5.8).

**Paso 3.** Para el estadístico de prueba K-S, se calculan  $D^+$  y  $D^-$  (véanse las columnas 7 y 8 de la tabla 5.8). Al final de estas columnas se muestran los máximos.

$$D^+ = \max_{1 \leq i \leq n} \left\{ \frac{i}{n} - F_0(\tilde{x}_i) \right\} = 0.0571; \quad D^- = \max_{1 \leq i \leq n} \left\{ F_0(\tilde{x}_i) - \frac{i-1}{n} \right\} = 0.0639$$

$$\text{Entonces } D = \max\{D^+, D^-\} = 0.0639.$$

**Paso 4.** Con el valor del tamaño de muestra y nivel de significancia dado, determinar el valor del estadístico de tablas (véase tabla 5.9),  $D_T(60, 0.05) = \frac{1.36}{\sqrt{60}} = 0.1756 > D = 0.0639$ . Al final, aplicamos la regla de decisión para  $D_T(60, 0.05) > D$ , y concluimos que al 5% de significancia y la muestra presentada, no existen evidencias para rechazar  $H_0$ , entonces los datos provienen de una población con distribución normal.

**Tabla 5.8** Valores para el cálculo del estadístico de prueba K-S

$x_i$	$i$	$\bar{x}_i$ ord.	$F_0$	$i/n$	$(i-1)/n$	$D^+$	$D^-$
10.1	1	-0.71	0.0029	0.0167	0.0000	0.0138	0.0029
4.6	2	2.11	0.0302	0.0333	0.0167	0.0031	0.0135
3.23	3	2.65	0.0437	0.0500	0.0333	0.0063	0.0104
8.75	4	3.23	0.0631	0.0667	0.0500	0.0036	0.0131
10.33	5	3.32	0.0667	0.0833	0.0667	0.0166	0.0000
5.75	6	3.84	0.0902	0.1000	0.0833	0.0098	0.0069

(Continúa) ➔

Tabla 5.8 Valores para el cálculo del estadístico de prueba K-S (Continuación)

$x_i$	$i$	$\bar{x}_i$ ord.	$F_0$	$i/n$	$(i-1)/n$	$D^+$	$D^-$
7.38	7	4.50	0.1283	0.1167	0.1000	-0.0116	0.0283
2.65	8	4.60	0.1350	0.1333	0.1167	-0.0017	0.0183
7.08	9	4.81	0.1497	0.1500	0.1333	0.0003	0.0164
9.14	10	4.95	0.1600	0.1667	0.1500	0.0067	0.0100
3.84	11	5.19	0.1788	0.1833	0.1667	0.0045	0.0121
6.98	12	5.31	0.1888	0.2000	0.1833	0.0112	0.0055
6.55	13	5.56	0.2105	0.2167	0.2000	0.0062	0.0105
8.43	14	5.75	0.2279	0.2333	0.2167	0.0054	0.0112
-0.71	15	5.93	0.2452	0.2500	0.2333	0.0048	0.0119
4.81	16	5.94	0.2461	0.2667	0.2500	0.0206	-0.0039
2.11	17	6.00	0.2521	0.2833	0.2667	0.0312	-0.0146
6.52	18	6.04	0.2560	0.3000	0.2833	0.0440	-0.0273
9.81	19	6.27	0.2796	0.3167	0.3000	0.0371	-0.0204
9.85	20	6.29	0.2817	0.3333	0.3167	0.0516	-0.0350
12.12	21	6.52	0.3063	0.3500	0.3333	0.0437	-0.0270
4.5	22	6.55	0.3096	0.3667	0.3500	0.0571	-0.0404
5.31	23	6.98	0.3582	0.3833	0.3667	0.0251	-0.0085
13.15	24	7.08	0.3698	0.4000	0.3833	0.0302	-0.0135
6.29	25	7.15	0.3781	0.4167	0.4000	0.0386	-0.0219
9.03	26	7.38	0.4055	0.4333	0.4167	0.0278	-0.0112
7.83	27	7.83	0.4605	0.4500	0.4333	-0.0105	0.0272
5.94	28	7.93	0.4729	0.4667	0.4500	-0.0062	0.0229
11.11	29	8.19	0.5051	0.4833	0.4667	-0.0218	0.0384
9.02	30	8.41	0.5323	0.5000	0.4833	-0.0323	0.0490
5.56	31	8.43	0.5348	0.5167	0.5000	-0.0181	0.0348
6	32	8.53	0.5471	0.5333	0.5167	-0.0138	0.0304
6.04	33	8.75	0.5741	0.5500	0.5333	-0.0241	0.0408
4.95	34	8.81	0.5814	0.5667	0.5500	-0.0147	0.0314
7.93	35	8.93	0.5959	0.5833	0.5667	-0.0126	0.0292
13.32	36	9.02	0.6067	0.6000	0.5833	-0.0067	0.0234
12.63	37	9.03	0.6079	0.6167	0.6000	0.0088	0.0079
7.15	38	9.14	0.6210	0.6333	0.6167	0.0123	0.0043
8.41	39	9.81	0.6972	0.6500	0.6333	-0.0472	0.0639
11.77	40	9.85	0.7015	0.6667	0.6500	-0.0348	0.0515

(Continúa) ➤

**Tabla 5.8** Valores para el cálculo del estadístico de prueba K-S (*Continuación*)

$x_i$	$i$	$\bar{x}_{i, \text{ord.}}$	$F_0$	$i/n$	$(i-1)/n$	$D^+$	$D^-$
3.52	41	10.04	0.7217	0.6833	0.6667	-0.0384	0.0550
11.78	42	10.10	0.7279	0.7000	0.6833	-0.0279	0.0446
11.28	43	10.18	0.7361	0.7167	0.7000	-0.0194	0.0361
10.48	44	10.33	0.7511	0.7333	0.7167	-0.0178	0.0344
8.93	45	10.42	0.7599	0.7500	0.7333	-0.0099	0.0266
8.19	46	10.48	0.7656	0.7667	0.7500	0.0011	0.0156
10.84	47	10.84	0.7986	0.7833	0.7667	-0.0153	0.0319
10.18	48	11.11	0.8213	0.8000	0.7833	-0.0213	0.0380
10.42	49	11.14	0.8237	0.8167	0.8000	-0.0070	0.0237
14.62	50	11.28	0.8348	0.8333	0.8167	-0.0015	0.0181
8.53	51	11.77	0.8698	0.8500	0.8333	-0.0198	0.0365
11.14	52	11.78	0.8705	0.8667	0.8500	-0.0038	0.0205
6.27	53	11.94	0.8807	0.8833	0.8667	0.0026	0.0140
11.94	54	12.12	0.8915	0.9000	0.8833	0.0085	0.0082
8.81	55	12.39	0.9063	0.9167	0.9000	0.0104	0.0063
5.93	56	12.63	0.9182	0.9333	0.9167	0.0151	0.0015
5.19	57	13.15	0.9400	0.9500	0.9333	0.0100	0.0067
12.39	58	13.32	0.9460	0.9667	0.9500	0.0207	-0.0040
13.43	59	13.43	0.9497	0.9833	0.9667	0.0336	-0.0170
10.04	60	14.62	0.9779	1.0000	0.9833	0.0221	-0.0054
					máximo	0.0571	0.0639

En la tabla 5.9 se muestran los valores del estadístico de prueba por tablas para diferentes valores del nivel de significancia.

**Tabla 5.9** Valores de tablas para el estadístico para diferentes valores del nivel significancia K-S

$n$	Nivel de significancia							
	0.001	0.002	0.005	0.01	0.02	0.05	0.10	0.20
1	0.99950	0.99900	0.99750	0.99500	0.99000	0.97500	0.95000	0.90000
2	0.97764	0.96838	0.95000	0.92929	0.90000	0.84189	0.77639	0.68337
3	0.92065	0.90000	0.86428	0.82900	0.78456	0.70760	0.63604	0.56481
4	0.85047	0.82217	0.77639	0.73424	0.68887	0.62394	0.56522	0.49265
5	0.78137	0.75000	0.70543	0.66853	0.62718	0.56328	0.50945	0.44698
6	0.72479	0.69571	0.65287	0.61661	0.57741	0.51926	0.46799	0.41037
7	0.67930	0.65071	0.60975	0.57581	0.53844	0.48342	0.43607	0.38148
8	0.64098	0.61368	0.57429	0.54179	0.50654	0.45427	0.40962	0.35831
9	0.60846	0.58210	0.54443	0.51332	0.47960	0.43001	0.38746	0.33910

(Continúa)

**Tabla 5.9** Valores de tablas para el estadístico para diferentes valores del nivel significancia K-S (*Continuación*)

<i>n</i>	Nivel de significancia							
	0.001	0.002	0.005	0.01	0.02	0.05	0.10	0.20
<b>10</b>	0.58042	0.55500	0.51872	0.48893	0.45562	0.40925	0.36866	0.32260
<b>11</b>	0.55588	0.53135	0.49539	0.46770	0.43670	0.39122	0.35242	0.30829
<b>12</b>	0.53422	0.51047	0.47672	0.44905	0.41918	0.37543	0.33815	0.29577
<b>13</b>	0.51490	0.49189	0.45921	0.43247	0.40362	0.36143	0.32549	0.28470
<b>14</b>	0.49753	0.47520	0.44352	0.41762	0.38970	0.34890	0.31417	0.27481
<b>15</b>	0.48182	0.45611	0.42934	0.40420	0.37713	0.33750	0.30397	0.26589
<b>16</b>	0.46750	0.44637	0.41644	0.39201	0.36571	0.32733	0.29472	0.25778
<b>17</b>	0.45540	0.43380	0.40464	0.38086	0.35528	0.31796	0.28627	0.25039
<b>18</b>	0.44234	0.42224	0.39380	0.37062	0.34569	0.30936	0.27851	0.24360
<b>19</b>	0.43119	0.41156	0.38379	0.36117	0.33685	0.30143	0.27136	0.23735
<b>20</b>	0.42085	0.40165	0.37451	0.35241	0.32866	0.29408	0.26473	0.23156
<b>21</b>	0.41122	0.39243	0.36588	0.34426	0.32104	0.28724	0.25858	0.22517
<b>22</b>	0.40223	0.38382	0.35782	0.33666	0.31394	0.28087	0.25283	0.22115
<b>23</b>	0.39380	0.37575	0.35027	0.32954	0.30728	0.27491	0.24746	0.21646
<b>24</b>	0.38588	0.36787	0.34318	0.32286	0.30104	0.26931	0.24242	0.21205
<b>25</b>	0.37743	0.36104	0.33651	0.31657	0.29518	0.26404	0.23768	0.20790
<b>26</b>	0.37139	0.35431	0.33022	0.30963	0.28962	0.25908	0.23320	0.20399
<b>27</b>	0.36473	0.34794	0.32425	0.30502	0.28438	0.25438	0.22898	0.20030
<b>28</b>	0.35842	0.34190	0.31862	0.29971	0.27942	0.24993	0.22497	0.19680
<b>29</b>	0.35242	0.33617	0.31327	0.29466	0.27471	0.24571	0.22117	0.19348
<b>30</b>	0.34672	0.33072	0.30818	0.28986	0.27023	0.24170	0.21756	0.19032
<b>31</b>	0.34129	0.32553	0.30333	0.28529	0.26596	0.23788	0.21412	0.18732
<b>32</b>	0.33611	0.32058	0.29870	0.28094	0.26189	0.23424	0.21085	0.18445
<b>33</b>	0.33115	0.31584	0.29428	0.27577	0.25801	0.23076	0.20771	0.18171
<b>34</b>	0.32641	0.31131	0.29005	0.27271	0.25429	0.22743	0.21472	0.17909
<b>35</b>	0.32187	0.30597	0.28600	0.26897	0.25073	0.22425	0.20185	0.17659
<b>36</b>	0.31751	0.30281	0.28211	0.26532	0.24732	0.22119	0.19910	0.17418
<b>37</b>	0.31333	0.29882	0.27838	0.26180	0.24404	0.21826	0.19646	0.17188
<b>38</b>	0.30931	0.29498	0.27483	0.25843	0.24089	0.21544	0.19392	0.16966
<b>39</b>	0.30544	0.29125	0.27135	0.25518	0.23785	0.21273	0.19148	0.16753
<b>40</b>	0.30171	0.28772	0.26803	0.25205	0.23494	0.21012	0.18913	0.16547
<b>41</b>	0.29811	0.28429	0.26482	0.24904	0.23213	0.20760	0.18687	0.16349
<b>42</b>	0.29465	0.28097	0.26173	0.24613	0.22941	0.20517	0.18468	0.16158
<b>43</b>	0.29130	0.27778	0.25875	0.24332	0.22679	0.20283	0.18257	0.15974
<b>44</b>	0.28806	0.27468	0.25587	0.24060	0.22426	0.20056	0.18051	0.15795

*(Continúa)*

**Tabla 5.9** Valores de tablas para el estadístico para diferentes valores del nivel significancia K-S (*Continuación*)

Nivel de significancia								
<i>n</i>	0.001	0.002	0.005	0.01	0.02	0.05	0.10	0.20
<b>45</b>	0.28493	0.27169	0.25308	0.23798	0.22181	0.19837	0.17856	0.15623
<b>46</b>	0.28190	0.26880	0.25038	0.23544	0.21944	0.19625	0.17665	0.15457
<b>47</b>	0.27896	0.26600	0.24776	0.23298	0.21715	0.19420	0.17481	0.15295
<b>48</b>	0.27611	0.26328	0.24523	0.23059	0.21493	0.19221	0.17301	0.15139
<b>49</b>	0.27339	0.26069	0.24281	0.22832	0.21281	0.19028	0.17128	0.14987
<b>50</b>	0.27067	0.25809	0.24039	0.22604	0.21068	0.18841	0.16959	0.14840
<i>n</i> > 50	$\frac{1.95}{\sqrt{n}}$	$\frac{1.85}{\sqrt{n}}$	$\frac{1.73}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$	$\frac{1.52}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.07}{\sqrt{n}}$

## Prueba de bondad de ajuste Kolmogorov-Smirnov con Minitab

La prueba de K-S requiere de varias operaciones, lo que la hace lenta en su aplicación, además de requerir la tabla 5.9. Pero debido a su importancia y popularidad los paquetes estadísticos tienen una aplicación de esta prueba. Nosotros vamos a revisar el paquete estadístico Minitab v.17.

### Ejemplo 5.6 Prueba K-S

Resuelva el ejemplo 5.5 con ayuda del paquete Minitab v.17.

#### Solución

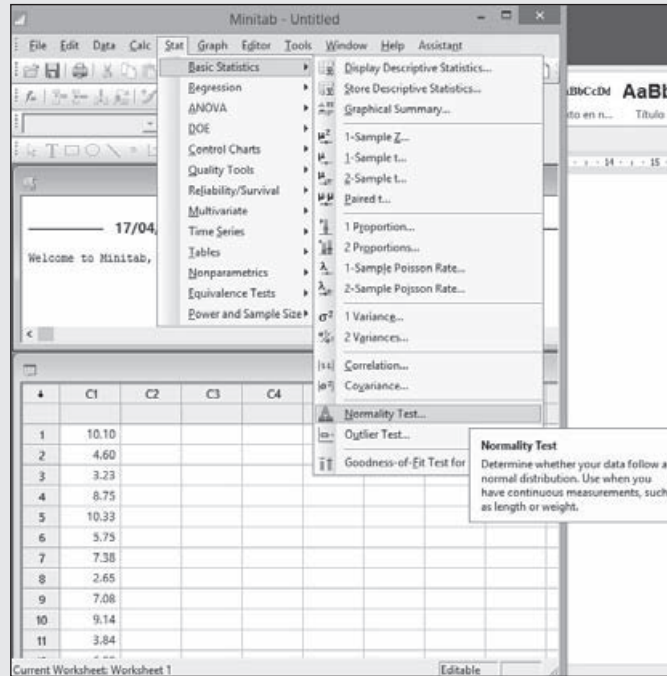
En la figura 5.5 se muestra la pantalla del paquete Minitab con las 60 observaciones.

	C1	C2	C3	C4	C5	C6	C7	C8
1	10.10							
2	4.60							
3	3.23							
4	8.75							
5	10.33							
6	5.75							
7	7.38							
8	2.65							
9	7.08							
10	9.14							
11	3.84							

**Figura 5.5** Valores de las observaciones capturados en el paquete Minitab.

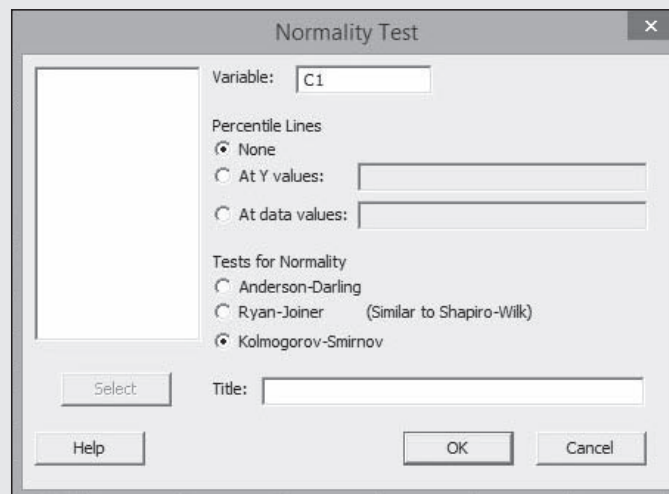


Seleccionar la pestaña **Stat**, elegir la opción **Basic Statistics** y escoger **Normality Test**, como se muestra en la figura 5.6.



**Figura 5.6** Pantalla de la opción del paquete Minitab para la prueba de normalidad.

En la pantalla que aparece de la prueba de normalidad en la opción de **Variable:** **C1** seleccionar la columna en la que aparecen las observaciones. Después, elija la prueba de normalidad que desee, hay disponibles tres tipos diferentes (véase la figura 5.7).



**Figura 5.7** Pantalla de la opción del paquete Minitab para la prueba K-S.

Después de seleccionar el tipo de prueba y presionar **OK**, aparece la respuesta (véase figura 5.8), en la que se puede apreciar el valor obtenido del estadístico de prueba  $D = 0.064$ , que le corresponde un valor- $p$  mayor a 0.15, del cual se concluye que a 5% de significancia no existen evidencias para rechazar la hipótesis nula.

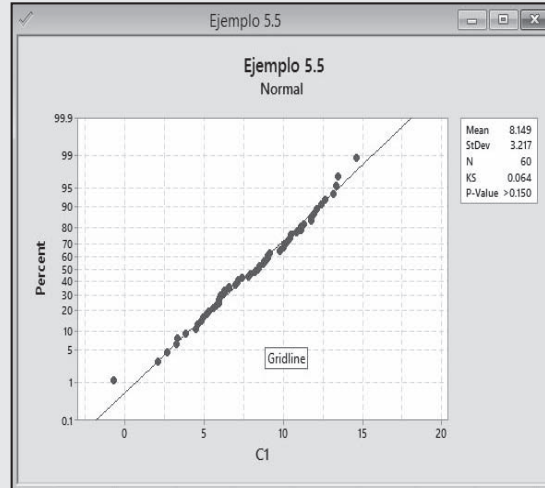


Figura 5.8 Pantalla de resultados de la prueba K-S.

## Prueba de bondad de ajuste Anderson-Darling con Minitab

Sean las  $n$  observaciones  $x_1, x_2, \dots, x_n$  y supóngase que se desea probar el contraste de hipótesis (5.1), mediante la prueba Anderson-Darling, que denotaremos por A-D. Es decir, tenemos que probar si las observaciones provienen de una función de distribución acumulada propuesta  $F_0$ , esto se puede realizar mediante la siguiente metodología.

**Paso 1.** Ordenar las  $n$  observaciones en forma no decreciente, denotemos por  $\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n$ .

**Paso 2.** Calcular los valores de la distribución teórica propuesta  $F_0$ , para los valores  $\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n$ .

**Paso 3.** Calcular el estadístico de prueba A-D dado en (5.2)

$$AD = -n - \sum_{i=1}^n \frac{(2i-1)}{n} \left\{ \ln[F_0(\tilde{x}_i)] + \ln[1 - F_0(\tilde{x}_{n+1-i})] \right\} \quad (5.2)$$

**Paso 4.** Con la distribución del estadístico de prueba AD, calcular el valor- $p$  y comparar con el nivel de significancia dado. Por último, aplicar la regla de decisión.

Rechazar  $H_0$ :  $x$ 's tiene un comportamiento  $F$ , al nivel de significancia  $\alpha$ ,  $p < \alpha$ .

En caso contrario, se concluye que, con la muestra presentada y un nivel de significancia  $\alpha$ , no existen evidencias para  $H_0$ .

### Ejemplo 5.7 Prueba A-D

Con los datos del ejemplo 5.2 y la prueba A-D, determine a un nivel de significancia de 5% si existen evidencias de normalidad en su distribución. Después, utilice el paquete Minitab y compare resultados.

#### Solución

En la primera columna de la tabla 5.10 se muestran las 60 observaciones.

**Paso 1.** Ordenar las 60 observaciones en forma no decreciente, denotemos por  $\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_{60}$  (véase la columna 3 de la tabla 5.10).

**Paso 2.** Calcular los valores de la distribución teórica propuesta  $F_0$ , para los valores  $\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n$  (véanse las columnas 4 y 6 de la tabla 5.10).

**Paso 3.** Para el estadístico de prueba A-D, al final de la columna 7 de tabla 5.10 se muestra la suma:

$$AD = -n - \sum_{i=1}^n \frac{(2i-1)}{n} \left\{ \ln[F_0(\bar{x}_i)] + \ln[1 - F_0(\bar{x}_{n+1-i})] \right\} = -60 - (-60.2080) = 0.2080$$

**Paso 4.** Se busca con un paquete el valor- $p$  de la estadística de A-D para 0.2080, le corresponde el valor  $0.860 > \alpha = 0.05$ . Por lo que, aplicamos la regla de decisión y concluimos que a 5% de significancia con la muestra presentada, no existen evidencias para  $H_0$ , los datos provienen de una población con distribución normal.

**Tabla 5.10** Valores para el cálculo del estadístico de prueba A-D

$x_i$	$i$	$\bar{x}_i$ ord.	$F_0(\bar{x}_i)$	$\bar{x}_{n-i+1}$	$F_0(\bar{x}_{n-i+1})$	$\frac{(2i-1)}{n} \left\{ \ln[F_0(\bar{x}_i)] + \ln[1 - F_0(\bar{x}_{n+1-i})] \right\}$
10.1	1	-0.71	0.0029	14.62	0.9779	-0.1606
4.6	2	2.11	0.0302	13.43	0.9497	-0.3244
3.23	3	2.65	0.0437	13.32	0.9460	-0.5041
8.75	4	3.23	0.0631	13.15	0.9400	-0.6505
10.33	5	3.32	0.0667	12.63	0.9182	-0.7817
5.75	6	3.84	0.0902	12.39	0.9063	-0.8751
7.38	7	4.50	0.1283	12.12	0.8915	-0.9260
2.65	8	4.60	0.1350	11.94	0.8807	-1.0322
7.08	9	4.81	0.1497	11.78	0.8705	-1.1173
9.14	10	4.95	0.1600	11.77	0.8698	-1.2259
3.84	11	5.19	0.1788	11.28	0.8348	-1.2326
6.98	12	5.31	0.1888	11.14	0.8237	-1.3045
6.55	13	5.56	0.2105	11.11	0.8213	-1.3669
8.43	14	5.75	0.2279	10.84	0.7986	-1.3865
-0.71	15	5.93	0.2452	10.48	0.7656	-1.3808
4.81	16	5.94	0.2461	10.42	0.7599	-1.4614
2.11	17	6.00	0.2521	10.33	0.7511	-1.5228
6.52	18	6.04	0.2560	10.18	0.7361	-1.5718
9.81	19	6.27	0.2796	10.10	0.7279	-1.5886
9.85	20	6.29	0.2817	10.04	0.7217	-1.6549
12.12	21	6.52	0.3063	9.85	0.7015	-1.6347
4.5	22	6.55	0.3096	9.81	0.6972	-1.6965
5.31	23	6.98	0.3582	9.14	0.6210	-1.4977
13.15	24	7.08	0.3698	9.03	0.6079	-1.5126
6.29	25	7.15	0.3781	9.02	0.6067	-1.5565
9.03	26	7.38	0.4055	8.93	0.5959	-1.5374
7.83	27	7.83	0.4605	8.81	0.5814	-1.4542
5.94	28	7.93	0.4729	8.75	0.5741	-1.4690

(Continúa) ➤

Tabla 5.10 Valores para el cálculo del estadístico de prueba A-D (Continuación)

$x_i$	$i$	$\bar{x}_i$ ord.	$F_0(\bar{x}_i)$	$\bar{x}_{n-i+1}$	$F_0(\bar{x}_{n-i+1})$	$\frac{(2i-1)}{n} \{ \ln[F_0(\bar{x}_i)] + \ln[1 - F_0(\bar{x}_{n-i+1})] \}$
11.11	29	8.19	0.5051	8.53	0.5471	-1.4014
9.02	30	8.41	0.5323	8.43	0.5348	-1.3725
5.56	31	8.43	0.5348	8.41	0.5323	-1.4090
6	32	8.53	0.5471	8.19	0.5051	-1.3717
6.04	33	8.75	0.5741	7.93	0.4729	-1.2949
4.95	34	8.81	0.5814	7.83	0.4605	-1.2947
7.93	35	8.93	0.5959	7.38	0.4055	-1.1934
13.32	36	9.02	0.6067	7.15	0.3781	-1.1533
12.63	37	9.03	0.6079	7.08	0.3698	-1.1674
7.15	38	9.14	0.6210	6.98	0.3582	-1.1498
8.41	39	9.81	0.6972	6.55	0.3096	-0.9383
11.77	40	9.85	0.7015	6.52	0.3063	-0.9483
3.32	41	10.04	0.7217	6.29	0.2817	-0.8870
11.78	42	10.10	0.7279	6.27	0.2796	-0.8930
11.28	43	10.18	0.7361	6.04	0.2560	-0.8531
10.48	44	10.33	0.7511	6.00	0.2521	-0.8361
8.93	45	10.42	0.7599	5.94	0.2461	-0.8264
8.19	46	10.48	0.7656	5.93	0.2452	-0.8316
10.84	47	10.84	0.7986	5.75	0.2279	-0.7496
10.18	48	11.11	0.8213	5.56	0.2105	-0.6858
10.42	49	11.14	0.8237	5.31	0.1888	-0.6516
14.62	50	11.28	0.8348	5.19	0.1788	-0.6231
8.53	51	11.77	0.8698	4.95	0.1600	-0.5283
11.14	52	11.78	0.8705	4.81	0.1497	-0.5164
6.27	53	11.94	0.8807	4.60	0.1350	-0.4761
11.94	54	12.12	0.8915	4.50	0.1283	-0.4498
8.81	55	12.39	0.9063	3.84	0.0902	-0.3505
5.93	56	12.63	0.9182	3.32	0.0667	-0.2856
5.19	57	13.15	0.9400	3.23	0.0631	-0.2394
12.39	58	13.32	0.9460	2.65	0.0437	-0.1920
13.43	59	13.43	0.9497	2.11	0.0302	-0.1606
10.04	60	14.62	0.9779	-0.71	0.0029	-0.0502
					Suma	-60.2080

Si se sigue el mismo procedimiento que en la prueba K-S en el paquete Minitab, pero seleccionando la prueba A-D, obtenemos los valores buscados (véase figura 5.9).

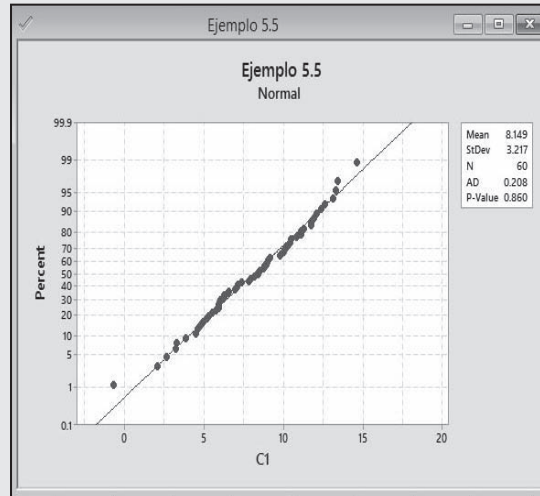


Figura 5.9 Pantalla de resultados de la prueba A-D.

## Ejercicios de repaso

### Preguntas de autoevaluación

- 5.1 ¿Qué es una prueba de bondad de ajuste?
- 5.2 ¿En qué se basa la prueba K-S para realizar una prueba de bondad de ajuste sobre un conjunto de datos?
- 5.3 ¿En qué se basa la prueba A-D para realizar una prueba de bondad de ajuste sobre un conjunto de datos?
- 5.4 ¿En qué se basa la prueba  $\chi^2$  para realizar una prueba de bondad de ajuste sobre un conjunto de datos?
- 5.5 ¿En qué se basa la prueba Q-Q para realizar una prueba de bondad de ajuste sobre un conjunto de datos?
- 5.6 ¿Es cierto que el resultado de todas las pruebas de bondad de ajuste que se apliquen a un mismo conjunto de datos y nivel de significancia siempre deben dar el mismo resultado? Justifique su respuesta.
- 5.7 Con los conceptos revisados en esta unidad, ¿cuál sería la metodología para presentar un informe estadístico sobre el comportamiento de las observaciones de la muestra?

- 5.8 ¿Qué es el valor- $p$ ?
- 5.9 ¿Cuál es la regla de decisión basada en el valor- $p$ , para la prueba K-S?
- 5.10 ¿Qué es un cuantil?
- 5.11 ¿Cuál es el contraste de hipótesis para una prueba de bondad de ajuste?
- 5.12 ¿Cuál es el estadístico de prueba para K-S?

### Ejercicios complementarios con grado de dificultad dos

- 5.13 Con los datos de la tabla 5.11 determine si existen evidencias de normalidad.
  - a) Mediante la técnica gráfica Q-Q con 5% de significancia.
  - b) Por medio de la prueba  $\chi^2$  con 5% de significancia.
  - c) Con las pruebas K-S y A-D y 5% de significancia.

Tabla 5.11

12.02	6.50	8.12	10.54	9.38	6.23	9.41	4.91	3.09	12.91	2.85	10.19
9.71	12.68	10.28	7.11	12.71	9.96	5.47	12.72	1.50	13.42	5.32	4.47
0.14	2.65	10.68	8.72	5.58	7.05	10.52	5.25	7.27	8.80	7.66	9.67
12.61	6.61	0.64	6.92	8.62	5.67	7.95	8.14	5.49	3.38	9.27	6.51
2.78	8.26	6.66	6.83	5.27	6.28	4.00	7.25	8.29	3.57	7.45	11.68

**5.14.** Distribuya los datos de la tabla 5.12 en 10 clases de frecuencias y trace su histograma de frecuencias para determinar una posible distribución de los datos. Luego, determine si existen evidencias de que los datos provienen de la distribución que proponga.

- Mediante la técnica gráfica Q-Q con 2% de significancia.
- Por medio de la prueba *ji* cuadrada con 2% de significancia.
- Con las pruebas K-S y A-D y 2% de significancia.

**Tabla 5.12**

10.66	10.21	0.57	18.21	9.11	18.98
7.52	27.19	2.57	4.09	7.44	27.52
50.38	9.23	1.47	8.93	3.65	21.48
0.37	26.83	15.29	16.06	63.83	35.99
25.89	22.57	0.26	33.87	8.29	8.92
48.73	12.79	11.55	12.12	18.62	5.32
54.80	1.15	14.70	8.74	16.28	3.92
29.63	70.36	23.51	14.56	30.27	3.31
3.25	9.20	0.12	27.55	12.44	6.77
28.22	26.98	34.87	67.96	7.98	12.51
29.41	11.61	95.83	4.19	54.26	14.66
4.06	17.11	81.07	12.59	0.81	7.11
3.14	4.88	21.64	1.92	31.02	10.00
108.13	10.74	3.17	9.71	21.96	12.57
39.73	3.34	46.17	5.59	20.16	4.07
17.01	82.00	22.85	21.36	12.17	18.00
2.13	0.91	9.03	18.56	37.51	0.77
48.45	11.17	26.04	1.82	7.19	1.48
0.41	8.77	32.24	12.51	50.69	8.56
2.93	2.90	37.32	39.33	1.89	57.94

**5.15** Los datos de la tabla 5.13 muestran los sueldos de 90 personas de una empresa de limpieza elegidas de manera aleatoria. Ordene los datos en 10 clases de frecuencia de igual longitud y determine si existen evidencias de que los datos provienen de una población normal.

- Mediante la técnica gráfica Q-Q con 10% de significancia.
- Por medio de la prueba *ji* cuadrada con 10% de significancia.

c) Con las pruebas K-S y A-D y 10% de significancia.

**Tabla 5.13**

3941	7037	4194	6324	3053
5735	3686	6132	3385	2529
5035	7480	5644	5057	2948
2773	3852	5330	5014	2829
2528	3162	6028	5722	6990
2557	4978	4678	6002	4265
5060	6167	7071	4325	6431
6732	6913	6392	7454	5386
4389	5622	5179	3233	7108
3583	4460	6919	2776	2399
5899	3599	3340	2762	5556
5755	3840	2876	5904	2769
2402	6947	7130	4928	1935
3649	4419	3931	2885	5178
5132	3613	7409	5549	3473
3968	4878	7045	3678	5317
3828	5718	6965	6368	4193
6358	6113	3487	6072	2839

**5.16** En la tabla 5.14 se muestran los errores tipográficos por página que comete una secretaria, en las primeras 100 páginas. Divida los datos en ocho clases de frecuencia de igual longitud y determine si existen evidencias de que los datos provienen de la distribución normal.

- Mediante la técnica gráfica Q-Q con 5% de significancia.
- Por medio de la prueba *ji* cuadrada con 5% de significancia.
- Con las pruebas K-S y A-D y 5% de significancia.

**Tabla 5.14**

0	2	3	2	1	5	2	1	6	3
1	5	6	2	3	2	2	2	4	5
5	3	2	6	7	1	3	7	2	3
4	4	5	8	1	3	4	7	3	8
10	0	5	3	2	4	4	6	7	8
9	2	4	6	2	3	4	7	6	4
5	4	6	7	7	2	1	3	8	2
4	5	6	2	7	2	5	5	1	8
3	4	7	8	2	8	1	3	4	4
3	5	6	2	4	2	6	8	1	7

**5.17** Los registros de la tabla 5.15 corresponden al tiempo de funcionamiento (en días) hasta que se presenta la primera falla de 55 teclados de computadora. Construya una distribución de frecuencias que contenga siete intervalos de clase de igual longitud y determine si existen evidencias de que los datos provienen de la distribución que proponga.

- Mediante la técnica gráfica Q-Q con 1% de significancia.
- Por medio de la prueba  $\chi^2$  cuadrada con 1% de significancia.
- Con las pruebas K-S y A-D y 1% de significancia.

Tabla 5.15

224	80	96	536	400	392	576	128	56	656	224
358	384	256	246	328	464	448	304	72	80	72
56	108	194	136	224	80	424	156	216	168	184
552	372	184	438	120	308	272	152	328	60	208
340	104	168	232	112	288	64	176	160	608	400

**5.18** En un experimento que medía el porcentaje de encogimiento al secar 50 especímenes de prueba de arcilla plástica, se obtuvieron los resultados que se muestran en la tabla 5.16.

Tabla 5.16

19.3	20.5	17.9	17.3	17.1	15.8	16.9	17.1	19.5	22.5
18.5	22.5	19.1	17.9	18.4	18.7	18.8	17.5	17.5	14.9
19.4	16.8	19.3	17.3	19.5	17.4	16.3	18.8	21.3	23.4
19.0	19.0	16.1	18.8	17.5	18.2	17.4	18.6	18.3	16.5
17.4	20.5	16.9	17.5	22.5	22.6	20.7	12.3	18.5	17.4

Construya una distribución de frecuencias que contenga siete intervalos de clase de igual longitud y determine si existen evidencias de que los datos provienen de una distribución normal.

- Mediante la técnica gráfica Q-Q con 5% de significancia.
  - Por medio de la prueba  $\chi^2$  cuadrada con 5% de significancia.
  - Mediante las pruebas K-S y A-D y 5% de significancia.
- 5.19** En un experimento de psicología se pide a varios individuos que memoricen cierta secuencia de palabras. A continuación se registran los tiempos (en segundos) que necesitan los participantes para lograrlo (véase tabla 5.17). Construya una distribución de frecuencias que contenga ocho intervalos de clase de igual longitud y determine si existen evidencias de que los datos provienen de la distribución que proponga.
- Mediante la técnica gráfica Q-Q con 10% de significancia.
  - Por medio de la prueba  $\chi^2$  cuadrada con 10% de significancia.

c) Con las pruebas K-S y A-D y 10% de significancia.

Tabla 5.17

100	107	34	57	66	30	79	84
126	89	128	100	88	61	108	79
129	46	107	109	32	106	122	41
88	85	149	75	105	50	99	50
69	87	64	85	126	100	102	112
127	102	129	88	123	98	110	93
103	149	90	145	96	146	119	76

118	77	135	95	130	138	52
37	93	116	45	57	112	73
70	96	98	117	97	99	62
79	43	90	114	53	123	100
78	118	135	110	64	62	107
135	58	73	80	125	88	142
93	99					

**5.20** Considere los datos de la tabla 5.18 que corresponden al porcentaje de algodón en el material usado para fabricar playeras para caballero. Construya una distribución de frecuencias que contenga nueve intervalos de clase de igual longitud y determine si existen evidencias de que los datos provienen de la distribución que proponga.

- Mediante la técnica gráfica Q-Q con 5% de significancia.
- Por medio de la prueba  $\chi^2$  cuadrada con 5% de significancia.
- Con las pruebas K-S y A-D y 5% de significancia.

Tabla 5.18

34.2	33.6	33.8	34.7	37.8	32.6	35.8	34.6
33.1	34.7	34.2	33.6	36.6	33.1	37.6	33.6
34.5	35.0	33.4	32.5	35.4	34.6	37.3	34.1
35.6	35.4	34.7	34.1	34.6	35.9	34.6	34.7
34.3	36.2	34.6	35.1	33.8	34.7	35.5	35.7
35.1	36.8	35.2	36.8	37.1	33.6	32.8	36.8
34.7	36.1	35.0	37.9	34.0	32.9	32.1	34.3
33.6	35.3	34.9	36.4	34.1	33.5	34.5	32.7

**5.21** En la tabla 5.19 se muestran los reportes de artículos defectuosos que fueron producidos diariamente por la línea A. Lleve a cabo una prueba de bondad de ajuste para normalidad de los datos, con un nivel de significancia de 5%. Utilice las diferentes pruebas revisadas en la unidad.



Tabla 5.19

37	35	38	35	35
33	29	38	40	44
23	36	43	28	33
34	32	32	35	36
37	36	41	32	27
33	34	33	31	32
32	35	31	45	32
33	37	35	46	33
38	30	42	27	33
37	38	26	30	33
36	34	26	36	36
38	33	39	32	41
34	31	31	24	45
33	34	29	36	33
37	31	40	39	40
35	29	29	39	42
38	33	29	37	41
42	37	43	39	39
35	33	33	36	26
40	44	34	31	31

**5.22** Los días de duración de 50 baterías para automóvil se muestran en la tabla 5.20. Lleve a cabo una prueba de bondad de ajuste para normalidad de los datos, con un nivel de significancia de 5%. Utilice las diferentes pruebas revisadas en la unidad.

Tabla 5.20

984.9	1 068.6	994.1	1 166.8	1 235.9
1 003.4	1 108.2	985.2	1 022.8	971.1
1 012.4	953.4	955.9	1 140.5	969.8
1 057.5	1 402.8	1 121.7	1 203.0	967.4
1 223.6	998.7	1 153.9	987.2	1 043.7
1 096.8	951.2	1 011.5	956.1	1 028.5
962.8	962.3	995.4	1 051.1	959.4
1 052.8	1 065.0	1 010.7	1 002.9	995.0
995.1	1 151.3	1 164.6	967.3	964.8
1 010.3	1 002.5	1 012.7	1 137.9	1 035.2

## Proyectos de la unidad 5

- I. En la hoja divisas del archivo Datos IPC Divisas.xlsx que se encuentra en la página de recursos SALI, está una base de datos tomada de la página: <http://economia.terra.com.mx/mercados/acciones/cambios.aspx?idtel=IB032MEXBOL>, del curso de diferentes monedas con respecto del dólar. La base de datos es a partir de febrero de 2013. Con esta información realice los siguientes proyectos:
  1. Con el curso diario del peso mexicano trace un histograma con 15 clases.
    - a) Con base en la gráfica obtenida proponga una distribución para el comportamiento de la muestra.
    - b) Realice una prueba de bondad de ajuste a 2% de significancia con las diferentes pruebas revisadas en la unidad y decida si la muestra obtenida tiene el comportamiento propuesto.
  2. Con el curso diario del peso uruguayo trace un histograma con 15 clases.
    - a) Con base en la gráfica obtenida proponga una distribución para el comportamiento de la muestra.
    - b) Realice una prueba de bondad de ajuste a 5% de significancia con las diferentes pruebas revisadas en la unidad y decida si la muestra obtenida tiene el comportamiento propuesto.
- II. La hoja IPC-42 Emp del archivo Datos IPC Divisas.xlsx, que se encuentra en la página de recursos SALI, contiene todos los IPC de 41 empresas que cotizan en México. La base de datos es a partir de febrero de 2013. Con esta información realice los siguientes proyectos:
  1. Con los valores del IPC del Grupo Alfa A<sup>®</sup> trace un histograma con 20 clases.
    - a) Con base en el histograma proponga una distribución para el comportamiento de la muestra.

- b)* Realice una prueba de bondad de ajuste a 1% de significancia con las diferentes pruebas revisadas en la unidad y decida si la muestra obtenida tiene el comportamiento propuesto.
- 2.** Con los valores del IPC de Liverpool® trace un histograma con 20 clases.
- a)* Con base en el histograma proponga una distribución para el comportamiento de la muestra.
- b)* Realice una prueba de bondad de ajuste a 5% de significancia con las diferentes pruebas revisadas en la unidad y decida si la muestra obtenida tiene el comportamiento propuesto.
- III.** En la hoja Divorcios por entidad, del archivo Datos de divorcios.xlsx que se encuentra en la página de recursos SALI, está una base de datos extraída del INEGI sobre los divorcios registrados en la República Mexicana para cada estado de 1985 a 2011.
- 1.** Con las cantidades de divorcios de la causa enajenación trace un histograma con nueve clases.
- a)* Con base en el histograma proponga una distribución para el comportamiento de la muestra.
- b)* Realice una prueba de bondad de ajuste a 5% de significancia con las diferentes pruebas revisadas en la unidad y decida si la muestra obtenida tiene el comportamiento propuesto.
- 2.** Con las cantidades de divorcios de la causa incitación a la violencia trace un histograma con 12 clases. Con base en el histograma ¿será posible ajustar una distribución para el comportamiento de la muestra? Explique su respuesta.

# Regresión lineal simple y múltiple

UNIDAD  
**6**



## Competencias específicas a desarrollar

- Identificar y aplicar los conceptos básicos del modelo de regresión simple.
- Establecer las condiciones para distinguir entre una regresión y una correlación.
- Identificar y aplicar los conceptos básicos del modelo de regresión múltiple.
- Identificar y aplicar los conceptos básicos del modelo de regresión no lineal.

## ¿Qué sabes?

- ¿Conoces cuáles son los supuestos en los errores para un modelo de regresión lineal?
- ¿Podrías explicar cómo son las variables independientes en un modelo de regresión lineal?
- ¿Sabes qué son los MELI?
- ¿Cómo identificas el problema de multicolinealidad?

## Introducción

Los datos numéricos que están relacionados abundan en todas partes; por ejemplo en los negocios, la economía, la medicina, la física, etcétera. Con frecuencia, se tiene la necesidad de examinar las relaciones entre diferentes variables. Los siguientes ejemplos muestran situaciones en las que dos variables se encuentran relacionadas.

1. La cantidad de anuncios de un producto determinado que se presenta en los medios de comunicación está relacionada con sus ventas.
2. El número de artículos que no cumplen con las normas de calidad está relacionado con el costo del producto.
3. El calentamiento de un cable en un circuito eléctrico está relacionado con el voltaje aplicado a éste y su tiempo de uso.
4. El rendimiento de un compuesto químico está afectado por la temperatura y la presión.
5. El precio de licitación para un proyecto de construcción de carreteras puede estar relacionado con su longitud y la cantidad de licitadores.

En los ejemplos anteriores se observa que es interesante investigar y proponer un modelo que defina la relación entre las variables que están analizándose, además de proporcionar una medida que dé a conocer el grado de asociación entre éstas. El modelo puede ser usado en diferentes procesos, por ejemplo, para **predicción, control u optimización** de la producción o ventas del producto.

En general, el análisis de regresión es una técnica estadística utilizada para la estimación de las relaciones entre las variables, que incluye muchas técnicas para modelar y analizar varias variables, cuando la atención se centra en la relación entre una variable dependiente y una o más variables independientes. De manera específica, el análisis de regresión ayuda a entender cómo influyen los cambios de las variables independientes en una variable dependiente o de respuesta, o incluso cómo influye en la variable de respuesta la variación de cualquiera de las variables independientes, cuando las demás se mantienen fijas. De manera más común, se dice que el análisis de regresión estima la esperanza condicional de la variable dependiente dadas las variables independientes. Es decir, conocer el valor medio de la dependiente cuando se fija un valor a cada una de las independientes.

El análisis de regresión es muy utilizado para la predicción, previsión y determinación de cómo están relacionadas las variables independientes con la variable dependiente, con el objetivo de explorar las formas de estas relaciones (inferir las relaciones causales entre las variables independientes y dependientes). En general, el modelo de regresión se define en términos de un número finito de parámetros desconocidos que se estiman a partir de los datos de la muestra y depende, en cierta medida, de hacer suposiciones acerca del proceso de estimación. Estos supuestos deben ser comprobables cuando se dispone de una muestra de tamaño considerable. Los modelos de regresión para la predicción a menudo son útiles, incluso cuando los supuestos son violados en forma moderada, aunque en estos casos pueden no funcionar de manera óptima.

En la actualidad, se sabe que la primera forma de regresiones lineales documentada fue el método de los mínimos cuadrados, publicado por el matemático francés Adrien-Marie Legendre (18 de septiembre de 1752-10 de enero de 1833), en 1805 en un apéndice del libro sobre la órbita de los cometas, titulado, *Nouvelles méthodes pour la détermination des orbites des comètes*. En el trabajo de Legendre, el método no tiene relación con la estadística, ya que lo introdujo como un método geodésico, creado para ayudar a los astrónomos en el problema de determinar, a partir de observaciones astronómicas, las órbitas de los cuerpos alrededor del Sol. Más tarde, en 1809, el matemático alemán Johann Carl Friedrich Gauss (30 de abril de 1777-23 de febrero de 1855) publicó sus resultados en *Theoria motus corporum coelestium*, en el cual introdujo el método de mínimos cuadrados mediante el uso de conceptos estadísticos, como la distribución normal. Después, en 1821, Gauss publicó un desarrollo adicional de la teoría de los mínimos cuadrados, en el que incluyó una versión del teorema de Gauss-Markov sobre los modelos lineales.

Etimológicamente, el término regresión fue acuñado por el estadístico británico Francis Galton (16 de febrero de 1822-17 de enero de 1911) en el siglo XIX, para describir un fenómeno biológico, al comparar la estatura de padres e hijos, cuyo resultado fue que los hijos cuyos padres tenían una estatura muy superior al valor medio tendían

a igualarse a éste, mientras que aquellos cuyos padres eran muy bajos tendían a reducir su diferencia respecto a la estatura media; es decir, *regresaban* al promedio. Para Galton, la regresión solo tenía este significado biológico; sin embargo, su trabajo fue ampliado tiempo después por los estadísticos George Udny Yule (18 de febrero de 1871-26 de junio de 1951), de origen escocés y Karl Pearson (27 de marzo de 1857-27 de abril de 1936), de Gran Bretaña, a un contexto estadístico más general. En lo que respecta al término lineal, aparece debido al tipo de funciones que intervienen en el modelo de regresión.

Entre 1950 y 1960, los economistas usaban calculadoras de mesa electromecánicas para calcular las regresiones; hasta antes de 1970, los cálculos en ocasiones tardaban un día entero para obtener el resultado de una regresión; hoy día, con la gran velocidad de los procesadores, estos cálculos se realizan en un tiempo mucho más reducido.

En la actualidad, los métodos de regresión constituyen un área de investigación activa. De unas décadas a la fecha, se han desarrollado nuevos métodos para la regresión robusta: regresión que implica respuestas correlacionadas como series de tiempo y curvas de crecimiento; regresión en la que el predictor o variables de respuesta son curvas, imágenes, gráficos u otros objetos de datos complejos; métodos de regresión para acomodar diversos tipos de datos que faltan (datos faltantes); regresión no paramétrica, métodos bayesianos para la regresión; regresión en la que las variables de predicción se miden con error; regresión con más variables predictoras que las observaciones y la inferencia causal con la regresión.

En esta unidad tratamos otra área de la estadística, que como ya hemos visto, lleva el nombre de regresión lineal; también revisamos los fundamentos para un análisis de regresión simple y múltiple, ambos obtenidos mediante el método más usual para determinar el mínimo de la suma de los cuadrados de los errores: **método de mínimos cuadrados**, introducido en 1805.

El desarrollo de la unidad lo hacemos en dos partes; en la primera, nos restringimos al análisis del comportamiento lineal entre dos variables, para observar si existe relación entre éstas: regresión lineal simple. En la segunda parte desarrollamos la generalización del problema al caso de varias variables independientes: regresión lineal múltiple.

Para el análisis de la regresión simple con los datos muestrales, primero se ubican los datos de dos variables en un gráfico de dispersión, para determinar si existe un comportamiento lineal, en tal caso se procede al cálculo de la correlación lineal para tener cuantitativamente su grado de relación; después, con el propósito de predecir los valores de una variable, se procede a crear el modelo **recta de regresión**, para *ajustar* los datos graficados en la dispersión, y con la evidencia muestral se aplica una prueba de hipótesis estadística para determinar si una relación muestral puede o no extenderse a toda la población.

También, vemos otro concepto estadístico para determinar la posible relación entre dos variables, refiriendo al coeficiente de correlación entre variables. En el desarrollo de la unidad entenderemos que el modelo de regresión lineal junto con el análisis de correlación constituyen una herramienta importante en todos los campos de la ciencia, cuyos resultados provengan de variables cuantitativas, tanto discretas como continuas.

## 6.1 Regresión lineal simple

Uno de los principales problemas que trata la estadística consiste en proponer modelos que ayuden a comprender un fenómeno aleatorio. En esta unidad tratamos un tipo de problema estadístico en el que deseamos conocer cómo influyen diferentes valores  $x_1, x_2, \dots, x_n$ , que han sido seleccionados de forma independiente de una variable controlable (no aleatoria), para predecir o estimar un valor medio o uno futuro de una **variable dependiente**  $y$ , también conocida como **variable de respuesta**.

Por ejemplo, el gerente de la empresa de juguetes  $A$  desea conocer cómo influye en sus ventas diarias el tiempo al aire en minutos, en el que se promociona el juguete en alguna televisora. Para tal efecto, decide contratar un espacio publicitario en el canal 5 de una importante televisora durante los siguientes  $n$  días, para que transmita  $x_1, x_2, \dots, x_n$  minutos al día, respectivamente, la propaganda del juguete. Y obtiene los datos que se muestran en la tabla 6.1.

**Tabla 6.1** Tiempo diario en minutos de publicidad del juguete A y volumen de ventas

Día $i$	Minutos al día de publicidad ( $x$ )	Volumen diario de ventas en millones de pesos ( $y$ )
1	$x_1$	$y_1$
2	$x_2$	$y_2$
$\vdots$	$\vdots$	$\vdots$
$n$	$x_n$	$y_n$

En este ejemplo consideramos la variable independiente  $x$ , cuyos valores son determinados por el gerente, quien decide cuántos minutos desea pagar al día para que el canal referido lo promocióne (note que  $x$  no es una variable aleatoria, pues sus valores son controlados por el gerente). La variable de respuesta que desea medir es el volumen diario de ventas en millones de pesos que recibe por la comercialización del juguete (note que esta variable sí es aleatoria, ya que sus valores no pueden ser controlados por el gerente o la televisora).

Algunas de las preguntas más comunes que puede tener el gerente en este ejemplo son: ¿será posible establecer alguna relación entre ambas variables  $x$ ,  $y$  mediante un modelo estadístico? En caso de ser posible, ¿cómo puede encontrar el mejor modelo que represente esta relación?

Cuando el gerente logre responder las preguntas formuladas, uno de los objetivos inmediatos consiste en poder determinar, con el modelo, para un valor particular de  $x$  (tiempo de publicidad en minutos por la televisora) un valor medio de las ventas diarias que se aproxime lo más posible al valor real de ventas que recibe por el juguete.

Los modelos que se emplean para relacionar una variable dependiente y con otra u otras variables independientes  $x_1, x_2, \dots, x_m$  se denominan **modelos de regresión** o **modelos estadísticos** porque expresan el valor medio de  $y$  para valores dados de  $x_1, x_2, \dots, x_m$ .

A partir de la definición anterior, y de acuerdo con la relación que se encuentre entre las variables, se clasifica el tipo de regresión. Por ejemplo, si la relación entre ambas variables  $x$  y  $y$  es lineal, al modelo se le llama **modelo de regresión lineal**.

En un modelo de regresión, a la variable por predecir o por modelar,  $y$ , la denominamos **variable dependiente** o **de respuesta**, y a las variables que se utilizan para predecir o modelar a  $y$  las denominamos **variables independientes** o **predictoras**.

Antes de continuar con la terminología de un modelo de regresión debemos tener en cuenta que para elaborar uno en dos variables,  $x$  y  $y$ , necesitamos tener conocimientos previos sobre el comportamiento de la dependencia entre éstas. Esto se puede lograr si trazamos un diagrama de dispersión que despliegue la relación  $(x, y)$  en términos gráficos. Así, con base en la correspondencia que muestre la gráfica será el tipo de modelo de regresión que se proponga, en esta parte de la unidad son de interés las relaciones lineales entre las variables  $x$  y  $y$ .

## Diagrama de dispersión

El diagrama de dispersión es una representación gráfica de dos variables cuantitativas que se analizan de manera simultánea, en general se denotan por  $x$  y  $y$ . La característica de estos gráficos es que los datos se presentan en forma de puntos, sin estar unidos por segmentos de recta. La escala del eje  $X$  contiene el rango de los valores necesarios para la variable  $x$ ; por su parte, el eje  $Y$  también tiene una escala adecuada para los valores de  $y$ . Los pares de datos se representan de manera gráfica en un sistema de dos dimensiones. Los diagramas de dispersión se pueden trazar en Excel.

## Ejemplo 6.1

Un gerente quiere saber si el volumen semanal de ventas en millones de pesos de su empresa se puede ajustar a una línea recta con el número de anuncios de publicidad para televisión. Con los datos de la tabla 6.2 trace una gráfica de dispersión.

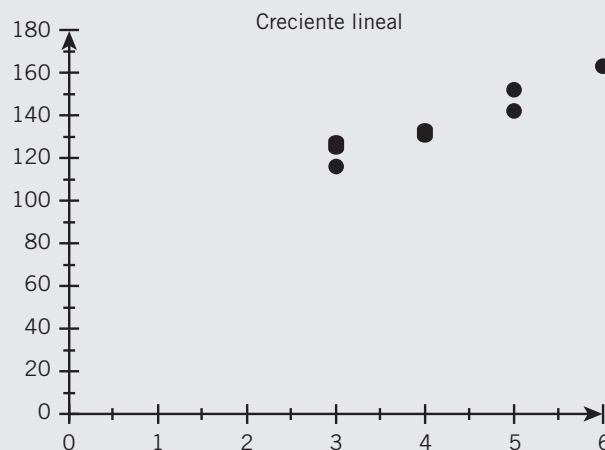
**Tabla 6.2** Anuncios y volumen de ventas

Observación $i$	Anuncios de publicidad ( $x$ )	Volumen semanal de ventas en millones de pesos ( $y$ )
1	3	125
2	5	152
3	4	131
4	4	133
5	5	142
6	3	116
7	3	127
8	6	163

### Solución

En este caso, los ocho datos puntuales  $(x, y)$  se pueden representar en un espacio de dos dimensiones para formar un diagrama de dispersión de datos. Primero, se escogen los anuncios de publicidad como la variable  $x$ , por lo que la escala en el eje  $X$  está en enteros. Luego, se determina que el volumen de ventas es la variable  $y$ , así que la escala del eje  $Y$  está en millones de pesos. Cada dato puntual se representa de manera gráfica en el sistema de coordenadas (véase figura 6.1).

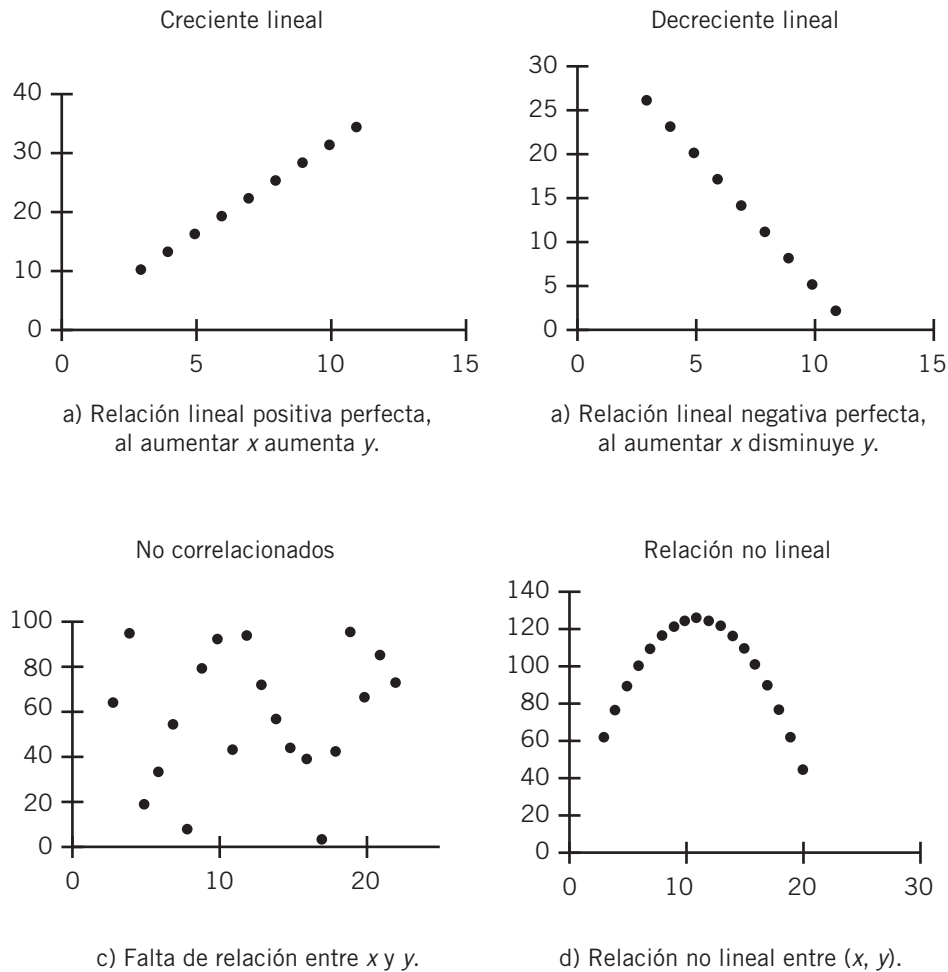
La ventaja de un diagrama de dispersión consiste en permitir visualizar la relación  $(x, y)$  en forma gráfica. Esto se aprecia en la figura 6.1, puesto que conforme aumentan los anuncios de publicidad también lo hace el volumen de ventas. Luego, puede existir una relación lineal entre  $x$  y  $y$ , pero antes de tomar la decisión sobre la existencia de una relación, se recomienda elegir una muestra mucho más grande y ver si el comportamiento se conserva.



**Figura 6.1** Diagrama de dispersión.



Antes se mencionó que una ventaja de los diagramas de dispersión es que permiten observar la relación entre las dos variables de interés. Es decir, si la relación es o no lineal o qué forma posible tiene. La figura 6.2 muestra algunos patrones de comportamiento para examinar una relación  $(x, y)$ .



**Figura 6.2** Diagramas de dispersión que muestran algunos patrones de la relación entre  $x$  y  $y$ .

## Supuestos de la variable dependiente en el análisis de regresión

Antes de iniciar el tratamiento del modelo de regresión, es necesario mencionar los objetivos principales y enumerar los supuestos bajo los cuales se rige el modelo. Como se dijo, los símbolos elegidos para las variables independiente y dependiente son  $x$  y  $y$ , respectivamente. Por otro lado, resulta que cuando se identifica solo una variable independiente, el análisis se llama **regresión simple**. Así, el tipo más sencillo de una curva de aproximación de un modelo de regresión es la línea recta. Ahora bien, podemos plantearnos la siguiente pregunta: ¿cuál es el objetivo principal de un análisis de regresión?

El objetivo principal del análisis de regresión es predecir el valor de una variable (dependiente) dado el valor de otra asociada (independiente).

El término análisis de regresión simple indica que los valores de la variable dependiente se predicen sobre la base de los de una sola variable independiente, mientras que el análisis de regresión múltiple se relaciona con la predicción de los valores de la dependiente sobre la base de los valores de dos o más variables independientes. Pero, una pregunta obligada antes de iniciar la búsqueda del mejor modelo lineal de ajuste es: ¿bajo qué supuestos de la variable dependiente se formula un modelo de regresión?

**Supuestos generales de la variable dependiente en el modelo de regresión lineal simple**

1. **La variable dependiente es una variable aleatoria.** Esto implica que aunque los valores de la variable independiente pueden ser designados, los de la dependiente deben obtenerse por medio del proceso de muestreo.
2. **Las variables independientes y dependientes están asociadas de manera lineal.**
3. Las varianzas de las distribuciones condicionales de la variable dependiente, dados valores diferentes de la variable independiente, son iguales (**homocedasticidad**).
4. La distribución condicional de la variable dependiente, dados valores diferentes de la variable independiente, es normal.

Es posible determinar si las variables se relacionan de manera lineal o no, al construir un diagrama de dispersión.

**Supuestos del error en un modelo lineal**

El tipo más sencillo de curva de aproximación en un modelo es la línea recta. Cuando se examina la relación de dos variables, en general, se hace con el propósito de usar una para pronosticar la otra. La mayor parte de los estudios de regresión se inician con el deseo de examinar y explicar el valor cambiante de esta variable, la cual, como vimos en el análisis de regresión, se llama variable dependiente. Por último, podemos decir que cuando se identifica solo una variable independiente, el análisis se llama regresión simple.

Un modelo de regresión lineal simple, para toda la población está dado en (6.1):

$$y = \beta_0 + \beta_1 x + \varepsilon \quad (6.1)$$

donde  $\varepsilon$  (letra griega *épsilon*) es un error aleatorio;  $\beta_0$  y  $\beta_1$  son parámetros desconocidos, tales que  $\beta_0$  = ordenada al origen (intersección con el eje  $Y$ ) y  $\beta_1$  = pendiente de la recta.

Una pregunta importante en el análisis de regresión de un problema de tipo empresarial es: ¿se puede pronosticar el valor exacto del ingreso sobre las ventas ( $y$ ) si se especifica el precio por unidad ( $x$ )?

Cuando se conoce el modelo de ajuste, es posible pronosticar el valor exacto. Las ventas dependen de muchas variables distintas al precio por unidad, como gastos de publicidad, época del año, estado de la economía, etcétera.

Un modelo de regresión lineal supone una relación y algún error aleatorio,  $\varepsilon$ , que representa al que ocurre cuando se usa una variable independiente para pronosticar la variable dependiente. Este término de error tiene en cuenta las variables independientes que afectan a  $y$ , pero que no están incluidas en el modelo. También considera el factor de variabilidad aleatoria o probabilística. Así podemos concluir que  $\varepsilon$  incluye dos tipos de error: el error del modelo (lo que significa que no todas las variables independientes relevantes están incluidas) y el error aleatorio.

La distribución probabilística de  $\varepsilon$  determina el grado en que el modelo de regresión describe la relación entre las variables independientes y dependientes.

**Supuestos generales sobre la distribución de probabilidad de  $\varepsilon$ :**

1. La distribución de probabilidad de  $\varepsilon$  es normal.
2. La varianza de la distribución de  $\varepsilon$  es constante para todos los valores de  $x$ .
3. La media de las distribuciones de probabilidad de  $\varepsilon$  es 0. Es decir, el valor medio de  $y$  para un valor dado de  $x$  es:

$$E(y) = E(\beta_0 + \beta_1 x + \varepsilon) = \beta_0 + \beta_1 x$$

4. Los valores de  $\varepsilon$  son independientes entre sí. Esta suposición implica que se ha seleccionado una muestra aleatoria de elementos de una población para medirlos.

## Ejemplo 6.2

Suponga que los gastos de publicidad semanales se usan para determinar el efecto de la publicidad sobre las ventas. La relación entre el ingreso por ventas ( $y$ ) y el gasto en publicidad ( $x$ ) se expresa como el modelo estadístico:

$$y = \beta_0 + \beta_1 x + \varepsilon = 0 + 20x + \varepsilon$$

Observe que la componente del modelo  $20x$  indica que si se gasta \$1 en publicidad, el ingreso por ventas será igual a \$20. Sin embargo, el error aleatorio en el modelo indica que el ingreso por ventas no se relaciona de manera precisa con el gasto de publicidad. Es decir, la componente de error aleatorio ( $\varepsilon$ ) indica que el ingreso por ventas puede depender de otras variables distintas al gasto de publicidad.

## 6.2 Método de mínimos cuadrados para optimizar el error

Hasta el momento hemos hablado acerca de los modelos de regresión lineal, pero aún no hemos visto cómo calcular sus coeficientes o parámetros. Para tal efecto requerimos de algún método de optimización para determinar cuál de todas las curvas de aproximación a una serie de datos puntuales es la que *mejor "ajusta"* a los datos.

Antes, indicamos que al ajustar los datos con un modelo se puede cometer un error aleatorio, de manera que la recta que mejor lo haga será aquella que proporcione un menor error. Así, si se recurre a un método de optimización de errores, por ejemplo al de mínimos cuadrados, podemos buscar el mínimo de la suma de los cuadrados de los errores.

En general, no se conocen los valores exactos de los parámetros de regresión,  $\beta_0$  y  $\beta_1$ , ni del error  $\varepsilon$ . Por tal razón, debido a que se trata de situaciones aleatorias en los errores y variables dependientes, buscamos estimaciones de estos parámetros a partir de datos muestrales, con lo cual determinamos la línea recta que *mejor ajusta* a este conjunto de puntos y la llamamos **recta de regresión muestral**:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad (6.2)$$

donde:

$\hat{y}$  = valor pronosticado de la variable dependiente.

$x$  = variable independiente.

$\hat{\beta}_0 = b_0$  = valor estimado de la ordenada al origen de la población.

$\hat{\beta}_1 = b_1$  = valor estimado de la pendiente de la recta poblacional.

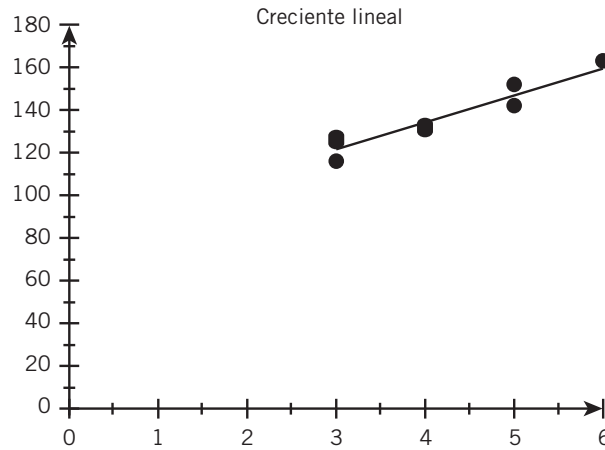
La recta determinada por la ecuación  $\hat{y} = b_0 + b_1 x$  debe pasar entre los puntos de los datos de manera que pueda predecirse el valor  $y$  para un valor dado de  $x$ . Obsérvese que  $\hat{y}$  (el valor pronosticado de la variable dependiente) en realidad es el valor promedio de  $y$  obtenido por los valores de  $x$ . Además, para cualquier valor específico de  $x$  en la muestra existen dos valores  $y$  asociados: el real observado de  $y$  (que corresponde al valor observado de  $x$ ) y la media pronosticada de  $y$  para el valor  $x$ .

La diferencia ( $y - \hat{y} = e$ ) mide el error que ocurre al pronosticarse la variable dependiente.

Como  $(x, y)$  es un punto en el diagrama de dispersión y  $(x, \hat{y})$  es un punto sobre la recta de regresión  $\hat{y} = b_0 + b_1 x$ , la recta de regresión muestral es la línea recta que mejor se ajusta a un conjunto de puntos  $(x, y)$ .

El método de mínimos cuadrados es un procedimiento matemático utilizado para encontrar la ecuación de la línea recta que minimiza la suma de los cuadrados de los errores de los pronósticos medidos en la dirección vertical ( $y$ ).

Por ejemplo, en el caso de los anuncios con volumen de ventas del ejemplo 6.1, la recta que mejor ajusta, según los datos, se presenta en la gráfica de la figura 6.3. Una pregunta que surge en este momento es: ¿cómo determinar la recta que mejor ajuste a los datos?



**Figura 6.3** Recta de mínimos cuadrados para el ajuste de los puntos  $(x, y)$ .

La deducción de las fórmulas necesarias para calcular el valor de los estimadores de los parámetros  $\beta_0$  (ordenada al origen) y  $\beta_1$  (pendiente de la recta de regresión) puede hacerse de varias formas, entre las más usuales están el *método de mínimos cuadrados* y el *método de máxima verosimilitud*. En el primero no se requieren restricciones, mientras que en el segundo se necesita que los errores tengan distribución normal con media cero y varianza constante.

### Teorema 6.1

Sean las parejas de observaciones  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ , donde  $x_i$  representa el valor de la variable independiente  $x$  con valor de respuesta  $y_i$ , entonces los valores de los estimadores de  $\beta_0$  y  $\beta_1$  que mejor ajustan un modelo lineal están dados en (6.3) y (6.4):

$$\hat{\beta}_1 = b_1 = \frac{n \sum_{i=1}^n y_i x_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2} \quad (6.3)$$

$$\hat{\beta}_0 = b_0 = \bar{y} - b_1 \bar{x} \quad (6.4)$$

En este momento cabe mencionar que cuando los errores aleatorios del modelo tienen distribución normal es el único caso en el que coinciden los resultados de mínimos cuadrados con el de máxima verosimilitud.

#### Demostración

Se pide encontrar los valores de los estimadores para los parámetros  $\beta_0$  y  $\beta_1$  que mejor ajusten los datos a una línea recta. Es decir, minimizar los errores  $y - \hat{y}$ , donde  $\hat{y} = b_0 + b_1 x$  y  $y = \beta_0 + \beta_1 x + \varepsilon$ .

Al sumar los cuadrados de los errores de todas las observaciones:

$$\begin{aligned} \sum_{i=1}^n e_i^2 &= \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2 = \sum_{i=1}^n (y_i^2 + b_0^2 + b_1^2 x_i^2 - 2b_0 y_i - 2b_1 y_i x_i + 2b_0 b_1 x_i) \\ &= \sum_{i=1}^n y_i^2 + n b_0^2 + b_1^2 \sum_{i=1}^n x_i^2 - 2b_0 \sum_{i=1}^n y_i - 2b_1 \sum_{i=1}^n y_i x_i + 2b_0 b_1 \sum_{i=1}^n x_i \end{aligned}$$

Si se deriva de manera parcial con respecto de  $b_0$  y  $b_1$  tenemos:

$$\frac{\partial}{\partial b_0} \sum_{i=1}^n e_i^2 = 2nb_0 - 2 \sum_{i=1}^n y_i + 2b_1 \sum_{i=1}^n x_i = 0$$

$$\frac{\partial}{\partial b_1} \sum_{i=1}^n e_i^2 = 2b_1 \sum_{i=1}^n x_i^2 - 2 \sum_{i=1}^n y_i x_i + 2b_0 \sum_{i=1}^n x_i = 0$$

Al despejar la variable  $b_0$  de la primera ecuación, tenemos:

$$b_0 = \bar{y} - b_1 \bar{x}$$

Si se sustituye esta expresión en la segunda ecuación y se divide entre 2, resulta:

$$b_1 \sum_{i=1}^n x_i^2 - \sum_{i=1}^n y_i x_i + \frac{\sum_{i=1}^n y_i - b_1 \sum_{i=1}^n x_i}{n} \sum_{i=1}^n x_i = 0$$

$$nb_1 \sum_{i=1}^n x_i^2 - n \sum_{i=1}^n y_i x_i + \sum_{i=1}^n x_i \sum_{i=1}^n y_i - b_1 \left( \sum_{i=1}^n x_i \right)^2 = 0$$

Luego, al despejar  $b_1$  resultan los dos estimadores:

$$\hat{\beta}_1 = b_1 = \frac{n \sum_{i=1}^n y_i x_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2}$$

$$\hat{\beta}_0 = b_0 = \bar{y} - b_1 \bar{x}$$

donde:

- $\sum_{i=1}^n x_i$  suma de valores de  $x$ ;  $\sum_{i=1}^n x_i^2$  suma de los cuadrados de los valores de  $x$ .
- $\sum_{i=1}^n y_i$  suma de valores de  $y$ ;  $\sum_{i=1}^n x_i y_i$  suma de productos de  $x$  y  $y$  para cada observación.
- $\bar{x}$  el promedio de los valores de  $x$ ;  $\bar{y}$  el promedio de los valores de  $y$ .
- $b_1$  estimador puntual de la pendiente a la recta de regresión  $\beta_1$ .
- $b_0$  estimador puntual de la ordenada al origen de la recta de regresión  $\beta_0$ .

### Ejemplo 6.3

El gerente de una empresa quiere saber si su volumen semanal de ventas en miles de dólares se puede ajustar a una línea recta con el número de anuncios de televisión para publicidad que se muestran en la tabla 6.2. Encuentre la recta de regresión que mejor ajuste las observaciones e interprete los valores.

#### Solución

Para el cálculo de los estimadores por mínimos cuadrados utilizamos las fórmulas del teorema 6.1, para lo cual completamos la tabla 6.2, agregando tres columnas para productos y cuadrados de las observaciones (véase tabla 6.3).

**Tabla 6.3** Productos y cuadrados de las observaciones de la tabla 6.2

Observación <i>i</i>	Anuncios de publicidad ( <i>x</i> )	Volumen semanal de ventas en miles de dólares ( <i>y</i> )	$x_i y_i$	$x_i^2$	$y_i^2$
1	3	125	375	9	15 625
2	5	152	760	25	23 104
3	4	131	524	16	17 161
4	4	133	532	16	17 689
5	5	142	710	25	20 164
6	3	116	348	9	13 456
7	3	127	381	9	16 129
8	6	163	978	36	26 569
<b>Sumas</b>	<b>33</b>	<b>1 089</b>	<b>4 608</b>	<b>145</b>	<b>149 897</b>

Entonces, los estimadores de los parámetros del modelo de regresión son:

$$\left\{ \begin{array}{l} b_1 = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2} = \frac{8(4608) - 33(1089)}{8(145) - 33^2} = 13.056 \\ b_0 = \bar{y} - b_1 \bar{x} = \frac{1}{8}(1089) - 13.056 \left( \frac{33}{8} \right) = 82.269 \end{array} \right.$$

La ecuación de la regresión lineal está dada por:

$$\hat{y} = 82.269 + 13.056x$$

Para la interpretación de los valores de la ecuación de regresión lineal, tenemos que la ordenada al origen,  $\hat{y} = 82.269$  es el valor esperado de  $y$  cuando  $x = 0$ . De la ecuación de regresión, se puede observar que el aumento de una unidad en  $x$  implica que el valor de  $y$  aumente en promedio 13.056. En términos prácticos, la ecuación de regresión sugiere que para cada comercial de televisión que se contrate se puede esperar un promedio de \$13.056 millones de ventas adicionales.

Esta información puede ser útil para planear el presupuesto de publicidad para años subsecuentes.

Los valores calculados para los estimadores también se pueden obtener con una calculadora de bolsillo y, por supuesto, con algún paquete estadístico.

#### Ejemplo 6.4

El analista de una empresa debe determinar si existe una relación positiva entre el material de desperdicio ( $x$ ), en miles de pesos, y las ventas ( $y$ ) de la línea complementaria con el uso de los desperdicios, en millones de pesos. Para ello, toma 12 observaciones con diferentes valores del desperdicio y calcula las ventas de línea complementaria, obteniendo los resultados que se muestran en las primeras tres columnas de la tabla 6.4. Trace un diagrama de dispersión y encuentre la recta de regresión que mejor ajuste las observaciones e interprete los valores obtenidos.

### Solución

En la tabla 6.4 se muestran los valores de las observaciones y los resultados de productos y cuadrados de cada variable.

**Tabla 6.4** Material de desperdicio y venta complementaria

Observación <i>i</i>	Material de desperdicio en miles de pesos ( <i>x</i> )	Ventas de línea complementaria en millones de pesos ( <i>y</i> )	$x_i y_i$	$x_i^2$	$y_i^2$
1	5.3	21	111.3	28.09	441
2	6.5	28	182	42.25	784
3	4.5	20	90	20.25	400
4	4.7	22	103.4	22.09	484
5	5.5	28	154	30.25	784
6	6.8	32	217.6	46.24	1024
7	7.2	35	252	51.84	1225
8	6.0	30	180	36	900
9	6.8	35	238	46.24	1225
10	5.1	24	122.4	26.01	576
11	4.6	17	78.2	21.16	289
12	5.7	24	136.8	32.49	576
<b>Sumas</b>	<b>68.7</b>	<b>316</b>	<b>1865.7</b>	<b>402.91</b>	<b>8708</b>

Con estos valores y las fórmulas (6.3) y (6.4) realizamos los cálculos de los estimadores para los parámetros correspondientes de la recta de regresión lineal muestral, de lo que se obtiene:

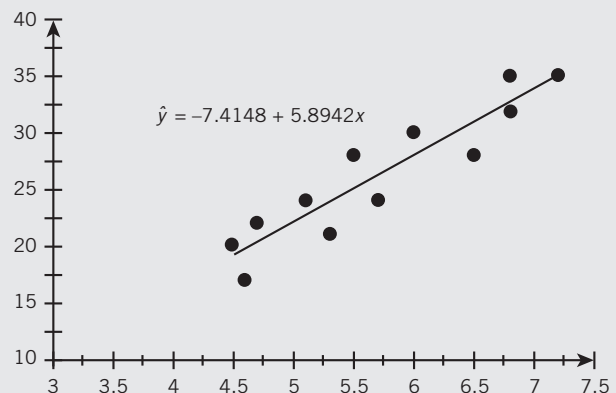
$$\left\{ \begin{array}{l} b_1 = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2} = \frac{12(1865.7) - 68.7(316)}{12(402.91) - 68.7^2} = 5.8942 \\ b_0 = \bar{y} - b_1 \bar{x} = \frac{1}{12}(316) - 5.8942 \left( \frac{68.7}{12} \right) = -7.4148 \end{array} \right.$$

La ecuación de la regresión lineal está dada por:

$$\hat{y} = -7.4148 + 5.8942x$$

Como  $b_1 = 5.8942$ , esto indica que existe una relación positiva entre el material de desperdicio y las ventas en la línea complementaria, lo que significa que un incremento en éstas se debe al aumento del material de desperdicio.

El diagrama de dispersión se muestra en la figura 6.4.



**Figura 6.4** Recta de mínimos cuadrados para el ajuste de los puntos (*x*, *y*).



## Ejercicios 6.1

1. Los datos que se muestran en la tabla 6.5 son las alturas en centímetros,  $x$ , y los pesos en kilogramos,  $y$ , de una muestra aleatoria de 10 empleadas de una gran empresa.

Tabla 6.5 Altura y peso de 10 empleadas de la empresa

Altura (cm)	173	170	165	173	163	170	168	165	163	173
Peso (kg)	57.1	56.6	61.9	64.8	59.0	67.2	60.0	63.4	56.6	62.4

- a) Calcule la recta de mínimos cuadrados e interprete los valores de  $b_0$  y  $b_1$ .  
 b) Estime el valor del peso medio para una empleada que tiene una estatura de 167 cm.
2. Las calificaciones de un grupo de nueve estudiantes en su reporte de medio año ( $x$ ) y en los exámenes finales ( $y$ ) dan los resultados de la tabla 6.6.

Tabla 6.6 Calificaciones de medio año y exámenes finales de los estudiantes

$x$	77	50	71	72	81	94	96	99	69
$y$	82	66	78	34	47	85	99	99	68

- a) Estime la recta de regresión e interprete los valores de  $b_0$  y  $b_1$ .  
 b) Estime la calificación final media de un estudiante que tuvo una calificación de medio año de 85 puntos.
3. El presidente de una agencia de autos piensa que el tiempo que un vendedor pasa con un cliente debe tener una relación positiva con el monto de lo que compra. Para corroborar si esta relación existe se reunieron los datos muestrales de la tabla 6.7.

Tabla 6.7 Tiempo de atención a un cliente y su monto de compras

Minutos que pasa un vendedor con un cliente ( $x$ )	108	132	62	95	58	134	87	78	120
Monto de la cuenta en miles de pesos ( $y$ )	856	825	651	348	294	242	78	112	159

- a) Trace un diagrama de dispersión y mencione qué tipo de relación existe entre las variables.  
 b) Determine la ecuación de regresión muestral e interprete los valores de  $b_0$  y  $b_1$ .  
 c) Estime el monto de compras medio de una persona atendida durante 80 minutos por un vendedor.
4. El contador de una mueblería debe estimar los costos generales según el número de sillas producidas. Los datos mensuales se recogen en la tabla 6.8, los cuales muestran los gastos generales y las sillas producidas en siete plantas diferentes.

Tabla 6.8 Gastos generales por sillas producidas en siete diferentes plantas

Número de sillas ( $x$ )	112	122	147	173	94	151	109
Gastos generales ( $y$ )	576	497	789	862	361	688	532

- a) Trace un diagrama de dispersión y mencione qué tipo de relación existe entre las variables.  
 b) Determine la ecuación de regresión muestral e interprete los valores de  $b_0$  y  $b_1$ .  
 c) Calcule la estimación puntual media de los gastos generales si se producen 150 sillas.
5. Para modelar la relación entre la resistencia de corte media  $E(y)$  de las puntas de albañilería y el esfuerzo de precompresión  $x$ , Riddington y Ghazali presentaron, en 1990, los resultados de una serie de pruebas de esfuerzo

con tabiques sólidos dispuestos en tripletas unidas con mortero, donde se varió el esfuerzo de precompresión para cada triplete y se registró la carga de corte máxima justo antes de la ruptura, resistencia de corte (véase tabla 6.9).

**Tabla 6.9** Resistencia de corte y esfuerzo de precompresión con tabiques sólidos

Prueba de triplete	1	2	3	4	5	6	7
Esfuerzo de precompresión	0	0.60	1.20	1.33	1.43	1.75	1.75
Resistencia al corte (N/mm <sup>2</sup> )	1.00	2.18	2.24	2.41	2.59	2.82	3.06

Fuente: J. R. Riddington y M. Z. Ghazali (1990). *Ice Proceedings*, vol. 89, issue 1, 1 de marzo de 1990, pp. 89-102.

- Trace un diagrama de dispersión y mencione qué tipo de relación existe entre las variables.
  - Determine la ecuación de regresión muestral e interprete los valores de  $b_0$  y  $b_1$ .
  - Calcule la resistencia media al corte para un esfuerzo de precompresión de 1.50.
6. Penner y Watts realizaron un experimento para determinar si el tiempo,  $y$ , que se requiere para taladrar una distancia de cinco pies en roca aumenta con la profundidad de la región de taladro,  $x$ . Una parte de los resultados del experimento se muestran en la tabla 6.10.

**Tabla 6.10** Resistencia de corte y esfuerzo de precompresión con tabiques sólidos

Obs.	Profundidad a la que se inicia el taladrado, $x$ , en pies	Tiempo de taladrar 5 pies, $y$ , en minutos	Obs.	Profundidad a la que se inicia el taladrado, $x$ , en pies	Tiempo de taladrar 5 pies, $y$ , en minutos
1	0	4.90	11	100	5.17
2	10	6.77	12	110	6.84
3	20	6.99	13	120	7.03
4	30	6.07	14	130	7.27
5	40	5.49	15	140	7.15
6	50	6.19	16	150	7.05
7	60	6.27	17	160	6.76
8	70	6.34	18	170	7.17
9	80	5.70	19	180	7.07
10	90	6.60	20	190	6.91

Fuente: R. Penner y D.G. Watts, (1991). "Minng Information", *The American Statistician*, vol. 45, núm. 1, febrero de 1991, p. 6.

- Trace un diagrama de dispersión y mencione qué tipo de relación existe entre las variables.
  - Determine la ecuación de regresión muestral e interprete los valores de  $b_0$  y  $b_1$ .
  - Calcule el tiempo medio de taladrar cinco pies cuando se inicia a una profundidad de 145 pies.
7. En la Ciudad de México se realizó un estudio para conocer si existe relación entre los ingresos de los trabajadores,  $x$ , y los pagos que hacen con tarjetas de crédito,  $y$ . Los resultados se muestran en la tabla 6.11.

**Tabla 6.11** Ingresos y pagos mensuales con tarjetas de crédito

Trabajadores	Ingresos mensuales, $x$ , en miles de pesos	Pagos mensuales en pesos con tarjetas de crédito, $y$	Trabajadores	Ingresos mensuales, $x$ , en miles de pesos	Pagos mensuales en pesos con tarjetas de crédito, $y$
1	6.8	1 575	16	10.1	3 142
2	6.9	1 912	17	10.2	3 179
3	7.4	2 439	18	10.4	3 291
4	7.4	2 451	19	10.5	3 296
5	7.8	2 569	20	10.6	3 345
6	7.9	2 636	21	10.9	3 357
7	8.2	2 643	22	11.4	3 427
8	8.5	2 800	23	11.8	3 446
9	8.5	2 916	24	11.8	3 539
10	8.9	2 955	25	12.6	3 638
11	9.1	2 961	26	12.8	3 685
12	9.3	2 964	27	13.1	3 859
13	9.4	2 987	28	13.5	3 920
14	9.6	3 029	29	13.8	4 028
15	9.8	3 061	30	15.3	4 136

- a) Trace un diagrama de dispersión y mencione qué tipo de relación existe entre las variables.
- b) Determine la ecuación de regresión muestral e interprete los valores de  $b_0$  y  $b_1$ .
- c) Calcule el pago medio con tarjetas de crédito para un trabajador con un sueldo de \$10 000 mensuales.
8. Se realizó un estudio para determinar si existe relación entre la edad de un trabajador y las semanas que queda desempleado después de ser despedido. Los resultados del estudio se muestran en la tabla 6.12.

**Tabla 6.12** Edad del trabajador y semanas que dura desempleado

Trabajador	Edad en años del trabajador ( $x$ )	Semanas desempleado ( $y$ )	Trabajador	Edad en años del trabajador ( $x$ )	Semanas desempleado ( $y$ )
1	24.0	1.0	11	34.0	10.5
2	25.0	2.0	12	35.0	11.0
3	25.5	2.9	13	36.0	12.0
4	26.0	3.4	14	36.5	12.1
5	27.0	4.9	15	38.0	12.1
6	27.0	6.4	16	38.5	13.2
7	28.0	6.5	17	40.0	13.5
8	28.0	9.3	18	40.5	13.8
9	30.0	9.4	19	42.0	14.1
10	31.0	10.0	20	42.0	14.4

- Trace un diagrama de dispersión y mencione qué tipo de relación existe entre las variables.
- Determine la ecuación de regresión muestral e interprete los valores de  $b_0$  y  $b_1$ .
- Calcule el tiempo medio que estará desempleado un trabajador de 32 años que acaba de ser despedido.

## 6.3 Error estándar de estimación y propiedades de los estimadores

Iniciamos esta sección con la definición del residual como la desviación vertical de la  $y$  observada a partir de la recta de regresión muestral que es conocida. Así, un residual es la diferencia entre un valor real  $y$  y el valor  $\hat{y}$  pronosticado por la ecuación de regresión muestral. Ya hemos visto que la ecuación  $\varepsilon = y - \hat{y}$  se usa para calcular un residual.

Cabe aclarar que el residual es diferente al término de error del modelo  $\varepsilon$ , el cual representa una variable aleatoria, mientras que el residual no es más que la desviación vertical de  $y$  a partir de la recta de regresión poblacional desconocida.

### Ejemplo 6.5

Para el problema del presidente de una empresa donde el volumen semanal de ventas se reporta y proporciona en miles de dólares y cuyos valores se muestran en la tabla 6.1, al ajustar la recta de regresión del número de anuncios de televisión para publicidad con el volumen de ventas (véase ejemplo 6.3) obtuvimos:

$$\hat{y} = 82.269 + 13.056x$$

Ahora, vamos a calcular los residuales con la recta de regresión ajustada.

### Solución

Para calcular los residuales, primero obtenemos los valores de los pronósticos según la recta de regresión lineal; después sus diferencias, como se muestra en la tabla 6.13.

**Tabla 6.13** Cálculo de residuales

Observación i	Anuncios de publicidad ( $x$ )	Volumen semanal de ventas en miles de dólares ( $y$ )	$\hat{y}_i = 82.269 + 13.056x_i$	$e_i = y_i - \hat{y}_i$
1	3	125	121.4	3.6
2	5	152	147.5	4.5
3	4	131	134.5	-3.5
4	4	133	134.5	-1.5
5	5	142	147.5	-5.5
6	3	116	121.4	-5.4
7	3	127	121.4	5.6
8	6	163	160.6	2.4

En la columna 4 de la tabla 6.13 se muestra el valor de  $y$  pronosticado que se calcula al sustituir el valor de  $x$  en la ecuación de regresión. Por ejemplo, el último valor de  $\hat{y}$  se calcula a partir de la ecuación de regresión con el valor muestral de  $x = 6$ :

$$\hat{y} = 82.269 + 13.056(6) = 160.6$$

El cálculo del último residual  $e_8 = y - \hat{y} = 163 - 160.6 = 2.4$  se encuentra al final de la columna 5 de la tabla 6.13. Ahora, trazamos la gráfica de los valores y la recta que ajusta los datos junto con sus residuales (véase figura 6.3).

Ya hemos visto en las medidas de dispersión que la desviación estándar de un conjunto de datos representa una medida de variabilidad o dispersión de los datos alrededor de la media. De igual manera, el error estándar de estimación se usa para medir la variabilidad o dispersión de los valores de  $y$  observados en la muestra alrededor de la recta de regresión.

Así, el error estándar de estimación lo denotaremos por  $s$  y lo calculamos mediante (6.5):

$$s = \sqrt{\frac{SCE}{n-2}} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^n e_i^2}{n-2}}, \text{ para } n > 2 \quad (6.5)$$

donde:

$$s = \text{error estándar de estimación} \quad SCE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2$$

$y_i$  = valores muestrales de  $y$

$\hat{y}_i$  = valores de  $y$  calculados con la ecuación de regresión

$n$  = tamaño de la muestra

El valor  $n - 2$  representa el número de grados de libertad de los residuales alrededor de la recta de regresión ajustada.

### Ejemplo 6.6

Usando los datos de la tabla 6.13 del ejemplo 6.5, calcule el error estándar de la estimación.

#### Solución

Como se puede ver, en la tabla 6.2 agregamos tres columnas; la primera para los valores pronosticados con la recta de regresión, la segunda para los residuales y la tercera para el cuadrado del residual (véase tabla 6.14).

**Tabla 6.14** Cálculo de residuales del ejemplo 6.6

Observación i	Anuncios de publicidad (x)	Volumen semanal de ventas en miles de dólares (y)	$\hat{y}_i = 82.269 + 13.056x_i$	$e_i = y_i - \hat{y}_i$	$e_i^2$
1	3	125	121.4	3.6	12.7
2	5	152	147.5	4.5	19.8
3	4	131	134.5	-3.5	12.2
4	4	133	134.5	-1.5	2.2
5	5	142	147.5	-5.5	30.8
6	3	116	121.4	-5.4	29.6
7	3	127	121.4	5.6	30.9
8	6	163	160.6	2.4	5.7
				<b>Suma</b>	<b>144</b>

Como se puede ver, la última columna contiene la suma de los cuadrados de los residuales  $y$ , por construcción, este valor minimiza el procedimiento de mínimos cuadrados para el ajuste de una línea recta a los datos.

Si se emplea la fórmula (6.5) para el error estándar de estimación y utilizamos los cálculos realizados en la tabla 6.14 tenemos:

$$s = \sqrt{\frac{SCE}{n-2}} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^n e_i^2}{n-2}} = \sqrt{\frac{144}{6}} = 4.9$$

donde,  $s = 4.9$  es el error estándar de la estimación para la recta de regresión, es decir, la desviación estándar o típica entre los valores muestrales  $y$ .

### Ejemplo 6.7

Al gerente de una empresa le fue entregado un reporte por parte de uno de los empleados acerca de la relación entre los costos generales y los distintos costos de operación de la empresa. Para su realización, el empleado utilizó datos de 12 meses consecutivos sobre costos generales y horas-máquina, con los que estableció la ecuación de regresión,  $\hat{y} = 72794 + 74.72x$  que representa las horas-máquina para estimar los costos generales y tiene anotado el error estándar de  $s = 9799$ . Sin embargo, el empleado ya no labora en la empresa y el gerente requiere de una interpretación de los resultados.

#### Solución

La ordenada al origen (\$72 794) es una estimación del total de los costos generales fijos mensuales. La pendiente de la ecuación de regresión (\$74.72) es la tasa de aplicación de los costos generales variables (es decir, \$74.72 por máquina cada hora). Para cada hora-máquina adicional ( $x$ ) se puede esperar que los costos generales medios mensuales,  $b_1$ , aumenten en promedio \$ 74.72.

Por otro lado, el tamaño del error estándar de la estimación, que es de \$9799, se puede interpretar como la cantidad en que por lo regular difieren los valores muestrales y costos generales de la estimación de regresión. Es decir, se puede esperar que los valores estimados o predecibles de  $y$  tengan un error similar.

Para finalizar la sección formularemos un resultado que muestra las propiedades estadísticas deseables de los estimadores puntuales. Para ello, agregamos la siguiente notación:

$$SC_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = n \text{cov}(x, y) \quad - n \text{ veces la covarianza entre } x \text{ y } y \quad (6.6)$$

$$SC_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = nS_n^2(x) \quad - n \text{ veces la varianza sesgada de } x \quad (6.7)$$

$$SC_y = \sum_{i=1}^n (y_i - \bar{y})^2 = nS_n^2(y) \quad - n \text{ veces la varianza sesgada de } y \quad (6.8)$$

### Teorema 6.2

Sean las parejas de observaciones  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ , donde las  $x_i$  representan los valores de la variable independiente,  $x$ ; con valores de respuesta,  $y_i$ , y con  $y = \beta_0 + \beta_1 x + \varepsilon$ , el modelo de regresión y las variables aleatorias,  $y, \varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ , cumplen con los supuestos de un modelo lineal. Entonces, los valores de los estimadores de  $\beta_0$  y  $\beta_1$  encontrados en el teorema 6.1 son los mejores estimadores lineales insesgados (MELI).

$$\hat{\beta}_1 = \frac{SC_{xy}}{SC_{xx}} = \frac{\text{cov}(x, y)}{S_n^2(x)} \Rightarrow E(\hat{\beta}_1) = \beta_1 \quad y \quad V(\hat{\beta}_1) = \frac{\sigma^2}{SC_{xx}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \Rightarrow E(\hat{\beta}_0) = \beta_0 \quad y \quad V(\hat{\beta}_0) = \left( \frac{\sigma^2}{n} \right) \frac{1}{SC_{xx}} \sum_{i=1}^n x_i^2$$

$$s^2 = \frac{SCE}{n-2} = \frac{\sigma^2 \chi_{n-2}^2}{n-2} \Rightarrow E(s^2) = \sigma^2$$

## Ejercicios 6.2

En los ocho ejercicios de la lista de ejercicios 6.1 calcule:

- Los residuales.
- El error estándar de estimación.

## 6.4 Prueba de hipótesis para el parámetro de la pendiente

En un modelo de regresión lineal (6.1), el parámetro de la pendiente representa los cambios que sufre la variable dependiente por cada uno en los valores de la variable independiente, mientras que el parámetro  $\beta_0$  es independiente con respecto a los cambios en  $x$ . En otras palabras, se requiere estudiar el comportamiento que tiene el MELI de  $\beta_1$  para determinar si es adecuado; es decir, si representa las variaciones en  $y$  por cada variación de  $x$ .

En el texto, en la parte de estadística inferencial, estudiamos varias técnicas para realizar inferencia sobre los parámetros basadas en los diferentes estimadores. Con los resultados del teorema 6.2 vemos que  $\hat{\beta}_1$  cumple varias propiedades importantes de un buen estimador que resumimos en el teorema 6.3 a fin de hacer inferencias sobre el parámetro.

### Teorema 6.3

Sean las parejas de observaciones  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ , donde las  $x_i$  representan los valores de la variable independiente  $x$  con valores de respuestas  $y_i$ , con  $y = \beta_0 + \beta_1 x + \varepsilon$ , el modelo de regresión y las variables aleatorias,  $y, \varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$  cumplen con los supuestos de un modelo lineal. Entonces, del teorema 6.2 tenemos:

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{SC_{xx}}\right) \Rightarrow \left(\frac{\hat{\beta}_1 - \beta_1}{\sigma}\right) \sqrt{SC_{xx}} \sim N(0, 1)$$

$$s^2 \sim \frac{\sigma^2}{n-2} \chi_{n-2}^2 \Rightarrow \left(\frac{\hat{\beta}_1 - \beta_1}{s}\right) \sqrt{SC_{xx}} \sim t_{n-2}$$

A partir del teorema 6.3 y las técnicas de inferencia que se estudian en las unidades 3 y 4 es posible deducir con facilidad las fórmulas para intervalos de confianza y pruebas de hipótesis del parámetro media de una distribución normal.

### Teorema 6.4

En las condiciones del teorema 6.3, los límites del intervalo al  $(1 - \alpha)$  100% de confianza para  $\beta_1$  están dados en (6.9):

$$\hat{\beta}_1 \pm \frac{s}{\sqrt{SC_{xx}}} t_{n-2, \alpha/2} \quad (6.9)$$

donde  $t_{n-2, \alpha/2}$  es el valor de la distribución t-Student con  $n - 2$  grados de libertad y área derecha igual  $\alpha/2$ .



## Teorema 6.5

En las condiciones del teorema 6.3, el estadístico de prueba para los contrastes de hipótesis de  $\beta_1$  está dado por

$$t_c = \left( \frac{\hat{\beta}_1 - \hat{\beta}_{10}}{s} \right) \sqrt{SC_{xx}} \text{ que tiene una distribución t-Student con } n - 2 \text{ grados de libertad.}$$

La prueba se puede realizar de la siguiente forma:

a) Establecer el contraste de hipótesis a probar.

$$\begin{array}{lll} \text{i) } & \begin{array}{l} H_0: \beta_1 = \hat{\beta}_{10} \\ H_1: \beta_1 \neq \hat{\beta}_{10} \end{array} & \text{ii) } \begin{array}{l} H_0: \beta_1 \geq \hat{\beta}_{10} \\ H_1: \beta_1 < \hat{\beta}_{10} \end{array} & \text{iii) } \begin{array}{l} H_0: \beta_1 \leq \hat{\beta}_{10} \\ H_1: \beta_1 > \hat{\beta}_{10} \end{array} \end{array}$$

b) Fijar el nivel de significancia  $\alpha$ .

c) En el estadístico de prueba  $t_c = \left( \frac{\hat{\beta}_1 - \hat{\beta}_{10}}{s} \right) \sqrt{SC_{xx}}$ ,  $\hat{\beta}_1 =$  coeficien-

te de regresión muestral,  $\hat{\beta}_{10} =$  coeficiente de regresión poblacional

hipotética,  $s = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$  error estándar de la estimación y

$SC_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 =$  suma de los cuadrados de las diferencias entre

cada  $x$  observada y la media de  $x$ .

### Regla de decisión

Para a), rechazar  $H_0: \beta_1 = \hat{\beta}_{10}$ , si CC:  $t_c < t_{tablas}(\alpha/2, n-2)$  o  $t_c > t_{tablas}(1-\alpha/2, n-2)$ .

Para b), rechazar  $H_0: \beta_1 \geq \hat{\beta}_{10}$ , si CC:  $t_c < t_{tablas}(\alpha, n-2)$ .

Para c), rechazar  $H_0: \beta_1 \leq \hat{\beta}_{10}$ , si CC:  $t_c > t_{tablas}(1-\alpha, n-2)$ .

d) Aplicar la regla de decisión.

Un contraste de hipótesis que se pide probar con frecuencia es  $H_0: \beta_1 = 0$  contra  $H_1: \beta_1 \neq 0$ ; si de la realización concluimos que no hay evidencias para rechazar  $H_0: \beta_1 = 0$ , estamos validando que el modelo tiene pendiente 0, luego un cambio en  $x$  no afecta a  $y$ . Es decir, el modelo de regresión lineal  $y = \beta_0 + \beta_1 x + \varepsilon$  no es adecuado porque variaciones en  $x$  no afectarían a la variable dependiente  $y$ . En caso contrario, se dice que la variable  $x$  sí describe a la variable  $y$ ; no obstante, se debe ver en qué proporción la hace.

## Ejemplo 6.8

En el ejemplo 6.2 se trata el problema de la empresa acerca del volumen semanal de ventas en miles de dólares:

1. Realice el contraste de hipótesis  $H_0: \beta_1 = 0$  contra  $H_1: \beta_1 \neq 0$ , con  $\alpha = 0.01$ .
2. Encuentre un intervalo de 99% de confianza para  $\beta_1$ .

### Solución

Utilizaremos los resultados de los teoremas 6.4 y 6.5.

1. Antes de aplicar la metodología, debemos consultar y considerar los valores calculados en la tabla 6.14 del ejemplo 6.6.

$$SC_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = 8.88, \quad \sum_{i=1}^n e_i^2 = 144.0 \quad \text{y} \quad s = 4.9$$

Entonces:

$$\frac{s}{\sqrt{SC_{xx}}} = \frac{4.9}{\sqrt{8.88}} = 1.64 \text{ (error estándar del coeficiente de regresión lineal).}$$

Ahora, seguimos los pasos de la metodología:

- i) El contraste de hipótesis  $H_0: \beta_1 = 0$  contra  $H_1: \beta_1 \neq 0$ .
- ii) Fijar el nivel de significancia  $\alpha = 0.01$ .

iii) El estadístico de prueba  $t_c = \left( \frac{\hat{\beta}_1 - \hat{\beta}_{10}}{s} \right) \sqrt{SC_{xx}}$ .

### Regla de decisión

Recordemos que  $(n - 2) = 8 - 2 = 6$  grados de libertad.

Rechazar  $H_0: \beta_1 = 0$ , si CC:  $t_c < t_{tablas}(0.01/2, 6) = -3.707$  o  $t_c > t_{tablas}(1 - 0.01/2, 6) = 3.707$

- iv) Aplicar la regla de decisión.

Si se recuerda la ecuación de regresión del ejemplo 6.3,  $\hat{y} = 82.269 + 13.056x$ , donde  $\hat{\beta}_1 = 13.056$ ; luego, el valor del estadístico  $t$  calculado es:

$$t_c = \left( \frac{\hat{\beta}_1 - \hat{\beta}_{10}}{s} \right) \sqrt{SC_{xx}} = \frac{13.056 - 0}{1.64} = 7.96$$

Por último,  $7.96 = t_c > 3.707$ , se rechaza la hipótesis nula  $H_0: \beta_1 = 0$ . Por tanto, se concluye que la recta de regresión poblacional no tiene pendiente 0, así que al 1% de significancia se puede decir que existe una relación lineal  $x$  y  $y$  en la población.

- b) Para el intervalo de confianza del teorema 6.4 y los cálculos realizados en el inciso anterior tenemos:

$$\hat{\beta}_1 = b_1 = 13.056, t_{n-2, \alpha/2} = 3.707 \text{ y } \frac{s}{\sqrt{SC_{xx}}} = 1.64$$

De este modo:

$$\hat{\beta}_1 \pm \frac{s}{\sqrt{SC_{xx}}} t_{n-2, \alpha/2} = 13.056 \pm 1.64(3.707)$$

Por tanto, podemos concluir que  $\beta_1 \in (6.977, 19.135)$  con una confianza de 99%.

## 6.5 Coeficientes de correlación y determinación

En la sección anterior revisamos el parámetro de la pendiente de la recta de regresión y vimos que éste proporciona información bastante útil sobre la asociación lineal que existe entre las variables  $x$  y  $y$ . Además, en la unidad 1 revisamos una medida adimensional para cuantificar la relación lineal entre las variables  $x$  y  $y$ , a la cual le dimos el nombre de coeficiente de correlación. Ahora bien, en esta sección analizamos tanto esta medida como el coeficiente de determinación, con el fin de indicar si existe una tendencia lineal entre las variables  $x$  y  $y$ .

### Coeficiente de correlación lineal

El coeficiente de correlación mide el grado de relación entre las variables, si sus valores satisfacen una ecuación, por ejemplo  $y = 4x + 3$ , se dice que las variables están correlacionadas de manera perfecta. Cuando se trata de dos variables solo se habla de correlación simple.

En general, se desea medir el grado de la relación entre  $x$  y  $y$ , así como observarla en un diagrama de dispersión. La medida que se usa para este propósito es el coeficiente de correlación, el cual es un valor numérico entre  $-1$  y  $1$  que mide la fuerza de la relación lineal entre dos variables cuantitativas.

Los coeficientes de correlación existen tanto para una población de valores como para cada muestra que se extrae de dicha población. El símbolo para el coeficiente de correlación de una población es  $\rho$ , para la muestra, por su parte el coeficiente de correlación se representa por la letra  $r$ ; tanto  $\rho$  como  $r$  adquieren valores entre  $-1$  y  $+1$ . Además, en casos particulares tenemos:

- $-1$  indica que existe una muy fuerte relación lineal entre las variables (negativa perfecta).
- $0$  indica que no hay relación lineal.
- $+1$  indica que existe muy fuerte relación lineal entre las variables (positiva perfecta).

Estos valores rara vez aparecen en situaciones reales, aunque constituyen un punto de referencia para evaluar el coeficiente de correlación de cualquier conjunto de datos.

El coeficiente de correlación de la muestra es un estimador puntual del coeficiente de correlación de la población, excepto que el tamaño de la muestra,  $n$ , sustituye al tamaño de la población,  $N$ , y se calcula por medio de (6.10):

$$r = \frac{SC_{xy}}{\sqrt{SC_{xx}SC_{yy}}} = \frac{\text{cov}(x, y)}{S_n(x)S_n(y)} = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \sqrt{n \sum_{i=1}^n y_i^2 - \left(\sum_{i=1}^n y_i\right)^2}} \quad (6.10)$$

donde  $S_n(x)$  y  $S_n(y)$  son las desviaciones estándar de la varianza sesgada. La fórmula anterior solo se usa para datos cuantitativos y se conoce como **coeficiente de correlación de Pearson**.

### Ejemplo 6.9

Usando los datos del ejemplo 6.4 que relaciona el material de desperdicio ( $x$ ) en miles de pesos y las ventas ( $y$ ) de la línea complementaria con el uso de los desperdicios, en millones de pesos, se obtuvieron los resultados de la tabla 6.4:

$$\sum_{i=1}^{12} x_i = 68.7; \quad \sum_{i=1}^{12} x_i^2 = 402.91; \quad \sum_{i=1}^{12} y_i = 316; \quad \sum_{i=1}^{12} y_i^2 = 8708; \quad \sum_{i=1}^{12} x_i y_i = 1865.7$$

Con estos resultados calcule el coeficiente de correlación entre el desperdicio y las ventas.

#### Solución

Con el uso de los valores de la tabla 6.4, y sustituyéndolos en la ecuación de correlación  $r$ , se tiene:

$$\begin{aligned} r &= \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \sqrt{n \sum_{i=1}^n y_i^2 - \left(\sum_{i=1}^n y_i\right)^2}} \\ &= \frac{12(1865.7) - 68.7(316)}{\sqrt{12(402.91) - 68.7^2} \sqrt{12(8708) - 316^2}} \\ &= \frac{679.2}{731.21} = 0.9289 \end{aligned}$$

El coeficiente de correlación para la muestra de 12 datos puntuales es  $r = 0.9289$ , lo cual indica una relación lineal positiva bastante fuerte entre el material de desperdicio y las ventas (el coeficiente de correlación verifica lo observado en el diagrama de dispersión de la figura 6.4).

El coeficiente de correlación muestra una medida para la dependencia lineal entre dos variables, pero no al modelo de la regresión lineal. Dicho en otras palabras, el coeficiente de correlación no implica causalidad. Es decir, con el coeficiente de correlación podemos concluir si existe una tendencia lineal entre  $x$  y  $y$ , pero no es posible concluir que un cambio en  $x$  causa un cambio en  $y$ .

### Ejemplo 6.10

Con los datos de la tabla 6.15, calcule la ecuación de la línea recta que mejor ajuste y el coeficiente de correlación lineal. Trace un diagrama de dispersión de los datos y concluya sobre los resultados obtenidos.

**Tabla 6.15** Observaciones de un experimento para el ejemplo 6.10

$x$	$y$
1.5	12.8
2.5	31.8
3.5	34.8
5.0	56.0
7.0	57.0
9.0	60.0
11.0	50.0
12.0	46.0
13.0	31.0
15.0	16.0

### Solución

Primero, calculamos las siguientes medidas descriptivas de las observaciones:

$$\begin{aligned}\bar{x} &= 7.9500 & \bar{y} &= 39.5250 \\ S_n^2(x) &= 20.2725 & S_n^2(y) &= 257.39313 \\ SC_{xx} &= 202.725 & SC_{yy} &= 2573.9313\end{aligned}$$

$$\text{cov}(x, y) = 4.18875$$

Luego, del teorema 6.2 tenemos los estimadores de la recta de regresión:

$$\begin{aligned}\hat{\beta}_1 &= \frac{\text{cov}(x, y)}{S_n^2(x)} = \frac{4.18875}{20.2725} = 0.2066 \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} = 39.525 - 0.2066(7.95) = 37.882\end{aligned}$$

La recta de mínimos cuadrados está dada por:

$$\hat{y} = 37.882 + 0.2066x$$

Ahora bien, el coeficiente de correlación lineal lo calculamos de (6.10):

$$r = \frac{\text{cov}(x, y)}{S_n(x)S_n(y)} = \frac{4.18875}{\sqrt{20.2725 \times 257.39313}} = 0.0580$$

El diagrama de dispersión de los datos se puede apreciar en la figura 6.5.

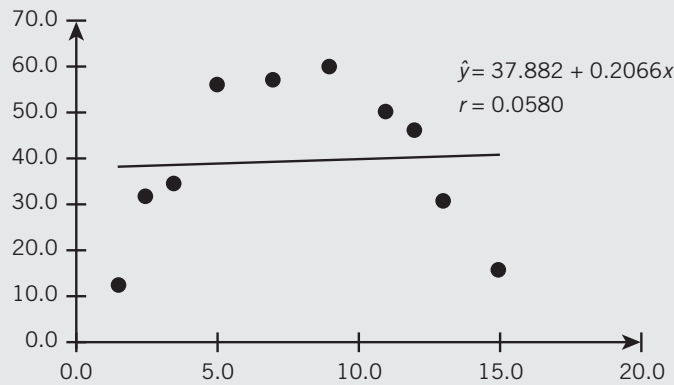


Figura 6.5 Diagrama de dispersión de los datos  $(x, y)$  con su mejor ajuste lineal.

### Conclusión

Con los resultados anteriores podemos apreciar que no existe dependencia lineal entre las variables  $x$  y  $y$ .

Pero esto no significa que las variables no estén relacionadas, de hecho en la figura 6.5 se aprecia una fuerte relación entre ambas, aunque no es lineal.

### Teorema 6.6

En las condiciones del teorema 6.3 se cumple:

$$r = \hat{\beta}_1 \sqrt{\frac{SC_{xx}}{SC_{yy}}}$$

En el análisis de correlación lineal, la pregunta a responder está relacionada con cuándo y bajo qué condiciones se puede concluir, con base en la evidencia muestral, si existe una relación lineal entre las dos variables continuas  $(x, y)$  de la población.

### Teorema 6.7

Sean las parejas de variables  $(x, y)$  que tiene una distribución conjunta normal bivariada (ambas marginales tienen distribución normal); entonces, el estadístico de prueba para los contrastes de hipótesis de  $\rho$  es  $t_c = \frac{r - \rho_0}{S_r}$  y tiene distribución t-Student con  $n - 2$  grados de libertad, donde  $S_r = \sqrt{\frac{1 - r^2}{n - 2}}$  error estándar del coeficiente de correlación,  $\rho_0 =$  coeficiente de correlación poblacional hipotético y  $r =$  coeficiente de correlación muestral.

Las pruebas de hipótesis se pueden llevar a cabo bajo la siguiente metodología.

a) Establecer el contraste de hipótesis a probar.

$$a) \begin{cases} H_0: \rho = \rho_0; \\ H_1: \rho \neq \rho_0 \end{cases} \quad b) \begin{cases} H_0: \rho \geq \rho_0; \\ H_1: \rho < \rho_0 \end{cases} \quad c) \begin{cases} H_0: \rho \leq \rho_0 \\ H_1: \rho > \rho_0 \end{cases}$$

b) Fijar el nivel de significancia  $\alpha$ .

c) Calcular el estadístico de prueba  $t_c = \frac{r - \rho_0}{S_r}$ .

**Regla de decisión:**

Para a), rechazar  $H_0: \rho = \rho_0$ , si CC:  $t_c < t_{tablas}(\alpha/2, n-2)$  o  $t_c > t_{tablas}(1 - \alpha/2, n-2)$ .

Para b), rechazar  $H_0: \rho \geq \rho_0$ , si CC:  $t_c < t_{tablas}(\alpha, n-2)$ .

Para c), rechazar  $H_0: \rho \leq \rho_0$ , si CC:  $t_c > t_{tablas}(1 - \alpha, n-2)$ .

d) Aplicar la regla de decisión.

### Ejemplo 6.11

Estudios recientes realizados por una universidad examinan el papel que juegan el diseño y la cultura organizacional en los distintos niveles de éxito que experimentan las tecnologías avanzadas de manufactura. Durante éstos se examinó la relación entre el perfil de valores competitivos de una organización y el número de características organizacionales. En los estudios se realizaron investigaciones en 332 colegios y universidades. Una de las muchas hipótesis probadas fue que un sistema de valores jerárquicos, que refleja los valores y normas asociadas con la burocracia, se correlaciona con la característica organizacional de formalización. La correlación entre el valor jerárquico numérico y el valor de formalización para la organización de la muestra fue  $r = 0.42$ . Pruebe al nivel de significancia  $\alpha = 0.05$  y el contraste de hipótesis:

$$H_0: \rho \leq 0.3$$

$$H_1: \rho > 0.3$$

**Solución**

Los datos son:

$$r = 0.42$$

$$\alpha = 0.05$$

$$\rho_0 = 0.3$$

$$n = 332$$

Y calculamos el error estándar:

$$S_r = \sqrt{\frac{1 - r^2}{n - 2}} = \sqrt{\frac{1 - 0.42^2}{332 - 2}} = 0.04996$$

Estamos ante una situación como la del inciso c) de la metodología propuesta para las pruebas de hipótesis del coeficiente de correlación lineal. Si se siguen los pasos anteriores para la prueba de hipótesis.

i)  $H_0: \rho \leq 0.3$  contra  $H_1: \rho > 0.3$ .

ii) Nivel de significancia  $\alpha = 0.05$ .

iii) Estadístico de prueba  $t_c = \frac{r - \rho_0}{S_r}$ , para la región de rechazo requerimos calcular:

$$t_{tablas}(1 - \alpha, n - 2) = t_{tablas}(0.95, 330) \approx 1.649$$

**Regla de decisión:**

rechazamos  $H_0$  para valores de  $t_c > 1.649$

iv) El valor del estadístico estará dado por:

$$t_c = \frac{r - \rho_0}{S_r} = \frac{0.42 - 0.3}{0.04996} = 2.40$$

Por último, con base en este valor, concluimos que la hipótesis nula,  $H_0: \rho \leq 0.3$  se rechaza al nivel de significancia 0.05.

Una correlación entre dos variables no significa necesariamente que una variable causa u ocasiona a la otra. Es decir, el análisis de correlación no se puede usar en forma directa para determinar la causalidad. Por otro lado, dos variables que se correlacionan en el sentido estadístico no lo hacen en forma directa de manera significativa. Por ejemplo, se podría determinar que en el sentido estadístico, la asistencia a la iglesia y el consumo de alcohol tiene una correlación alta en ciertas ciudades grandes. Sin embargo, puede que no sea posible determinar cuál es la variable dependiente y cuál la independiente.

**Ejemplo 6.12**

En el ejemplo 6.1, donde se aborda el tema de la empresa con el volumen semanal de ventas en miles de dólares, se quiere saber si existe una correlación con el número de anuncios de televisión para publicidad, cuyos valores se encuentran en la tabla 6.2 del ejemplo 6.3, en la cual:

$$\sum_{i=1}^{12} x_i = 33; \quad \sum_{i=1}^{12} x_i^2 = 145; \quad \sum_{i=1}^{12} y_i = 1089; \quad \sum_{i=1}^{12} y_i^2 = 149897; \quad \sum_{i=1}^{12} x_i y_i = 4608$$

¿Se puede decir que la correlación en la población es menor a 0.98 con base en una muestra tan pequeña? Utilice un nivel de significancia de 0.05.

**Solución**

Para contestar esta pregunta, primero necesitamos probar el contraste de hipótesis:

$$H_0: \rho \geq 0.98$$

$$H_1: \rho < 0.98$$

Para esto, realizamos una prueba de hipótesis con un nivel de significancia de 0.05. Pero, primero necesitamos calcular el coeficiente de correlación muestral:

$$\begin{aligned} r &= \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \sqrt{n \sum_{i=1}^n y_i^2 - \left(\sum_{i=1}^n y_i\right)^2}} \\ &= \frac{8(4608) - 33(1089)}{\sqrt{8(145) - 33^2} \sqrt{8(149897) - 1089^2}} \\ &= 0.956. \end{aligned}$$

Luego, el error estándar de  $r$  es:

$$S_r = \sqrt{\frac{1 - r^2}{n - 2}} = \sqrt{\frac{1 - 0.956^2}{8 - 2}} = 0.1198$$

Ya tenemos los datos  $r = 0.956$ ,  $\alpha = 0.05$ ,  $\rho_0 = 0.98$ ,  $n = 8$  y  $S_r = 0.1198$ ; ahora bien, debemos seguir los pasos anteriores para la prueba de hipótesis.

i)  $H_0: \rho \geq 0.98$  contra  $H_1: \rho < 0.98$ .

ii) Nivel de significancia  $\alpha = 0.05$ .

iii) Estadístico de prueba  $t_c = \frac{r - \rho_0}{S_r}$ . Para la región de rechazo requerimos  $t_{tablas}(\alpha, n - 2) = t_{tablas}(0.05, 6) \approx -1.943$ .

**Regla de decisión:**

rechazamos  $H_0$  para valores  $t_c < -1.943$ .

iv) El valor del estadístico estará dado por:

$$t_c = \frac{r - \rho_0}{S_r} = \frac{0.956 - 0.98}{0.1198} = -0.20 > -1.943$$

Por último, con base en este valor concluimos al 5% de significancia  $H_0: \rho \geq 0.98$ ; no se rechaza. Es decir, existe una fuerte evidencia sobre la correlación lineal entre variables.

**Conclusión:** se tiene una relación directa entre el número de anuncios de televisión para publicidad y el volumen semanal de ventas.

## Coeficiente de determinación

Otra forma de medir la predicción de  $y$ , mediante la ecuación de regresión  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ , se obtiene al calcular la reducción de la proporción entre  $SCE$  y  $SC_{yy}$ , mediante  $1 - \frac{SCE}{SC_{yy}}$ . A este estadístico lo llamamos **coeficiente de determinación** y lo denotamos por  $r^2 = 1 - \frac{SCE}{SC_{yy}}$ . No es una coincidencia que se use la misma letra para el coeficiente de determinación simple ( $r^2$ ) que para el de correlación ( $r$ ), ya que el de determinación simple  $r^2$  es igual al cuadrado del coeficiente de correlación  $r$ . Por ejemplo, si el de correlación entre dos variables es  $r = 0.80$ , entonces  $r^2 = 0.64$  o 64%.

La fórmula para calcular el coeficiente de determinación simple mediante los residuales está dada en (6.11):

$$r^2 = 1 - \frac{SCE}{SC_{yy}} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (6.11)$$

donde:

$y$  = valores muestrales de  $y$ .

$\hat{y}$  = promedio de los valores de  $y$ .

$\hat{y}$  = valores estimados de  $y$  con la ecuación de regresión.

El cociente  $\frac{SCE}{SC_{yy}}$  representa el porcentaje de la variabilidad de  $y$  que todavía no se puede explicar en la ecuación de regresión. Como este cociente es el porcentaje no explicado por  $x$  en la ecuación de regresión, entonces 1 menos este valor constituye el porcentaje explicado. Entonces, el coeficiente de determinación simple  $r^2$  mide el porcentaje de variabilidad en  $y$ , que se explica cuando se usa  $x$  para predecir  $y$ .

### Ejemplo 6.13

En el ejemplo 6.1, en el que se trata el caso de la empresa con volumen semanal de ventas en miles de dólares, el coeficiente de determinación  $r^2$  lo obtenemos usando los residuales.



**Solución**

Primero, completamos la tabla 6.2 con los residuales, con lo cual obtenemos los resultados de la tabla 6.16:

**Tabla 6.16** Cálculo de residuales del ejemplo 6.11

Observación $i$	Anuncios de publicidad ( $x$ )	Volumen semanal ventas en miles de dólares ( $y$ )	$\hat{y}_i$	$e_i$	$e_i^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
1	3	125	121.4	3.6	12.7	-11.13	123.77
2	5	152	147.5	4.5	19.8	15.88	252.02
3	4	131	134.5	-3.5	12.2	-5.13	26.27
4	4	133	134.5	-1.5	2.2	-3.13	9.77
5	5	142	147.5	-5.5	30.8	5.88	34.52
6	3	116	121.4	-5.4	29.6	-20.13	405.02
7	3	127	121.4	5.6	30.9	-9.13	83.27
8	6	163	160.6	2.4	5.7	26.88	722.27
<b>Sumas</b>	<b>33</b>	<b>1 089</b>			<b>144.0</b>		<b>1 656.9</b>

El promedio estará dado por:

$$\bar{y} = \frac{1089}{8} = 136.13$$

Luego, el coeficiente de determinación resulta:

$$r^2 = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{144}{1656.9} = 0.913$$

Del ejemplo 6.10 obtuvimos  $r = 0.956$ , entonces se cumple que  $r^2 \cong 0.913$ .

Esto indica que 91.3% de la variabilidad se puede explicar con el modelo y solo 8.7% queda sin explicación.

Como el coeficiente de correlación  $r$  es un número entre  $-1$  y  $+1$ , entonces  $r^2$  debe ser un número entre 0 y 1, esto se cumple para todos los porcentajes expresados en forma decimal y se debe a que  $r^2$  se interpreta como un porcentaje de la variabilidad de  $y$  que  $x$  puede explicar;  $r^2$  es uno de los estadísticos que se consulta con más frecuencia en el análisis de regresión porque refleja en forma breve y exacta la habilidad de la variable predictora elegida  $x$ , para explicar la variabilidad de  $y$  en un estudio de regresión simple.

### Ejercicios 6.3

En los siguientes ejercicios:

- Calcule el coeficiente de correlación.
- Calcule el coeficiente de determinación simple.
- Realice la prueba de hipótesis que se pide en cada caso.

1. Ejercicios 6.1.1 y 6.2.1.

$H_0: \rho \leq 0.70$  a un nivel de significancia de 0.05.

2. Ejercicios 6.1.2 y 6.2.2.

- $H_0: \rho \geq 0.80$  a un nivel de significancia de 0.05.
3. Ejercicios 6.1.3 y 6.2.3.  
c)  $H_0: \rho = 0.90$  a un nivel de significancia de 0.10.
  4. Ejercicios 6.1.4 y 6.2.4.  
c)  $H_0: \rho \leq 0.85$  a un nivel de significancia de 0.05.
  5. Ejercicios 6.1.5 y 6.2.5.  
c)  $H_0: \rho \leq 0.70$  a un nivel de significancia de 0.05.
  6. Ejercicios 6.1.6 y 6.2.6.  
c)  $H_0: \rho \geq 0.80$  a un nivel de significancia de 0.05.
  7. Ejercicios 6.1.7 y 6.2.7.  
c)  $H_0: \rho \leq 0.90$  a un nivel de significancia de 0.10.
  8. Ejercicios 6.1.8 y 6.2.8.  
c)  $H_0: \rho \leq 0.85$  a un nivel de significancia de 0.05.

## 6.6 Intervalos de confianza para la predicción y estimación

Cuando el modelo de regresión lineal  $y = \beta_0 + \beta_1 x + \varepsilon$  es representativo del experimento nos sirve para conocer (dentro del rango de valores de la variable independiente,  $x$ , en el que fue construido) los valores de la variable de respuesta,  $y$ , con un cierto error  $\varepsilon$ . Antes, vimos que en general el modelo no lo vamos a conocer, pero sí lo podemos estimar por medio de  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ , donde en realidad para cada valor dado de la variable independiente,  $x_0$ , estimamos un valor de  $E(y)$ .

De igual manera, con el modelo  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$  para cada valor dado de la variable independiente,  $x_0$ , predecimos un valor particular de  $y$ . Es decir, tenemos el mismo valor en ambos casos, de la estimación y predicción para un valor  $x_0$ .

La diferencia entre los dos resultados anteriores no es numérico, ya que resulta el mismo valor, reside en su objetivo. El **valor estimado** representa al valor con el que se estima  $E(y)$ , después de realizar *una gran cantidad* de veces el experimento para un mismo valor de la variable independiente  $x_0$ , se considera el promedio de la variable de respuesta en cada repetición del experimento y este valor es estimado con  $\hat{\beta}_0 + \hat{\beta}_1 x_0$ , mientras que en la otra situación simplemente se elige un valor  $x_0$  para **predecir** un valor particular de  $y$ . Para diferenciar de qué valor hablamos vamos a utilizar subíndices; para la estimación será  $y_e$  y para el valor de predicción  $y_p$ .

Es de suponerse que una de las aplicaciones que tiene mayor relevancia de un modelo de regresión consiste en proporcionar un intervalo de confianza, ya sea para  $y_e$  o  $y_p$ . Es decir, en un problema práctico será de interés conocer tanto un valor puntual de estimación o de predicción, como un intervalo de confianza correspondiente. En este caso, para construir el intervalo de confianza requerimos conocer la distribución de cada una de las variables (véase el teorema 6.8).

### Teorema 6.8

Sean las parejas de observaciones  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ , donde las  $x_i$  representan los valores de la variable independiente  $x$  con valores de respuesta  $y_i$ , con  $y = \beta_0 + \beta_1 x + \varepsilon$ , el modelo de regresión y las variables aleatorias,  $y, \varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ , cumplen con los supuestos de un modelo lineal con,  $\sigma^2$ , la varianza homogénea. Entonces para un valor particular,  $x_0$ , con valor estimado o predicho,  $\hat{y}_0$ , se cumple:

$$y_e \sim N(\hat{y}_0, \sigma_e^2) \text{ en donde } \sigma_e^2 = \sigma^2 \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SC_{xx}} \right) = \frac{\sigma^2}{n} \left( 1 + \frac{(x_0 - \bar{x})^2}{S_n^2(x)} \right)$$

$$y_p \sim N(\hat{y}_0, \sigma_p^2) \text{ en donde } \sigma_p^2 = \sigma^2 \left( 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SC_{xx}} \right) = \frac{\sigma^2}{n} \left( n + 1 + \frac{(x_0 - \bar{x})^2}{S_n^2(x)} \right)$$

Debido a que  $\sigma^2$  rara vez es conocida su valor es estimado con  $s^2 = \frac{SCE}{n-2}$ , entonces:

$$\frac{y_e - \hat{y}_0}{s_e} \sim t_{n-2} \text{ en donde } s_e^2 = s^2 \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SC_{xx}} \right) = \frac{s^2}{n} \left( 1 + \frac{(x_0 - \bar{x})^2}{S_n^2(x)} \right)$$

$$\frac{y_p - \hat{y}_0}{s_p} \sim t_{n-2} \text{ en donde } s_p^2 = s^2 \left( 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SC_{xx}} \right) = \frac{s^2}{n} \left( n + 1 + \frac{(x_0 - \bar{x})^2}{S_n^2(x)} \right)$$

Con el teorema 6.8 y los intervalos de confianza estudiados en la unidad 3, resulta muy sencillo determinar la fórmula para el intervalo de confianza de un valor estimado o predicho para el valor particular de  $x = x_0$ .

### Teorema 6.9

En las condiciones del teorema 6.8, para un valor particular  $x = x_0$  un intervalo al  $(1 - \alpha)$  100% de confianza para el valor estimado o predicho  $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$  es:

a) Cuando se desconoce  $\sigma^2$  (caso más usual):

$$y_e \in \left( \hat{y}_0 - s_e t_{\alpha/2, n-2}, \hat{y}_0 + s_e t_{\alpha/2, n-2} \right),$$

$$y_p \in \left( \hat{y}_0 - s_p t_{\alpha/2, n-2}, \hat{y}_0 + s_p t_{\alpha/2, n-2} \right)$$

donde  $s_e^2$ ,  $s_p^2$  y  $s^2$  se calculan como se indica en el teorema 6.8 y  $t_{\alpha/2, n-2}$  es un valor de la distribución t-Student con  $n - 2$  grados de libertad cuya área derecha es igual a  $\alpha/2$ .

b) Cuando se conoce  $\sigma^2$ :

$$y_e \in \left( \hat{y}_0 - \sigma_e Z_{\alpha/2}, \hat{y}_0 + \sigma_e Z_{\alpha/2} \right)$$

$$y_p \in \left( \hat{y}_0 - \sigma_p Z_{\alpha/2}, \hat{y}_0 + \sigma_p Z_{\alpha/2} \right)$$

donde,  $\sigma_e^2$ ,  $\sigma_p^2$  se calculan como se indica en el teorema 6.8 y  $Z_{\alpha/2}$  es un valor de la distribución normal estándar cuya área derecha es igual  $\alpha/2$ .

### Ejemplo 6.14

En el caso del analista de una empresa del ejemplo 6.4 en el que se determina una relación entre el material de desperdicio ( $x$ ) en miles de pesos y las ventas ( $y$ ) de la línea complementaria con el uso de los desperdicios en millones de pesos, los datos se muestran en la tabla 6.4. Determine un intervalo de confianza a 95% para la estimación y predicción del valor correspondiente a:

- Un desperdicio de \$4900.
- Un desperdicio de \$6100.

#### Solución

En el ejemplo 6.4 obtuvimos la ecuación de regresión lineal  $\hat{y} = -7.415 + 5.894x$ ; así, con base en la ecuación de predicción, podemos obtener la tabla de residuales.

Tabla 6.17 Cálculo de residuales del ejemplo 6.14

Observación $i$	Material de desperdicio en miles ( $x$ )	Ventas de línea complementaria en millones de pesos ( $y$ )	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$	$(x_i - \bar{x})^2$
1	5.3	21	23.823	7.970	0.181
2	6.5	28	30.896	8.387	0.601
3	4.5	20	19.108	0.796	1.501
4	4.7	22	20.287	2.935	1.051
5	5.5	28	25.002	8.988	0.051
6	6.8	32	32.664	0.441	1.156
7	7.2	35	35.022	0.000	2.176
8	6.0	30	27.949	4.207	0.076
9	6.8	35	32.664	5.456	1.156
10	5.1	24	22.644	1.838	0.391
11	4.6	17	19.697	7.276	1.266
12	5.7	24	26.181	4.756	0.001
<b>Sumas</b>	<b>68.7</b>	<b>316</b>	<b>315.938</b>	<b>53.050</b>	<b>9.603</b>

De la tabla anterior tenemos:

$$\bar{x} = \frac{68.7}{12} = 5.725; SC_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = 9.603; s^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{53.050}{12-2} = 5.305$$

Para el valor de la distribución t-Student, tenemos  $1 - \alpha = 0.95$ ; luego,  $\alpha/2 = 0.025$  con  $n - 2 = 12 - 2 = 10$  grados de libertad. Es decir  $t_{\alpha/2, n-2} = t_{0.025, 10} = 2.228$ .

Al sustituir estos valores en la fórmula del teorema 6.9, el intervalo de estimación está dado por:

$$\left( \hat{y}_0 - 2.228 \sqrt{5.305 \left( \frac{1}{12} + \frac{(x_0 - 5.725)^2}{9.603} \right)}, \hat{y}_0 + 2.228 \sqrt{5.305 \left( \frac{1}{12} + \frac{(x_0 - 5.725)^2}{9.603} \right)} \right)$$

Mientras que el intervalo de predicción estará dado por:

$$\left( \hat{y}_0 - 2.228 \sqrt{5.305 \left( 1 + \frac{1}{12} + \frac{(x_0 - 5.725)^2}{9.603} \right)}, \hat{y}_0 + 2.228 \sqrt{5.305 \left( 1 + \frac{1}{12} + \frac{(x_0 - 5.725)^2}{9.603} \right)} \right)$$

Pero aún falta sustituir los valores para cada inciso de la pregunta.

a) Para el valor  $x_0 = 4.9$ , tenemos  $\hat{y}_0 = -7.415 + 5.894(4.9) = 21.466$ . Entonces, el intervalo a 95% de confianza para la estimación de  $E(y)$  está dado por:

$$\left( 21.466 - 2.228 \sqrt{5.305 \left( \frac{1}{12} + \frac{(4.9 - 5.725)^2}{9.603} \right)}, 21.466 + 2.228 \sqrt{5.305 \left( \frac{1}{12} + \frac{(4.9 - 5.725)^2}{9.603} \right)} \right)$$

Por último, el intervalo de confianza para la estimación de  $E(y)$  correspondiente a  $x_0 = 4.9$  con 95% de confianza está dado por:

$$(19.450, 23.481).$$

El intervalo a 95% de confianza para la predicción de  $y$  está dado por:

$$\left( 21.466 - 2.228 \sqrt{5.305 \left( 1 + \frac{1}{12} + \frac{(4.9 - 5.725)^2}{9.603} \right)}, 21.466 + 2.228 \sqrt{5.305 \left( 1 + \frac{1}{12} + \frac{(4.9 - 5.725)^2}{9.603} \right)} \right)$$

Por último, el intervalo de confianza para la predicción de  $y$  correspondiente a  $x_0 = 4.9$  con 95% de confianza está dado por:

$$(15.952, 26.979).$$

b) Para el valor  $x_0 = 6.1$ , tenemos  $\hat{y}_0 = -7.415 + 5.894(6.1) = 28.538$ . Entonces, el intervalo de confianza a 95% para la estimación de  $E(y)$  está dado por:

$$\left( 28.538 - 2.228 \sqrt{5.305 \left( \frac{1}{12} + \frac{(6.1 - 5.725)^2}{9.603} \right)}, 28.538 + 2.228 \sqrt{5.305 \left( \frac{1}{12} + \frac{(6.1 - 5.725)^2}{9.603} \right)} \right)$$

Por último, el intervalo de confianza para la estimación de  $E(y)$  correspondiente a  $x_0 = 6.1$  con 95% de confianza está dado por:

$$(26.932, 30.135).$$

El intervalo a 95% de confianza para la predicción de  $y$  está dado por:

$$\left( 28.538 - 2.228 \sqrt{5.305 \left( 1 + \frac{1}{12} + \frac{(6.1 - 5.725)^2}{9.603} \right)}, 28.538 + 2.228 \sqrt{5.305 \left( 1 + \frac{1}{12} + \frac{(6.1 - 5.725)^2}{9.603} \right)} \right)$$

Al final, el intervalo de confianza para la predicción de  $y$  correspondiente a  $x_0 = 6.1$  con 95% de confianza está dado por:

$$(23.161, 33.916)$$

Del ejemplo anterior podemos apreciar que, en efecto, el rango del intervalo de confianza para la estimación es inferior que el rango del intervalo de confianza para las predicciones. Ahora bien, en el siguiente ejemplo podremos apreciar que ambos intervalos crecen cuando los valores de  $x_0$  se alejan del promedio  $\bar{x}$  y que tienden a ser iguales mientras más se alejan.

### Ejemplo 6.15

Con Excel calcule los intervalos para la estimación y predicción a 95% de confianza para diferentes valores del desperdicio, varíe  $x_0 = 2.6, 2.8, 3.0, 3.2, \dots, 11.0$ .

#### Solución

En el ejemplo 6.14 obtuvimos una representación para los intervalos de confianza donde se debe sustituir el valor de  $x_0$  y calcular  $\hat{y}_0$  mediante la ecuación de estimación de la regresión  $\hat{y}_0 = -7.415 + 5.894x_0$ . Con estos dos valores, y al sustituirlos en la representación correspondiente, tenemos el intervalo para el valor estimado de  $E(y)$  a 95% de confianza que le corresponderá a cada valor de  $x_0$  al sustituir los valores de  $\hat{y}_0 = -7.415 + 5.894x_0$  en:

$$\left( \hat{y}_0 - 2.228 \sqrt{5.305 \left( \frac{1}{12} + \frac{(x_0 - 5.725)^2}{9.603} \right)}, \hat{y}_0 + 2.228 \sqrt{5.305 \left( \frac{1}{12} + \frac{(x_0 - 5.725)^2}{9.603} \right)} \right)$$

El intervalo para el valor de predicción de  $y$  a 95% de confianza que le corresponderá a cada valor de  $x_0$  al sustituir los valores de  $\hat{y}_0 = -7.415 + 5.894x_0$  lo tenemos en:

$$\left( \hat{y}_0 - 2.228 \sqrt{5.305 \left( 1 + \frac{1}{12} + \frac{(x_0 - 5.725)^2}{9.603} \right)}, \hat{y}_0 + 2.228 \sqrt{5.305 \left( 1 + \frac{1}{12} + \frac{(x_0 - 5.725)^2}{9.603} \right)} \right)$$

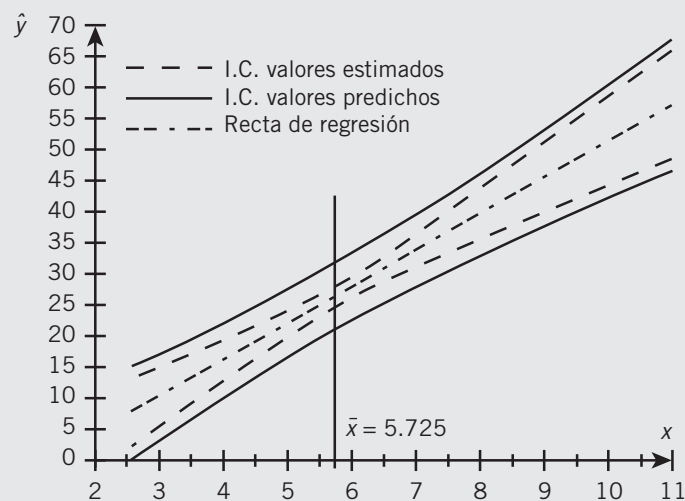
Ahora, se realizan los cálculos para cada valor de  $x_0 = 2.6, 2.8, 3.0, 3.2, \dots, 11.0$  y los resultados se muestran en la tabla 6.18. Después, se traza la gráfica de los límites para cada uno de los dos intervalos de confianza calculados (véase figura 6.6).

**Tabla 6.18** Valores de desperdicio, recta de regresión y límites para cada intervalo de confianza

$x_0$	Valor estimado o predicho	Intervalo de confianza, valor estimado		Intervalo de confianza, valor predicho	
$x_0$	$\hat{y}_0 = -7.415 + 5.894x_0$	$\hat{y}_0 - s_e t_{\alpha/2, n-2}$	$\hat{y}_0 + s_e t_{\alpha/2, n-2}$	$\hat{y}_0 - s_p t_{\alpha/2, n-2}$	$\hat{y}_0 + s_p t_{\alpha/2, n-2}$
2.6	7.909	2.527	13.292	0.472	15.346
2.8	9.088	4.023	14.153	1.878	16.299
3.0	10.267	5.518	15.016	3.275	17.259
3.2	11.446	7.010	15.882	4.663	18.229
3.4	12.625	8.499	16.750	6.040	19.209
3.6	13.803	9.985	17.621	7.407	20.200
3.8	14.982	11.467	18.497	8.762	21.202
4.0	16.161	12.943	19.379	10.104	22.218
4.2	17.340	14.412	20.268	11.432	23.248
4.4	18.519	15.871	21.166	12.744	24.293
4.6	19.697	17.317	22.078	14.041	25.354
4.8	20.876	18.745	23.007	15.320	26.433
5.0	22.055	20.148	23.962	16.581	27.529
5.2	23.234	21.516	24.951	17.822	28.645
5.4	24.413	22.836	25.989	19.044	29.781
5.6	25.591	24.096	27.087	20.246	30.937
5.8	26.770	25.284	28.257	21.428	32.113
6.0	27.949	26.399	29.499	22.588	33.310
6.2	29.128	27.451	30.805	23.729	34.527
6.4	30.307	28.451	32.162	24.850	35.764
6.6	31.485	29.413	33.558	25.951	37.020
6.8	32.664	30.348	34.980	27.034	38.294
7.0	33.843	31.264	36.422	28.100	39.586
7.2	35.022	32.165	37.878	29.149	40.895
7.4	36.201	33.056	39.345	30.182	42.219

**Tabla 6.18** Valores de desperdicio, recta de regresión y límites para cada intervalo de confianza (*Continuación*)

	Valor estimado o predicho	Intervalo de confianza, valor estimado		Intervalo de confianza, valor predicho	
$x_0$	$\hat{y}_0 = -7.415 + 5.894x_0$	$\hat{y}_0 - s_e t_{\alpha/2, n-2}$	$\hat{y}_0 + s_e t_{\alpha/2, n-2}$	$\hat{y}_0 - s_p t_{\alpha/2, n-2}$	$\hat{y}_0 + s_p t_{\alpha/2, n-2}$
7.6	37.379	33.939	40.820	31.201	43.558
7.8	38.558	34.816	42.300	32.207	44.909
8.0	39.737	35.689	43.785	33.201	46.273
8.2	40.916	36.558	45.274	34.183	47.648
8.4	42.095	37.424	46.765	35.156	49.034
8.6	43.273	38.287	48.259	36.118	50.428
8.8	44.452	39.149	49.755	37.073	51.832
9.0	45.631	40.009	51.253	38.019	53.243
9.2	46.810	40.868	52.752	38.958	54.661
9.4	47.989	41.725	54.252	39.891	56.086
9.6	49.167	42.582	55.753	40.818	57.516
9.8	50.346	43.437	57.255	41.740	58.952
10.0	51.525	44.292	58.758	42.657	60.393
10.2	52.704	45.147	60.261	43.569	61.839
10.4	53.883	46.000	61.765	44.477	63.288
10.6	55.061	46.854	63.269	45.382	64.741
10.8	56.240	47.707	64.774	46.282	66.198
11.0	57.419	48.559	66.279	47.180	67.658

**Figura 6.6** Intervalos de confianza para la estimación y predicción de valores  $x_0 = 2.6, 2.8, 3.0, 3.2, \dots, 11.0$ .

En la figura 6.6 se muestran los límites de los intervalos de confianza para las estimaciones y predicciones correspondientes a los valores de  $x_0 = 2.6, 2.8, 3.0, 3.2, \dots, 11.0$ . En ésta se puede apreciar cómo crece el rango del intervalo de confianza mientras más se aleja  $x_0$  del valor de  $\bar{x} = 5.725$ .

**Observaciones del análisis de regresión y correlación**

- En el análisis de regresión, un valor de  $y$  no se puede estimar de manera legítima si el valor de  $x$  está fuera del rango de valores que sirvió como base para la ecuación de regresión.
- Si el cálculo de la estimación de  $y$  involucra la predicción de un resultado, la información histórica que sirvió como base de la ecuación de regresión puede no ser pertinente para eventos futuros.
- El uso de una predicción o un intervalo de confianza se basa en la suposición de que las distribuciones asociadas a  $y$  son normales y tienen varianzas iguales.
- Un coeficiente de correlación significativo no indica necesariamente que exista una estrecha relación entre dos variables, pero sí puede indicar un buen enlace con otros eventos.
- Una correlación *significativa* no es necesariamente una correlación importante.

**Ejercicio 6.4**

Encuentre un intervalo de 95% de confianza para cada una de las predicciones hechas en los ocho ejercicios de la lista de ejercicios 6.1.

**6.7 Regresión lineal múltiple**

Los modelos de regresión lineal simple que acabamos de revisar con dos parámetros se pueden generalizar para  $m$  variables independientes  $x_1, x_2, \dots, x_m$ ; de esta forma, el modelo (6.1) se extiende a (6.12) con  $m + 1$  parámetros:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m + \varepsilon \quad (6.12)$$

El material estudiado para la regresión lineal simple juega un papel muy importante para entender qué es un modelo de regresión lineal. Con los diagramas de dispersión ilustramos algunos tipos de relaciones y el problema de la correlación entre variables. Por tanto, en esta sección vemos una generalización del modelo de regresión lineal simple que inicia con el planteamiento matricial del problema, los supuestos para poder aplicar el modelo, formulación de resultados y algunos ejemplos. También vemos un resumen de los problemas que suelen ocurrir al aplicar un modelo de regresión múltiple. Debido al gran trabajo que se requiere en los cálculos, las operaciones serán realizadas en Excel.

**Planteamiento general del modelo de regresión lineal múltiple**

En el modelo (6.12) se requiere conocer el valor de los  $m + 1$  parámetros, para lo cual se proporcionan  $n$  valores de las variables independientes y se obtienen las ecuaciones (6.13):

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \dots + \beta_m x_{1m} + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \dots + \beta_m x_{2m} + \varepsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \dots + \beta_m x_{nm} + \varepsilon_n \end{aligned} \quad (6.13)$$

Los errores deben cumplir los supuestos de un modelo de regresión lineal simple.

1. La distribución de probabilidad de  $\varepsilon_i \sim N(0, \sigma^2)$ .
2. La media de las distribuciones de probabilidad de  $E(\varepsilon_i) = 0$ . Es decir, el valor medio de  $y$  para un valor dado de cada variable independiente es:

$$E(y) = E(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m + \varepsilon) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m$$

3. La varianza de  $\varepsilon_i$  es constante para todos los valores de  $x$ ,  $E(\varepsilon_i^2) = \sigma^2$ .



4. Los valores de  $\varepsilon$  son independientes entre sí,  $E(\varepsilon_i \varepsilon_j) = 0$  para  $i \neq j$ .

A partir de estos supuestos las variables de respuesta deben ser tales que:

$$y_i \sim N(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_m x_{im}, \sigma^2)$$

La representación matricial de (6.13) se obtiene al introducir las matrices:

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{pmatrix} \quad \text{y} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Entonces, podemos representar el modelo (6.3) en su forma matricial (6.14):

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (6.14)$$

El valor esperado de la matriz de los errores es  $E(\boldsymbol{\varepsilon}) = \mathbf{0}$  y

$$\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^t = \begin{pmatrix} \varepsilon_1^2 & \varepsilon_1 \varepsilon_2 & \cdots & \varepsilon_1 \varepsilon_n \\ \varepsilon_1 \varepsilon_2 & \varepsilon_2^2 & \cdots & \varepsilon_2 \varepsilon_n \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_1 \varepsilon_n & \varepsilon_2 \varepsilon_n & \cdots & \varepsilon_n^2 \end{pmatrix} \Rightarrow E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^t) = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix} = \sigma^2 \mathbf{I},$$

donde  $\mathbf{I}$  es la matriz identidad de orden  $n \times n$ . El modelo de regresión (6.13) se representa con los supuestos de los errores en (6.15):

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad E(\boldsymbol{\varepsilon}) = \mathbf{0} \quad \text{y} \quad E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^t) = \sigma^2 \mathbf{I} \quad (6.15)$$

Para poder estimar los parámetros de (6.15) se requiere tener  $n \geq m + 1$  observaciones diferentes de las variables independientes. En términos de las matrices se necesita que el rango  $(\mathbf{X}) = m + 1$ , es decir, la matriz debe ser de rango completo. Esto equivale a pedir que la matriz simétrica  $\mathbf{X}^t \mathbf{X}$  sea invertible (no singular).

### Teorema 6.10

Sea el modelo de regresión lineal múltiple (6.15), donde  $\mathbf{X}^t \mathbf{X}$  es una matriz no singular, entonces los MELI para  $\boldsymbol{\beta}$  están dados en (6.16):

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{Y} \quad (6.16)$$

La demostración se puede hacer por mínimos cuadrados, al encontrar que el estimador para  $\boldsymbol{\beta}$  está dado por  $\hat{\boldsymbol{\beta}} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{Y}$ ; incluso se puede demostrar que la suma de los cuadrados de los errores  $SCE = \sum_{i=1}^n \varepsilon_i^2 = \boldsymbol{\varepsilon}^t \boldsymbol{\varepsilon} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^t (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$  tiene un mínimo absoluto  $\hat{\boldsymbol{\beta}} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{Y}$ . Después, se prueba que estos estimadores tienen la menor varianza. Durante el desarrollo de la demostración se obtiene que:

$$SCE = \mathbf{Y}^t \mathbf{Y} - \hat{\boldsymbol{\beta}}^t \mathbf{X}^t \mathbf{Y} \quad (6.17)$$

Mientras que el estimador de la varianza,  $\sigma^2$ , de un modelo de regresión lineal múltiple (6.15) está dado en (6.18), donde  $m + 1$  es la cantidad de parámetros del modelo:

$$s^2 = \frac{SCE}{n - (m + 1)} = \frac{\mathbf{Y}^t \mathbf{Y} - \hat{\boldsymbol{\beta}}^t \mathbf{X}^t \mathbf{Y}}{n - (m + 1)} \quad (6.18)$$

## Generalización de resultados de la regresión lineal y prueba F

En esta subsección presentamos los resultados que se muestran para la regresión lineal simple, pero en el caso multivariado. Como vimos en la subsección anterior, estos resultados requieren del análisis multivariado y teoría de matrices, pero esto queda fuera de los objetivos del texto. Por esta razón, solo resumiremos los principales resultados.

Varios de los resultados que resumimos sobre los parámetros del modelo de regresión (6.15) se relacionan con la matriz inversa de  $\mathbf{X}'\mathbf{X}$  (véase 6.19).

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} c_{00} & c_{01} & \cdots & c_{0m} \\ c_{10} & c_{11} & \cdots & c_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2} & \cdots & c_{mm} \end{pmatrix} \quad (6.19)$$

### Teorema 6.11

Cuando los supuestos de un modelo de regresión lineal múltiple (6.15) se cumplen:

- a) La distribución de los MELI para  $i = 0, 1, \dots, m$  se muestra en (6.20) y la distribución del estimador de la varianza en (6.21):

$$\hat{\beta}_i \sim N(\beta_i, c_{ii}\sigma^2) \Rightarrow \left( \frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{c_{ii}}} \right) \sim N(0, 1) \quad (6.20)$$

$$s^2 \sim \frac{\sigma^2}{n - (m + 1)} \chi_{n - (m + 1)}^2 \Rightarrow \left( \frac{\hat{\beta}_i - \beta_i}{s\sqrt{c_{ii}}} \right) \sim t_{n - (m + 1)} \quad (6.21)$$

- b) Los límites del intervalo a  $(1 - \alpha)$  100% de confianza para  $\beta_i$  están dados en (6.22) para  $i = 0, 1, \dots, m$ :

$$\hat{\beta}_i \pm t_{n - (m + 1), \alpha/2} s \sqrt{c_{ii}} \quad (6.22)$$

donde  $t_{n - (m + 1), \alpha/2}$  es el valor de la distribución t-Student con  $n - (m + 1)$  grados de libertad y área derecha igual  $\alpha/2$ .

#### c) Prueba t

Esta prueba consiste en los siguientes pasos:

- i) Establecer el contraste de hipótesis a probar.

$$a) \begin{cases} H_0: \beta_i = \hat{\beta}_{i0} \\ H_1: \beta_i \neq \hat{\beta}_{i0} \end{cases} \quad c) \begin{cases} H_0: \beta_i \geq \hat{\beta}_{i0} \\ H_1: \beta_i < \hat{\beta}_{i0} \end{cases} \quad e) \begin{cases} H_0: \beta_i \leq \hat{\beta}_{i0} \\ H_1: \beta_i > \hat{\beta}_{i0} \end{cases}$$

En estas pruebas  $t_c = \left( \frac{\hat{\beta}_i - \hat{\beta}_{i0}}{s\sqrt{c_{ii}}} \right)$  es el estadístico de prueba que tiene una distribución t-Student con  $n - (m + 1)$  grados de libertad.

- ii) Fijar el nivel de significación  $\alpha$ .

iii) En el estadístico de prueba  $t_c = \left( \frac{\hat{\beta}_i - \hat{\beta}_{i0}}{s\sqrt{c_{ii}}} \right)$ ,  $\hat{\beta}_i$  = coeficiente de regresión muestral,  $\hat{\beta}_{i0}$  = coeficiente de regresión poblacional hipotética,  $s$  error estándar del estimador  $\sigma$ , está dado en (6.18).

**Regla de decisión**

a) Rechazar  $H_0 : \beta_i = \hat{\beta}_{i0}$ , si CC:

$$t_c < t_{tablas}(\alpha/2, n - (m + 1)) \text{ o } t_c > t_{tablas}(1 - \alpha/2, n - (m + 1))$$

Con el valor  $p$ , rechazar  $H_0$  al nivel de significancia  $\alpha$  cuando  $\alpha > p$ .

b) Rechazar  $H_0 : \beta_i \geq \hat{\beta}_{i0}$ , si CC:  $t_c < t_{tablas}(\alpha, n - (m + 1))$  o con el valor  $p$  rechazar  $H_0$  al nivel de significancia  $\alpha$  cuando  $\alpha > p$ .

c) Rechazar  $H_0 : \beta_i \leq \hat{\beta}_{i0}$ , si CC:  $t_c > t_{tablas}(1 - \alpha, n - (m + 1))$  o con el valor  $p$  rechazar  $H_0$  al nivel de significancia  $\alpha$  cuando  $\alpha > p$ .

iv) Aplicar la regla de decisión.

Nota: Estas pruebas de hipótesis se conocen como pruebas  $t$ .

d) La distribución del estadístico de la combinación lineal de los estimadores de los parámetros:

$$\mathbf{a}^t \hat{\boldsymbol{\beta}} = a_0 \hat{\beta}_0 + a_1 \hat{\beta}_1 + \dots + a_m \hat{\beta}_m \sim N(\mathbf{a}^t \boldsymbol{\beta}, \mathbf{a}^t (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{a} \sigma^2),$$

donde  $\mathbf{a}^t = (a_0, a_1, \dots, a_m)$  vector de constantes en la combinación lineal.

e) Los límites del intervalo a  $(1 - \alpha)$  100% de confianza para una estimación de  $E(y)$  en los valores particulares de las variables independientes  $x_{01}, x_{02}, \dots, x_{0m}$  están dados en (6.23):

$$\mathbf{X}_0^t \hat{\boldsymbol{\beta}} \pm t_{n-(m+1), \alpha/2} s \sqrt{\mathbf{X}_0^t (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}_0} \quad (6.23)$$

donde  $t_{n-(m+1), \alpha/2}$  es el valor de la distribución t-Student con  $n - (m + 1)$  grados de libertad y área derecha igual a  $\alpha/2$  y  $\mathbf{X}_0^t = (1, x_{01}, x_{02}, \dots, x_{0m})$ .

f) Los límites del intervalo a  $(1 - \alpha)$  100% de confianza para una predicción  $y$  en los valores particulares de las variables independientes  $x_{01}, x_{02}, \dots, x_{0m}$  están dados en (6.24):

$$\mathbf{X}_0^t \hat{\boldsymbol{\beta}} \pm t_{n-(m+1), \alpha/2} s \sqrt{1 + \mathbf{X}_0^t (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}_0} \quad (6.24)$$

donde  $t_{n-(m+1), \alpha/2}$  es el valor de la distribución t-Student con  $n - (m + 1)$  grados de libertad y área derecha igual a  $\alpha/2$  y  $\mathbf{X}_0^t = (1, x_{01}, x_{02}, \dots, x_{0m})$ .

Con este resumen de resultados para el caso multivariado casi hemos terminado de conocer los conceptos necesarios para comprender los modelos de regresión lineal múltiple. Solo falta tratar cómo decidir si un modelo es adecuado o no.

**Coeficiente de determinación ajustado**

En el caso lineal para medir cuánto explicaba la variable  $x$  a la variable  $y$ , se podía utilizar el coeficiente de determinación (6.11), pero en los modelos de regresión lineal múltiple el uso del coeficiente de determinación tiene el problema de que su valor aumenta conforme se agregan variables al modelo, aunque éstas no contribuyan con información a la predicción de  $y$ . Entonces, el uso de este coeficiente puede no ser tan determinante en los modelos de regresión múltiple. Para evitar, en parte, este problema, se puede utilizar el coeficiente de determinación múltiple ajustado,  $r_a^2$ , definido en (6.25):

$$r_a^2 = 1 - \frac{n-1}{n-(m+1)} \left( \frac{SCE}{SC_{yy}} \right) = 1 - \frac{(n-1)(1-r^2)}{n-(m+1)} \quad (6.25)$$

Antes de empezar a tomar decisiones sobre lo idóneo de un modelo de regresión múltiple mediante alguno de los dos coeficientes, primero debemos tener en cuenta que tanto  $r^2$  como  $r_a^2$  son estadísticos de una muestra y que un investigador no puede depender de estos valores para decidir si el modelo que obtuvo es útil o no para predecir  $y$ .

## Prueba $F$ , análisis de varianza

En una situación general de un modelo de regresión, la forma más objetiva para un decisor sobre la validez de un modelo de regresión que explique los valores de  $y$  consiste en utilizar la prueba  $F$ .

La prueba  $F$  para un modelo de regresión es una prueba de hipótesis en la cual el contraste de hipótesis está dado en (6.26):

$$\begin{cases} H_0: \beta_1 = \beta_2 = \dots = \beta_m = 0 \\ H_1: \text{Por lo menos un } \beta_i \neq 0 \end{cases} \quad (6.26)$$

Por otro lado, el estadístico de prueba para (6.26) está dado por (6.27):

$$F_c = \left( \frac{r^2}{1-r^2} \right) \left( \frac{n-(m+1)}{m} \right) = \left( \frac{\hat{\beta}^t \mathbf{X}^t \mathbf{Y} - n\bar{y}^2}{\mathbf{Y}^t \mathbf{Y} - \hat{\beta}^t \mathbf{X}^t \mathbf{Y}} \right) \left( \frac{n-(m+1)}{m} \right) \quad (6.27)$$

La regla de decisión se muestra en (6.28):

$$\text{Rechazar } H_0 \text{ al nivel de significancia } \alpha \text{ cuando } F_c > F_{t,\alpha}(m, n-(m+1)), \quad (6.28)$$

donde  $F_{t,\alpha}(m, n-(m+1))$  es el valor de la distribución  $F$  de Snedecor con  $m$  grados de libertad en el numerador y  $n-(m+1)$  grados de libertad del denominador.

En los paquetes estadísticos a esta prueba se le conoce como análisis de varianza, ANOVA, denotada también como ANAVA o ANDEVA. Su representación de salida en los paquetes se muestra en la tabla 6.19.

**Tabla 6.19** ANOVA para el modelo de regresión

<i>FV</i>	<i>GL</i>	<i>SC</i>	<i>CM</i>	$F_c$
Modelo	$m$	$SCM = \hat{\beta}^t \mathbf{X}^t \mathbf{Y} - n\bar{y}^2$	$CMM = \frac{SCM}{m}$	$\frac{CMM}{CME}$
Error	$n - (m + 1)$	$SCE = \mathbf{Y}^t \mathbf{Y} - \hat{\beta}^t \mathbf{X}^t \mathbf{Y}$	$CME = \frac{SCE}{n - (m + 1)} = s^2$	
Total	$n - 1$	$SCT = \mathbf{Y}^t \mathbf{Y} - n\bar{y}^2$		

Donde:

FV: fuentes de variación.

GL: grados de libertad.

SC: suma de cuadrados.

CM: cuadrados medios.

SCM: suma de cuadrados del modelo.

SCT: suma de cuadrados totales.

CMM: suma de cuadrados medios del modelo.

CME: suma de cuadrados medios de los errores.

En la tabla 6.19 falta una columna que en general presenta los paquetes en su salida, la cual está destinada al valor de tablas  $F_{t,\alpha}(m, n-(m+1))$  para tomar una decisión (6.28) si se rechaza o no la hipótesis nula o el valor  $p$ ; en este caso, la regla de decisión se muestra en (6.29):

$$\text{Rechazar } H_0 \text{ al nivel de significancia } \alpha, \text{ cuando } \alpha > p. \quad (6.29)$$

## Uso de Excel para la regresión lineal múltiple

En la subsección anterior estudiamos los resultados teóricos que se requieren para las regresiones lineales aplicables, en general, para regresiones simples o múltiples. Como se ha podido constatar en esta unidad, los modelos de regresión son muy importantes, pero requieren de cálculos bastante laboriosos, por lo que vamos a introducir la hoja de cálculo de Excel de Microsoft para realizar los cálculos de las regresiones lineales múltiples.

A continuación se muestran los pasos para utilizar Excel para los modelos de regresión. La versión que explicamos es Office 2013.

1. En la pestaña de **DATOS**, en la parte superior derecha aparece la opción Herramientas de análisis de datos (véase figura 6.7).



Figura 6.7

### Nota


Si esta opción no aparece, primero se tiene que activar, al hacer clic con el ratón en la pestaña superior  de la barra de herramientas de acceso rápido (véase figura 6.8).



Figura 6.8

Aparecen diferentes opciones, entre éstas la opción: Más comandos... (véase figura 6.9).

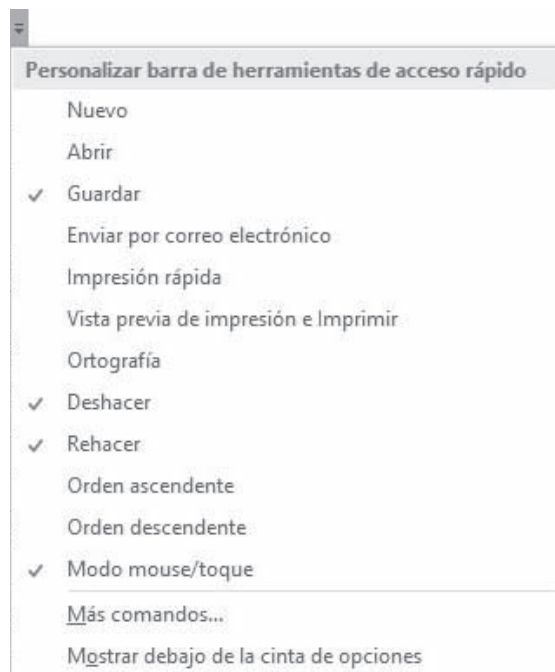


Figura 6.9

Seleccionar esta opción; luego, en la ventana que aparece en la parte izquierda, elegir la opción Complementos. Al hacerlo, aparece una nueva ventana con la siguiente opción en la parte de abajo (véase figura 6.10).

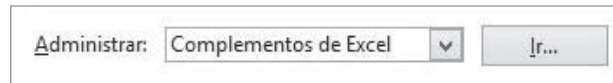


Figura 6.10

Presionar enter en el cuadro **Ir...** y aparecerá el cuadro de diálogo Complementos, en el cual se debe seleccionar la opción **Herramientas para análisis** (véase figura 6.11).

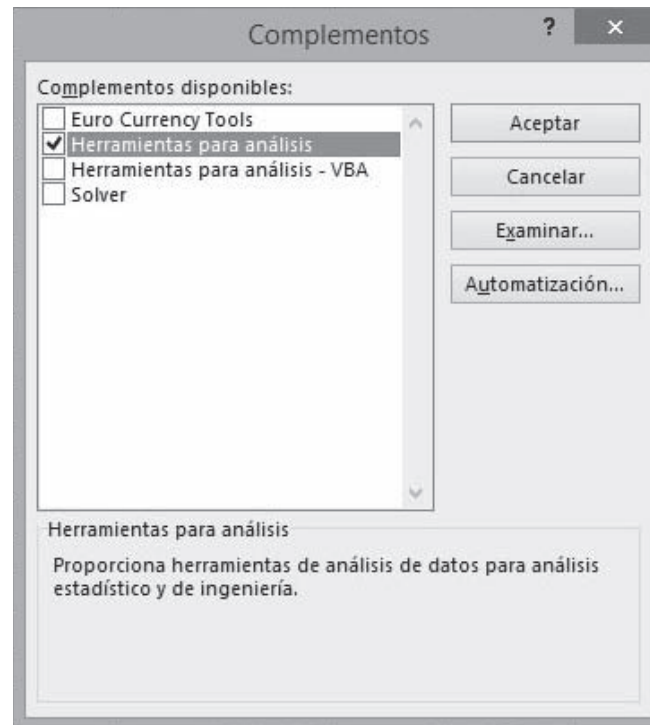


Figura 6.11

Después de elegir **Aceptar**, volver a la barra de acceso rápido e iniciar con el paso 1.

Al presionar enter en esta opción, aparece el cuadro de diálogo en el que se busca la opción Regresión (véase figura 6.12).

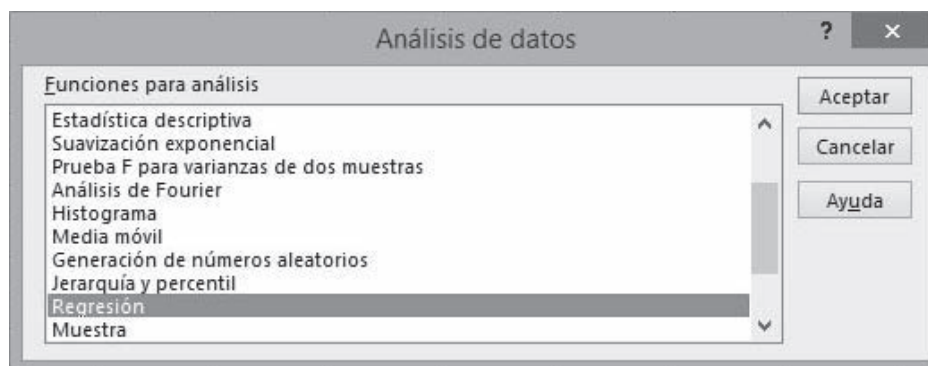


Figura 6.12

2. Después de presionar enter aparece otro cuadro de diálogo, del cual se eligen las opciones:

- **Rango  $Y$  de entrada.** Se seleccionan los valores de la variable dependiente  $y$ , capturados con anticipación en una columna.
- **Rango  $X$  de entrada.** Se seleccionan los valores de las variables independientes proporcionados en la muestra, los valores de cada variable independiente han sido capturados con anterioridad en columnas independientes.
- **Rótulos.** Cuando la columna de datos tiene un título y se desea que los resultados del parámetro correspondiente queden etiquetados con ese nombre, se debe seleccionar esta opción.
- **Nivel de confianza.** Se selecciona el nivel de confianza para los intervalos de confianza de los parámetros (véase la fórmula 6.22). De manera predeterminada, siempre proporciona intervalos con 95% de confianza, pero se puede elegir cualquier otro nivel de confianza.
- **Constante igual a cero.** Esta opción se elige cuando se desea que el estimador del parámetro  $\beta_0$  valga cero (véase figura 6.13).

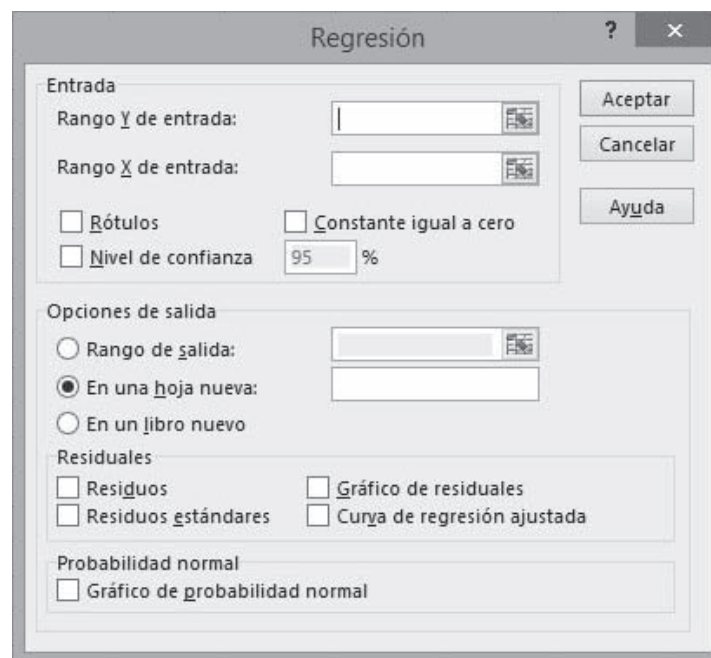


Figura 6.13

### Opciones de salida

- **Rango de salida.** Se elige al menos una celda de la misma hoja en la cual se quiere que el programa muestre los resultados.
- **En una hoja nueva.** Cuando los resultados de salida que se muestran en pantalla deben aparecer en una hoja nueva.
- **En un libro nuevo.** Cuando los resultados de salida que se muestran en pantalla deben aparecer en otro libro.

### Residuales

- **Residuos.** Se activa cuando se desea que en la salida de resultados aparezcan los residuos,  $e_i = y_i - \hat{y}_i$  para cada valor de la muestra.
- **Residuos estándares.** Se activa cuando se desea que en la salida de resultados aparezcan los residuos,  $\frac{y_i - \hat{y}_i}{S_{n-1}(e)}$  para cada valor de la muestra, donde  $S_{n-1}(e)$  representa la desviación estándar de los residuos.

- **Gráfica de residuales.** Muestra en la salida la gráfica de los residuales de cada una de las variables independientes. Con esta gráfica podemos comprobar de manera empírica si existe independencia en los errores.
  - **Curva de regresión ajustada.** Muestra en la salida la gráfica de la regresión ajustada para cada una de las variables independientes. Con esta gráfica podemos comprobar de manera empírica si existe independencia en los errores.
  - **Gráfico de probabilidad normal.** Muestra en la salida la gráfica de los valores estimados de la variable de respuesta con los cuantiles normales. Su interpretación es similar a la gráfica Q-Q vista en la unidad 5. Con esta gráfica podemos comprobar si existe normalidad en la variable de respuesta.
3. Cuando se termina la selección de todas las opciones del menú Regresión que se desea conocer se presiona **Aceptar**, la salida muestra todos los resultados pedidos de un modelo de regresión, incluso la prueba de ANOVA (véase tabla 6.19) para decidir si el modelo es adecuado o no.

## Solución de un modelo de regresión lineal múltiple

En esta subsección vemos un ejemplo para ilustrar los resultados de los modelos de la regresión lineal múltiple. La revisión del ejemplo es un ejercicio completo de análisis de datos por medio de un modelo de regresión lineal. Los cálculos se realizan con ayuda de Excel.

### Ejemplo 6.16

En un hospital se lleva a cabo un estudio de la relación que existe entre la *satisfacción* del paciente ( $y$ ), su *edad* ( $x_1$ , en años), la *gravedad de su enfermedad* ( $x_2$ , un índice) y el *nivel de ansiedad* ( $x_3$ , un índice). El investigador que realizó el estudio seleccionó a 23 pacientes en forma aleatoria y reunió los datos que se muestran en la tabla 6.20, en los cuales los valores grandes de  $y$ ,  $x_2$  y  $x_3$  denotan mayor satisfacción, mayor gravedad y más ansiedad, respectivamente.

**Tabla 6.20** Datos de la satisfacción de los pacientes del hospital

Paciente	$y$	$x_1$	$x_2$	$x_3$
1	58	50	51	2.3
2	57	48	46	2.5
3	66	60	48	2.7
4	70	61	44	2.9
5	79	68	43	3.2
6	53	49	54	2.5
7	66	52	50	2.7
8	57	45	48	2.4
9	62	51	62	2.9
10	67	52	50	2.6
11	65	57	48	2.5
12	67	58	53	2.7
13	47	38	55	1.9
14	59	49	51	2.8

Continúa 



**Tabla 6.20** Datos de la satisfacción de los pacientes del hospital (*Continuación*)

Paciente	$y$	$x_1$	$x_2$	$x_3$
15	67	53	54	2.6
16	66	73	49	2.9
17	69	59	56	3.1
18	78	68	46	3.5
19	60	52	49	2.4
20	59	55	51	2.3
21	69	71	52	2.9
22	64	48	58	2.8
23	60	43	50	2.3

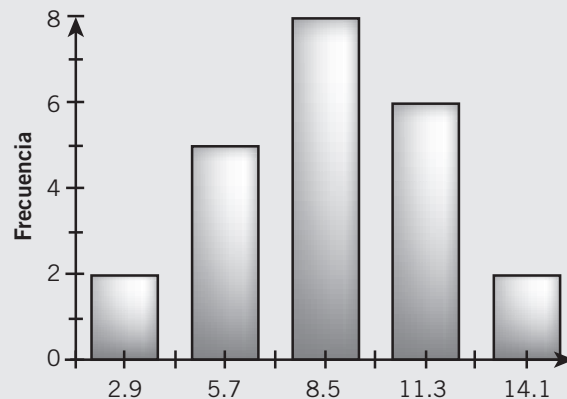
- a) Trace un histograma con cinco clases de frecuencia de igual longitud con los valores de la variable de respuesta. Concluya sobre la posible normalidad de las observaciones.

### Solución

En este caso, construimos las clases de frecuencia de los valores de la satisfacción con base en el método propuesto en la unidad 1. Primero, calculamos los valores máximo y mínimo de las observaciones para el rango y la longitud de clases, que es el cociente del rango entre la cantidad de clases. Con estos valores construimos las clases de frecuencia (véase tabla 6.21) y trazamos su histograma (véase figura 6.14).

**Tabla 6.21** Resultados de las clases de frecuencia

	Clase	liminf	limsup	Frecuencia
Máx = 79	1	47	53.4	2
Mín = 47	2	53.4	59.8	5
Clases = 5	3	59.8	66.2	8
Longitud = 6.4	4	66.2	72.6	6
	5	72.6	79	2

**Figura 6.14** Histograma de las clases de frecuencia de los valores de satisfacción.

### Conclusión

El histograma de las clases de frecuencia de los valores de satisfacción (variable de respuesta del estudio), muestra cierta normalidad en los datos. Este resultado será corroborado con la gráfica de normalidad en el inciso j).

b) Calcule con ayuda de Excel la matriz de correlación. Interpretela y comente los principales resultados.

### Solución

La matriz de correlaciones es una matriz simétrica que se calcula con las variables independientes y muestra la correlación lineal entre cada par de variables independientes. Para el caso general del modelo (6.14), la matriz de correlaciones se muestra en (6.30):

$$\Sigma_r = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1m} \\ r_{12} & 1 & \cdots & r_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1m} & r_{2m} & \cdots & 1 \end{pmatrix} \quad (6.30)$$

donde  $r_{ij}$  es el coeficiente  $a$  de correlación lineal entre las variables independientes  $i$  con  $j$  que se puede calcular con (6.10) para las variables.

El cálculo con Excel para la matriz de correlación se realiza de forma similar como se explicó para el modelo de regresión sobre el uso de Excel, pero en el último cuadro de diálogo, **Análisis de datos**, se utiliza la opción Coeficiente de correlación (véase figura 6.15).

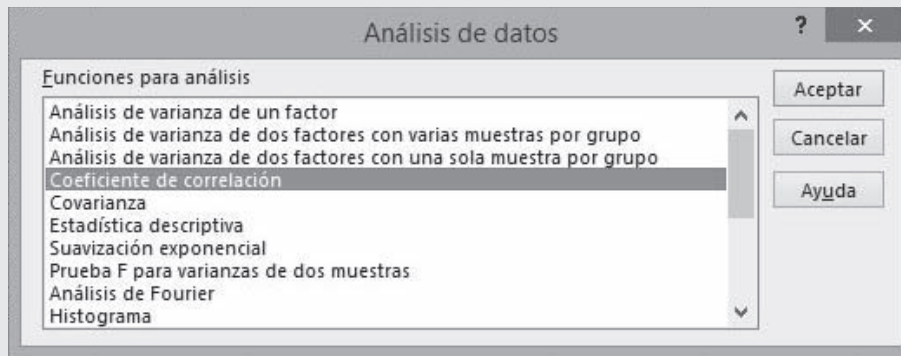


Figura 6.15

Después de realizar los cálculos se obtiene la matriz de la tabla 6.22.

Tabla 6.22 Matriz de correlación de las tres variables independientes

	$x_1$	$x_2$	$x_3$
$x_1$	1	-0.3674	-0.7460
$x_2$	-0.3674	1	0.1579
$x_3$	-0.7460	0.1579	1

### Conclusión

En la tabla 6.22 se observa una correlación inversa lineal fuerte entre las variables edad ( $x_1$ ) y nivel de ansiedad ( $x_3$ ), mientras que en las otras dos parejas de variables su corrección lineal es baja. De hecho, entre nivel de ansiedad y gravedad de la enfermedad ( $x_2$ ) la correlación muestra cierto grado de independencia lineal.

- c) Ajuste el modelo de regresión lineal múltiple a las tres variables independientes. ¿Cuáles son los valores de los tres coeficientes de correlación y cómo se interpretan?

### Solución

Si seguimos los pasos de la subsección sobre el uso de Excel para los cálculos del modelo de mejor ajuste con las tres variables independientes, obtenemos la tabla 6.23, con el resumen de resultados para los tres coeficientes, correlación, determinación y determinación ajustado que se calculan con (6.10), (6.11) y (6.25), respectivamente.

**Tabla 6.23** Resumen de los coeficientes de correlación

Estadísticos de la regresión	
Coefficiente de correlación múltiple	0.90293443
Coefficiente de determinación $R^2$	0.81529058
$R^2$ ajustado	0.78612593
Error típico	3.37957859
Observaciones	23

### Conclusión

El coeficiente de correlación múltiple es alto y, como se esperaba, mayor al coeficiente de determinación, el cual, a la vez, es mayor al coeficiente de determinación ajustado. En las subsecciones anteriores ya vimos que el mejor indicador para el ajuste es el coeficiente de correlación ajustado, valor considerablemente alto 0.786; entonces, podemos suponer que el modelo de regresión múltiple que se obtenga debe ser adecuado, pero esta aseveración debe ser reforzada con el ANOVA. El error típico mostrado en Excel es el valor  $s$ , el estimador de la desviación estándar.

- d) Formule el contraste de hipótesis para el análisis de varianza del modelo de regresión lineal múltiple y concluya sobre el contraste de hipótesis.

### Solución

El contraste de hipótesis que deseamos probar para el ANOVA del modelo de regresión lineal múltiple está dado en (6.20) por:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \\ H_1 : \text{Por lo menos un } \beta_i \neq 0 \end{cases}$$

Al realizar los cálculos del inciso anterior en Excel, en la salida de respuestas también se proporcionó el ANOVA para este problema, que se muestra en la tabla 6.24. En este caso, los cálculos se hacen con las fórmulas de la tabla 6.19.

**Tabla 6.24** ANOVA del modelo de regresión

FV	GL	SC	CM	F	valor p
Regresión	3	957.860088	319.286696	27.954757	3.5465E-07
Residuos	19	217.009478	11.4215514		
Total	22	1174.86957			

Por tanto, de la última columna de la tabla 6.24 se concluye que a un nivel de significancia  $\alpha > 3.55 \times 10^{-7}$  se tiene que rechazar la hipótesis nula  $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ .

### Conclusión

Para un nivel de significancia  $\alpha > 3.55 \times 10^{-7}$  rechazar la hipótesis; entonces, podemos decir que las variables independientes sí explican el modelo a cualquier nivel de significancia  $\alpha > 3.55 \times 10^{-7}$ .

- e) Proporcione el mejor modelo de regresión lineal que ajusta a las tres variables independientes. ¿Cómo se interpreta cada uno de los  $\hat{\beta}_i$  para  $i = 1, 2, 3$ ?

### Solución

Al realizar los cálculos del inciso c) en Excel de Microsoft, también se proporcionó en la salida de respuestas la prueba  $t$  del problema. Los cálculos para los estimadores se realizan con la fórmula (6.16) y los resultados que proporciona Excel se muestran en la segunda columna de la tabla 6.25.

**Tabla 6.25** Prueba  $t$  del modelo de regresión

FV	Coefficientes	Error típico	Estadístico $t$	Probabilidad	Inferior 95%	Superior 95%	Inferior 99%	Superior 99%
Beta 0	104.0468	19.4461	5.3505	0.0000	63.3457	144.7479	48.4129	159.6808
Beta 1	0.2032	0.1299	1.5638	0.1344	-0.0688	0.4752	-0.1685	0.5750
Beta 2	-0.2840	0.1776	-1.5990	0.1263	-0.6557	0.0877	-0.7921	0.2241
Beta 3	-13.5739	3.1744	-4.2761	0.0004	-20.2180	-6.9298	-22.6557	-4.4922

La columna **Coefficientes** de la prueba  $t$  de la tabla 6.25 muestra los valores de los estimadores  $\hat{\beta}_i$  para  $i = 0, 1, 2, 3$  del modelo de regresión estimado dado por:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 = 104.0468 + 0.2032x_1 - 0.2840x_2 - 13.5739x_3$$

### Conclusión

Los modelos de regresión no son del tipo causa-efecto, entonces la interpretación individual de cada parámetro, en general, no es correcta. En la subsección: Problemas en la regresión lineal múltiple, que se halla más adelante, discutimos con mayor detalle este problema. Por el momento, solo podemos concluir:

- Para  $\hat{\beta}_1 = 0.2032$ , si fuera posible modificar la variable independiente *edad* y mantener constantes los valores de las variables independientes *gravedad de la enfermedad* y *ansiedad*, entonces por cada año de edad de variación en la edad del paciente, su satisfacción variaría de manera proporcional en 0.2032 unidades. Al aumentar la edad del paciente en un año la satisfacción aumentaría en 0.2032 unidades en promedio; de igual manera, al disminuir la edad la satisfacción disminuiría en 0.2032 unidades en promedio; este resultado se justifica con el modelo de regresión dentro del rango de edades de los pacientes de 38 a 79 años (valores: menor y mayor de las edades de la muestra).
- Para  $\hat{\beta}_2 = -0.2840$ , si fuera posible modificar la variable independiente, índice de gravedad *de la enfermedad*, y mantener constantes la *edad* e índice de *ansiedad*, entonces por cada unidad de variación en el índice de la *gravedad de la enfermedad del paciente*, su satisfacción variaría inversamente proporcional en 0.2840 unidades. Por su parte, al aumentar el índice de *gravedad de la enfermedad* en una unidad, la satisfacción disminuiría en 0.2840 unidades en promedio; de la misma forma, al disminuir el índice de *gravedad de la enfermedad* la satisfacción aumentaría en 0.2840 unidades en promedio y este resultado se justifica con el modelo dentro del rango del índice de la enfermedad de los pacientes de 43 a 62 unidades (valores: menor y mayor del índice de gravedad de la enfermedad de la muestra).
- Para  $\hat{\beta}_3 = -13.5739$ , si fuera posible modificar la variable independiente, índice *ansiedad*, y mantener constantes la *edad* e índice de *gravedad de la enfermedad del paciente*, entonces por cada unidad de variación en el índice de *ansiedad del paciente*, su satisfacción variaría de manera inversamente proporcional

en 13.5739 unidades. Por su parte, al aumentar el índice de *ansiedad* en una unidad, la satisfacción disminuiría en 13.5739 unidades en promedio; de igual manera, al reducir el índice de *ansiedad*, la satisfacción aumentaría en 13.5739 unidades en promedio y este resultado se justifica con el modelo dentro del rango del índice de *ansiedad* de los pacientes de 1.9 a 3.5 unidades (valores: menor y mayor del índice *ansiedad* de la muestra).

- f) Realice la prueba  $t$  a 5% de significancia para los contrastes de hipótesis y concluya sobre los resultados.

$$\begin{cases} H_0: \beta_i = 0 \\ H_1: \beta_i \neq 0 \end{cases} \text{ con } i = 1, 2, 3$$

### Solución

La prueba  $t$  se realiza con la metodología propuesta en el inciso c) del teorema 6.11. En la tabla 6.25 de Excel se muestran en la columna **Probabilidad**, los valores de  $p$ , para decidir si se rechaza o no  $H_0: \beta_i = 0$  para  $i = 1, 2, 3$ .

- Se obtuvo  $p_1 = 0.1344 > 0.05 = \alpha$ , se concluye que no hay evidencia para rechazar  $H_0: \beta_1 = 0$ .
- Se obtuvo  $p_2 = 0.1263 > 0.05 = \alpha$ , se concluye que no hay evidencia para rechazar  $H_0: \beta_2 = 0$ .
- Se obtuvo  $p_3 = 0.0004 < 0.05 = \alpha$ , se concluye que se rechaza  $H_0: \beta_3 = 0$ .

### Conclusión

De la prueba  $t$  aplicada a cada uno de los tres parámetros se tiene que las variables independientes *edad* y *gravedad de la enfermedad* a 5% de significancia no influyen en la satisfacción de los pacientes, no así los índices de *ansiedad*.

- g) Calcule un intervalo de 95 y 99% de confianza para los parámetros  $\beta_i$  del modelo con  $i = 1, 2, 3$ .

### Solución

Los intervalos de confianza para los parámetros se calculan con la fórmula (6.22), del inciso b) del teorema 6.11. En las cuatro últimas columnas de la tabla 6.25 de Excel se observan los límites de estos intervalos, los cuales se muestran en la tabla 6.26.

**Tabla 6.26** Intervalos de confianza para los parámetros del modelo de regresión

Parámetro	Inferior 95%	Superior 95%	Inferior 99%	Superior 99%
$\beta_0$	63.3457	144.7479	48.4129	159.6808
$\beta_1$	-0.0688	0.4752	-0.1685	0.5750
$\beta_2$	-0.6557	0.0877	-0.7921	0.2241
$\beta_3$	-20.2180	-6.9298	-22.6557	-4.4922

### Conclusión

De la prueba  $t$  aplicada a cada uno de los tres parámetros, se tiene que las variables independientes *edad* y *gravedad de la enfermedad* a 5% de significancia no influyen en la satisfacción de los pacientes; no así los índices de *ansiedad*.

- h) Calcule un intervalo a 95% de confianza para la estimación de la satisfacción media cuando  $x_{01} = 55$ ,  $x_{02} = 60$  y  $x_{03} = 2.6$ .

En la tabla de la prueba  $t$ , la columna de los errores típicos es igual a  $s\sqrt{c_{ii}}$ , la columna del estadístico  $t$  de la prueba es  $t_c = \frac{\hat{\beta}_i}{s\sqrt{c_{ii}}}$  y la columna de probabilidad es el valor  $p$  de la prueba  $t$ .

**Solución**

Los intervalos de confianza para la estimación están dados en la fórmula (6.23):

$$X_0^t \hat{\beta} \pm t_{n-(m+1), \alpha/2} s \sqrt{X_0^t (X^t X)^{-1} X_0},$$

donde:

$$X_0^t = (1, 55, 60, 2.6)$$

$$\hat{\beta}^t = (104.0468, 0.2032, -0.2840, -13.5739)$$

Para el intervalo a 95% de confianza se obtiene  $1 - \alpha = 0.95$ , si se despeja  $\alpha = 0.05$ , se tiene  $\alpha/2 = 0.025$ . Por tanto,  $t_{n-(m+1), \alpha/2} = t_{19, 0.025} = 2.093$ . En la tabla 6.23 tenemos que el error típico es  $s = 3.3796$ ; aunque también puede obtenerse de la tabla 6.24 del ANOVA, en la columna del CME  $s = \sqrt{11.4216} = 3.3796$ . Al calcular la matriz  $X^t X$  obtenemos:

$$X^t X = \begin{pmatrix} 23.0 & 1260.0 & 1168.0 & 62.8 \\ 1260.0 & 70808.0 & 63663.0 & 3389.10 \\ 1168.0 & 63663.0 & 59748.0 & 3194.50 \\ 62.8 & 3389.1 & 3194.5 & 174.12 \end{pmatrix}$$

Solo falta calcular la matriz inversa, que resulta:

$$(X^t X)^{-1} = \begin{pmatrix} 33.108446276 & -0.1938049520 & -0.2075441860 & -4.3612805020 \\ -0.193804952 & 0.0014783147 & 0.0007670479 & 0.0270529535 \\ -0.207544186 & 0.0007670479 & 0.0027615140 & 0.0092609481 \\ -4.361280502 & 0.0270529535 & 0.0092609481 & 0.8822602354 \end{pmatrix}$$

Si se utiliza  $X_0^t \hat{\beta}$  para el valor estimado de  $E(y)$ , tenemos  $\hat{E}(y) = 62.8907$ . Por otro lado, el intervalo de confianza para el valor estimado se calcula con la fórmula (6.23); así, al sustituir valores tenemos:

$$X_0^t \hat{\beta} - t_{n-(m+1), \alpha/2} s \sqrt{X_0^t (X^t X)^{-1} X_0} = 59.1985$$

$$X_0^t \hat{\beta} + t_{n-(m+1), \alpha/2} s \sqrt{X_0^t (X^t X)^{-1} X_0} = 66.5828$$

**Conclusión**

Es decir,  $E(y) \in (59.1985, 66.5828)$  con una confianza de 95%.

- i) Obtenga un intervalo a 95% de confianza de predicción para la satisfacción de un nuevo paciente si:  $x_{01} = 55$ ,  $x_{02} = 60$  y  $x_{03} = 2.6$ .

**Solución**

Los intervalos de confianza para el valor predicho están dados en la fórmula (6.24):

$$X_0^t \hat{\beta} \pm t_{n-(m+1), \alpha/2} s \sqrt{1 + X_0^t (X^t X)^{-1} X_0}$$

Si utiliza  $X_0^t \hat{\beta}$  para el valor predicho, tenemos  $\hat{y} = 62.8907$ . El intervalo de confianza se calcula con la fórmula (6.24); así, al sustituir valores tenemos:

$$X_0^t \hat{\beta} - t_{n-(m+1), \alpha/2} s \sqrt{1 + X_0^t (X^t X)^{-1} X_0} = 54.9115$$

$$X_0^t \hat{\beta} + t_{n-(m+1), \alpha/2} s \sqrt{1 + X_0^t (X^t X)^{-1} X_0} = 70.8698$$

### Conclusión

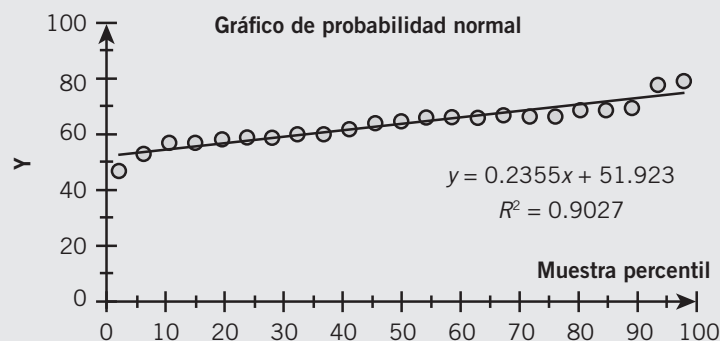
Es decir,  $y \in (54.9115, 70.8698)$  con una confianza de 95%.

j) Trace la gráfica de normalidad.

En la subsección anterior se mostró que la gráfica de normalidad es otra de las opciones que proporciona Excel para hacer un análisis de los modelos de regresión múltiple. Con éstos, mediante la técnica Q-Q, es posible verificar si existe normalidad en la variable de respuesta. Los resultados para este ejemplo se muestran en la tabla 6.27 y en la figura 6.16, donde podemos apreciar que existe una dependencia lineal entre los valores de la variable de respuesta y los cuantiles normales con  $r^2 = 90.27\%$ .

**Tabla 6.27** Cuantiles para normalidad

Percentil normal	Valor de Y	Percentil normal	Valor de Y
2.17	47	54.35	66
6.52	53	58.70	66
10.87	57	63.04	66
15.22	57	67.39	67
19.57	58	71.74	67
23.91	59	76.09	67
28.26	59	80.43	69
32.61	60	84.78	69
36.96	60	89.13	70
41.30	62	93.48	78
45.65	64	97.83	79
50.00	65		



**Figura 6.16** Gráfica de normalidad.

### Conclusión

De la figura 6.16 y el coeficiente de determinación 0.9027 se concluye que la variable de respuesta tiene una distribución normal, si se corrobora el resultado del inciso a).

### k) Conclusiones generales

¿Existe una relación de regresión? Para poder indicarlo de manera objetiva se debe explicar si el modelo es adecuado.

Después de realizar un análisis de regresión para la variable de satisfacción, vemos que se cumple el supuesto de normalidad. El modelo de regresión múltiple obtenido con las variables independientes *edad* y los *índices de gravedad de la enfermedad* y *ansiedad*,  $\hat{y} = 104.0468 + 0.2032x_1 - 0.2840x_2 - 13.5739x_3$ , sí explica a la variable de respuesta. Al parecer, sin embargo, en el modelo se puede reducir la cantidad de variables, ya que tanto en la prueba *t* como en los intervalos de confianza para los parámetros obtuvimos que  $\beta_1$  y  $\beta_2$  no influyen de manera significativa en los cambios de *y*.

Por último, para poder tomar una decisión acerca de qué tan viable es el modelo propuesto, falta analizar los residuales para verificar los supuestos de los errores. Además, tenemos que decidir si la cantidad de variables independientes es la adecuada o se puede reducir el modelo, estos temas los tratamos en las siguientes subsecciones.

## Análisis de residuales en la regresión lineal múltiple

El estudio de los residuales en los modelos de regresión juega un papel muy importante para validar un modelo, lo cual se debe a que los residuos deben satisfacer los supuestos mencionados con anterioridad y bajo los que se construyen los modelos lineales. Para verificar los supuestos de los errores, utilizamos los datos de la muestra, calculamos los residuales y verificamos que cumplan los supuestos siguientes.

### Independencia y valor esperado cero de los errores

El supuesto de que  $E(\varepsilon) = 0$  lo verificamos al calcular el promedio; el valor debe estar cercano a 0.

Para la aleatoriedad e independencia de los errores, su gráfica no debe mostrar comportamientos de dependencia entre los valores de la variable independiente y los residuales. En la figura 6.17 se muestran comportamientos correlacionados de los residuales que no son deseables.

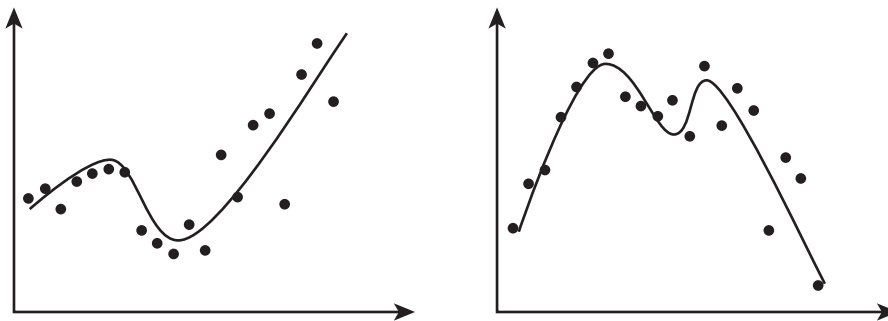
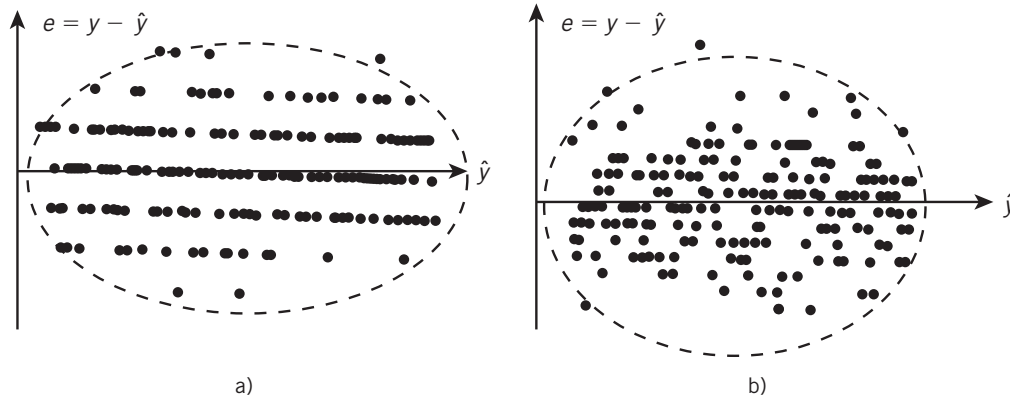


Figura 6.17 Residuales correlacionados.

### Varianza constante de los errores

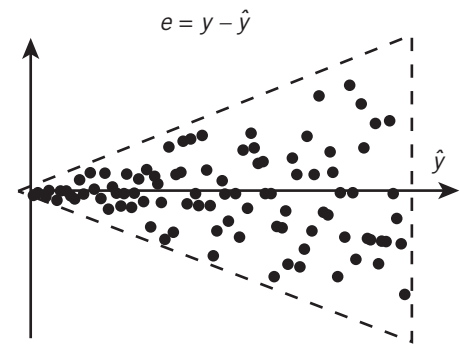
Otro de los supuestos de los errores en un modelo de regresión es que la varianza con respecto a los valores de la variable independiente debe ser constante. Una forma de verificar este supuesto consiste en graficar los residuales ( $e = y - \hat{y}$ ) contra la variable independiente o valores pronosticados ( $\hat{y}$ ) y que su dispersión no aumenta o disminuye de manera considerable conforme se incrementan o disminuyen los valores de la variable independiente. Por ejemplo, en el caso de una variable de respuesta tipo binomial no cumplen con el supuesto de varianzas constantes, sus gráficas son del tipo que se muestra en la figura 6.18.





**Figura 6.18** Gráfica de  $e = y - \hat{y}$  contra  $\hat{y}$  para variables de respuesta binomial. a)  $p = 0.5$  y  $n = 8$ , b)  $p = 0.5$  y  $n = 18$ .

En la figura 6.18 a) se muestran los residuos contra los valores pronosticados para  $p = 0.5$  y  $n = 8$ , en la cual se puede apreciar que no existen variaciones considerables en éstos y que se forma una figura elíptica, en la que al inicio y final ocurren valores próximos a cero y en el centro hay mayor variabilidad, al aumentar el tamaño del parámetro  $n$ . En la figura 6.18 b) se puede ver que los residuos se hacen más homogéneos, pero aún se conservan las características del inciso a). En la figura 6.19 se muestra la gráfica de residuos contra valores pronosticados para una variable de respuesta de Poisson, la figura se asemeja al caso de errores multiplicativos que se trata más adelante. En esta figura se aprecia con claridad que las varianzas crecen conforme lo hace el valor de la variable independiente o el valor pronosticado, lo que ocasiona que los errores no tengan la misma varianza y no se cumpla con este supuesto, entonces el modelo de regresión lineal construido con estas observaciones tampoco sería el adecuado.



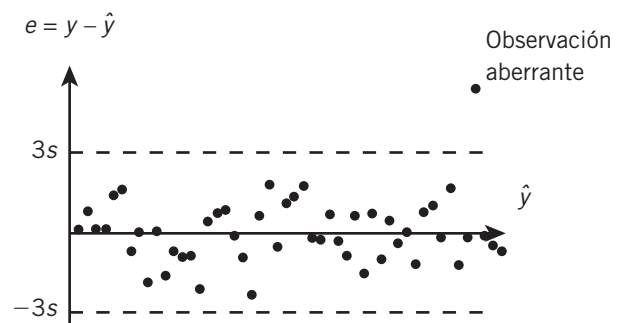
**Figura 6.19** Gráfica de  $e = y - \hat{y}$  contra  $\hat{y}$  para variables de respuesta de Poisson o con errores multiplicativos.

## Observaciones atípicas o aberrantes

Cuando el investigador lleva a cabo un experimento pueden resultar mediciones de la variable de respuesta que difieran de manera considerable del resto de las demás observaciones; en una situación como ésta se dice que la observación queda fuera del rango. Una forma de detectar estas observaciones es mediante los residuos, ya que se considera un rango permisible  $|e| = |y - \hat{y}| < 3s$ ; donde  $s$  es error estándar de estimación para  $\sigma$ , dado en (6.5) o (6.18). Las observaciones que quedan fuera de este rango se llaman **aberrantes** o **atípicas**. Para detectarlas se traza el diagrama de dispersión de los residuos y dos rectas horizontales a las distancias  $\pm 3s$ , todas las observaciones que queden fuera de este intervalo son aberrantes (véase figura 6.20).

Las observaciones aberrantes suelen influir en el modelo; en ocasiones a tal grado que no es posible obtener un modelo de regresión que explique los resultados del experimento. Entonces surge la duda sobre qué hacer con estas observaciones.

Cuando aparecen observaciones aberrantes se recomienda revisar el experimento para detectar si estos valores no se deben a fallas en la captura de las ob-



**Figura 6.20** Residuales de un modelo con observaciones aberrantes.

servaciones. En caso de existir errores deben corregirse o eliminar la observación aberrante y volver a calcular el modelo de regresión.

Cuando la observación aberrante no es causada por errores en el experimento, entonces la respuesta sobre qué hacer con ella es una decisión del investigador. Algunos recomiendan eliminarla y volver a calcular el modelo, pero otros prefieren trabajar con ella ya que argumentan que el valor observado forma parte de los resultados posibles del experimento y eliminarlo alteraría su naturaleza.

En este texto recomendamos:

- Cuando la cantidad de observaciones aberrantes es muy pequeña, podemos calcular el modelo de regresión sin estos valores y verificar si el modelo resultante explica los resultados del experimento.
- En el punto anterior, puede ser que no se eliminen todos los valores aberrantes, sino solo los que queden más alejados del resto.
- Cuando las observaciones aberrantes se repiten en forma periódica, entonces se recomienda descomponer la base de datos en dos, una con las observaciones que quedaron dentro del intervalo  $\pm 3s$  y la otra con las que quedan fuera del intervalo, y ajustar dos modelos, uno para cada base de datos.

### Ejemplo 6.17

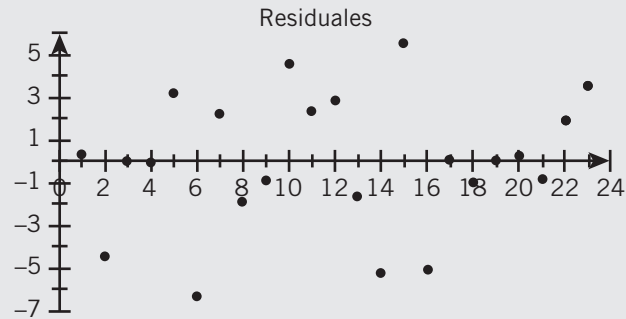
Con los datos del ejemplo 6.16 realice un análisis de los residuales. ¿Hay evidencia de posibles observaciones aberrantes?

#### Solución

Después de realizar los cálculos en Excel, en la tabla 6.28 se muestran los valores de los residuales y residuales estandarizados (definidos al final del punto 3 de la subsección sobre el uso de Excel) y en la figura 6.21, en la cual se aprecia la gráfica de dispersión de los valores de residuales correspondientes y en la que no existen observaciones aberrantes. Al final de la tabla 6.28 se muestran los valores del promedio y varianza muestral de las tres columnas de la tabla y podemos verificar que se cumple que el promedio es casi cero.

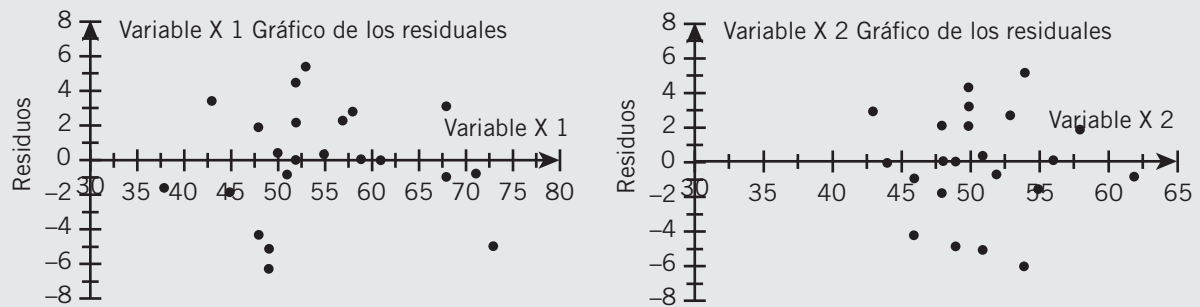
Tabla 6.28 Residuales

Observaciones	Pronóstico para Y	Residuos	Residuos estándares
1	57.6456	0.3544	0.1128
2	61.3738	-4.3738	-1.3926
3	65.9592	0.0408	0.0130
4	70.0130	-0.0130	-0.0041
5	75.7916	3.2084	1.0215
6	59.3053	-6.3053	-2.0076
7	63.7656	2.2344	0.7114
8	58.8389	-1.8389	-0.5855
9	62.8695	-0.8695	-0.2768
10	62.4082	4.5918	1.4620
11	62.6348	2.3652	0.7531
12	64.1329	2.8671	0.9129
13	48.6417	-1.6417	-0.5227
14	64.2294	-5.2294	-1.6650
15	61.4755	5.5245	1.7590
16	71.0317	-5.0317	-1.6021
17	68.9138	0.0862	0.0275
18	79.0119	-1.0119	-0.3222
19	59.9774	0.0226	0.0072
20	58.6616	0.3384	0.1077
21	69.7733	-0.7733	-0.2462
22	62.0384	1.9616	0.6246
23	56.5071	3.4929	1.1121
<b>Promedio</b>	63.6957	0.0000	0.0000
<b>Varianza</b>	43.5391	9.8641	1.0000

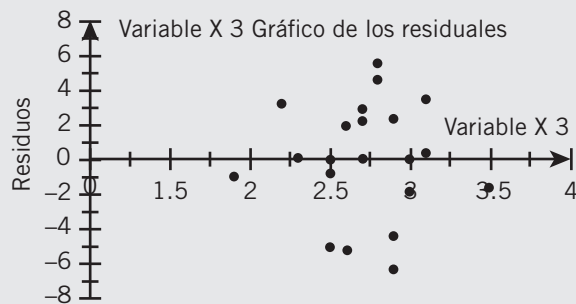


**Figura 6.21** Diagrama de dispersión de los residuales.

En la salida de Excel de la regresión también podemos activar la opción para los gráficos de los residuales contra cada una de las variables independientes; estas gráficas se muestran en las figuras 6.22 y 6.23.



**Figura 6.22** Gráfica de  $e = y - \hat{y}$  contra  $x_1$  y  $x_2$ .



**Figura 6.23** Gráfica de  $e = y - \hat{y}$  contra  $x_3$ .

### Conclusión

A partir de la tabla 6.28 y las figuras 6.21 a 6.23 podemos concluir que los residuales no presentan puntos aberrantes y cumplen con los supuestos de que los errores no están correlacionados,  $E(e) = 0$ , y tienen varianza constante.

## Problemas en la regresión lineal múltiple

Cuando construimos un modelo de regresión lineal, podemos enfrentarnos a diferentes problemas que lo hacen incorrecto o no adecuado para describir el comportamiento de la variable de respuesta. Por esta razón, antes de proponer el modelo debemos revisar que se cumplan los siguientes puntos:

### 1. Estimabilidad de parámetros

Cuando se quiere ajustar un modelo de regresión lineal a un conjunto de observaciones se requiere al menos la misma cantidad de valores diferentes de las variables independientes que la cantidad de parámetros del modelo que se quiere ajustar. Por ejemplo, si buscamos ajustar las observaciones con un modelo de regresión con dos parámetros,  $y = \beta_0 + \beta_1 x + \varepsilon$ , mínimo requerimos al menos dos valores diferentes de la variable independiente  $x$ , para poder determinar la recta de regresión. Para el caso de tres parámetros,  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ , se requieren al menos tres valores diferentes de las variables independientes.

Por otro lado, hay que recordar que la estimación de la varianza requiere de  $n - (m + 1) > 0$  grados de libertad, esto implica que la cantidad de observaciones de la muestra para el modelo de regresión debe ser  $n > m + 1$ ; es decir, se requiere al menos una cantidad mayor a la cantidad de parámetros de valores de la variable independiente. Cuando no se cumple esto se dice que los parámetros no son estimables.

En general, para no tener problemas en los ajustes se recomienda una mayor cantidad de observaciones que de parámetros que además contengan una cantidad mayor a  $m + 1$  de valores diferentes de las variables independientes.

### 2. Extrapolación

Los modelos de regresión que ajustan a las observaciones solo pueden utilizarse para estimar a  $E(y)$  o predecir valores particulares de  $y$ , cuando los valores de las variables independientes están dentro del rango de valores en el que se construyó el modelo de regresión lineal. Nunca deben emplearse valores fuera de estos rangos, ya que los resultados teóricos avalan solo resultados dentro de estos rangos de valores.

Lo anterior se debe a que los valores de la variable de respuesta no siempre siguen el mismo comportamiento para cualquier valor de la variable independiente. Por ejemplo, suponga que se desea medir la conductividad del calor en un metal, a ciertas temperaturas el incremento puede ser lineal, pero al alcanzar una cierta temperatura el metal se deforma y la conductividad cambia su comportamiento, ya que el metal puede comenzar a derretirse.

En conclusión, cuando se quiere hacer estimaciones de  $E(y)$  o predicciones de  $y$  fuera de los rangos de valores de las variables independientes no se trata de un problema de regresión, puede ser que se desea hacer pronósticos y para esto existen otras técnicas diferentes a las regresiones, que no abordaremos en este texto.

### 3. Interpretación de parámetros

Cuando los parámetros del modelo son estimables y se ha obtenido el modelo de regresión, antes de aplicarlo es conveniente saber qué se puede hacer con éste y en qué rango de valores de las variables independientes es válido el resultado, además de interpretar de manera correcta las estimaciones de los parámetros.

Con respecto a la primera parte, podemos decir que los modelos de regresión solo deben utilizarse para estimar a  $E(y)$  o predecir valores particulares de  $y$ , cuando los valores de las variables independientes están dentro del rango de valores en los que se construyó el modelo de regresión. Por otra parte, los valores de la variable de respuesta deben considerarse como resultado de los cambios en todas las variables predictoras, pero no deben contemplarse como los resultados de una relación causa-efecto entre cada variable predictora y la variable de respuesta. Esto último nos dice que no debemos considerar las estimaciones de los  $\hat{\beta}_i$  como el efecto proporcional que ejerce  $x_i$  sobre  $E(y)$ , ya que  $E(y)$  está influenciado por todas las variables independientes y algunas otras que no fueron consideradas en el modelo. Para hacer válida la consideración de proporcionalidad entre  $\hat{\beta}_i$  y  $E(y)$  debemos suponer que podemos cambiar solo un valor de una de las variables, mientras las otras permanecen sin modificaciones, el problema es que esta consideración en la práctica no siempre se puede llevar a cabo, ya que como veremos en el siguiente punto las variables predictoras pueden estar correlacionadas.

Para terminar este punto, debemos tener claro que los modelos de regresión no son causa-efecto, sino que tienen como objetivo establecer si un conjunto de variables es capaz de explicar el comportamiento promedio de una variable de respuesta. Un modelo causa efecto se lleva a cabo mediante un diseño de experimentos, pero este es un tema que queda fuera de los objetivos del texto.

### 4. Multicolinealidad y modelos parsimoniosos

Cuando estudiamos el coeficiente de determinación se explicó que un problema grave de estos modelos de regresión multivariado es que cuando se agregan más variables al modelo el coeficiente aumenta su valor, aun cuando

las nuevas variables no tengan relación con la de respuesta, y que esto se podía disminuir con el coeficiente de determinación ajustado. Además puede ocurrir que alguna de las variables que se agreguen al modelo esté correlacionada con otra u otras variables predictoras.

Lo anterior se debe a que las variables predictoras son independientes en el sentido de que sus valores son asignados por el investigador en los experimentos que realiza, pero entre ellas puede existir una correlación fuerte; es decir, la independencia entre variables predictoras no es en sentido estadístico de variables aleatorias independientes (hay que recordar que las predictoras no son variables aleatorias). Por ejemplo, suponga que el investigador quiere conocer si las variables:

$x_1$ : consumo de litros de gasolina de un carro

$x_2$ : potencia del carro

$x_3$ : índice de aceleración del carro, que toma valores entre 1 y 5, uno para aceleración ideal (sin forzar motor) y 5 para la aceleración con máximo forzamiento del motor

Esta variable influye en la variable de respuesta  $y$ , rendimiento del carro. En este caso, queda claro que las tres variables predictoras propuestas están correlacionadas, entonces el modelo de regresión lineal múltiple que ajuste al problema con las tres variables tendrá parámetros de sobra.

Cuando en un modelo de regresión lineal múltiple aparecen variables predictoras con una fuerte correlación, se dice que existe el **problema de multicolinealidad**.

La multicolinealidad aparece cuando:

- Al menos un par de variables predictoras están correlacionadas fuertemente.
- La prueba ANOVA indica que el modelo con todas las variables es adecuado, pero la prueba  $t$  muestra que algunos parámetros no son significativos en el modelo, no se rechaza la hipótesis nula  $H_0: \beta_i = 0$ .
- Algunos estimadores de los parámetros tienen signo contrario al que se esperaría.

El problema de la multicolinealidad debe corregirse, ya que puede repercutir en que:

- El modelo queda sobreparametrizado, tiene más parámetros de los que se requieren. Esto contradice la teoría de los modelos, en la cual el **principio de parsimonia** establece que un buen modelo tiene la menor cantidad de parámetros y explica de manera adecuada al fenómeno, a este modelo se le llama **modelo parsimonioso**.
- Cuando se estiman los parámetros pueden resultar valores alejados del parámetro o hasta valores contradictorios, lo que ocasiona serios problemas al interpretar los resultados del modelo.

Para detectar la multicolinealidad existen diferentes métodos programados en algunos paquetes estadísticos como SAS (por su nombre en inglés: Statistical Analysis Systems - Sistemas de Análisis Estadístico), que contienen varias funciones para detectar y eliminar la multicolinealidad. La forma más efectiva de encontrar correlaciones entre las variables se basa en la rama de la estadística conocida como análisis multivariado, mediante los llamados componentes principales que muestran el grado de aportación de cada una de las variables predictoras; sin embargo, su estudio queda fuera de los objetivos del texto.

El problema de multicolinealidad es más extenso de lo que podemos exponer aquí. Una forma factible de realizarse, pero que no es muy recomendable llevar a cabo cuando se tiene un número considerable de variables independientes, consiste en buscar las diferentes combinaciones de los modelos que se pueden hacer con las variables predictoras y después elegir el modelo parsimonioso. Esta forma de resolver un problema de multicolinealidad requiere de mucho trabajo cuando la cantidad de pará-

Con el ejemplo de los carros es posible explicar porqué el modelo de regresión no debe considerarse como la representación de un problema causa-efecto, ya que en las condiciones del problema de los carros no sería factible que se altere una sola variable predictora y las demás permanezcan fijas. Esto explica en cierta medida el problema anterior sobre la interpretación de los parámetros, en la cual se mencionó que las estimaciones de los parámetros no deben considerarse como efectos individuales que influyen en forma proporcional en el valor de  $E(y)$ .

metros aumenta, ya que en total se requieren  $2^{m+1} - 1$  combinaciones de modelos en donde  $m + 1$  es la cantidad de parámetros del modelo completo.

En general, en la práctica para detectar el problema de multicolinealidad se recomienda:

- No hacer todas las combinaciones posibles de modelos, pero sí utilizar las correlaciones entre variables predictoras y revisar las pruebas  $F$  y  $t$  para algunas de estas combinaciones donde es más evidente el problema de multicolinealidad.
- Utilizar algún paquete estadístico como SAS, que simplifique el trabajo. Aunque el problema de estos paquetes es su elevado costo.

### Ejemplo 6.18

Con los datos del ejemplo 6.16 y un análisis de los resultados de la tabla 6.22 sobre los coeficientes de correlación realice el ANOVA y la prueba  $t$  para otra combinación de variables; luego, compárelos con los resultados del problema 6.16 y decida si este nuevo modelo es parsimonioso. Calcule los intervalos con una confianza de 95 y 99% para los parámetros. Determine si hay multicolinealidad en estos datos.

#### Solución

En la tabla 6.22 podemos apreciar que existe fuerte correlación lineal entre las variables *edad*  $x_1$  y la variable índice de gravedad de la *enfermedad*  $x_2$ , pero no así entre las variables  $x_2$  y el índice de *ansiedad*  $x_3$ . Entonces, debemos calcular el modelo de regresión de mejor ajuste con estas dos variables.

**Tabla 6.29** Resumen de los coeficientes de correlación para  $x_2$  y  $x_3$

Estadísticos de la regresión	
Coefficiente de correlación múltiple	0.8897
Coefficiente de determinación $R^2$	0.7915
$R^2$ ajustado	0.7707
Error típico	3.4996
Observaciones	23

**Tabla 6.30** ANOVA del modelo de regresión para  $x_2$  y  $x_3$

<i>FV</i>	<i>GL</i>	<i>SC</i>	<i>CM</i>	<i>F</i>	<i>valor p</i>
Regresión	2	929.9275	464.9637	37.9652	1.5515E-07
Residuos	20	244.9421	12.2471		
Total	22	1174.8696			

Por tanto, de la última columna de la tabla 6.30 se concluye que a un nivel de significancia  $\alpha > 1.5515 \times 10^{-7}$  se tiene que rechazar la hipótesis  $H_0: \beta_2 = \beta_3 = 0$ .

#### Conclusión

Para un nivel de significancia  $\alpha > 1.5515 \times 10^{-7}$  rechazar la hipótesis nula, entonces podemos decir que las variables independientes  $x_2$  y  $x_3$  sí explican el modelo, para cualquier nivel de significancia  $\alpha > 1.5515 \times 10^{-7}$ .

Tabla 6.31 Prueba  $t$  del modelo de regresión para  $x_2$  y  $x_3$ 

FV	Coefficientes	Error típico	Estadístico $t$	Probabilidad	Inferior 95%	Superior 95%	Inferior 99%	Superior 99%
Beta 0	130.6870	9.7115	13.4569	1.750E-11	110.4291	150.9449	103.0544	158.3196
Beta 2	-0.3894	0.1701	-2.2888	0.0331	-0.7443	-0.0345	-0.8735	0.0947
Beta 3	-17.2926	2.1776	-7.9411	1.307E-07	-21.8350	-12.7502	-23.4886	-11.0965

La columna *Coefficientes* de la prueba  $t$  de la tabla 6.31 muestra los valores de los estimadores  $\hat{\beta}_2$  y  $\hat{\beta}_3$  del modelo de regresión estimado dado por:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 = 130.6870 - 0.3894x_2 - 17.2926x_3$$

Si se comparan los resultados del modelo completo contra éste de dos variables, podemos concluir que ambos son casi igual de significativos, aunque sería preferible trabajar con el último, ya que es más parsimonioso que el primero. Además de concluir sobre la evidencia fuerte de multicolinealidad entre las variables *edad* e índice de *gravedad de la enfermedad*.

## Regresión curvilínea

En muchos experimentos, la variable de respuesta que se estudia no siempre es lineal. Por ejemplo, la conducción del calor en una barra metálica puede tener un comportamiento lineal en la variable independiente hasta cierta  $T_0$ , y a partir de este valor su comportamiento cambia a uno de tipo polinomial o de otro tipo, pero los parámetros del modelo  $\beta_0, \beta_1, \beta_2, \dots, \beta_m$  aún son lineales y los errores aditivos, como en los modelos:

1.  $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_m x^m + \varepsilon$  con  $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_m x^m$
2.  $y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 x^2 + \beta_4 z^2 + \beta_3 xz + \varepsilon$  con  $E(y) = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 x^2 + \beta_4 z^2 + \beta_3 xz$
3.  $y = \beta_0 + \beta_1 x^{q_1} + \beta_2 x^{q_2} + \dots + \beta_m x^{q_p} + \varepsilon$  con  $E(y) = \beta_0 + \beta_1 x^{q_1} + \beta_2 x^{q_2} + \dots + \beta_m x^{q_p}$
4.  $y = \beta_0 + \beta_1 \log(x_1) + \beta_2 \log(x_2) + \varepsilon$  con  $E(y) = \beta_0 + \beta_1 \log(x_1) + \beta_2 \log(x_2)$
5. Etcétera.

El modelo de regresión para este tipo de experimentos no es lineal, pero al cumplir con la propiedad aditiva en los errores, es posible hacer un cambio de variable para resolverlo como un modelo lineal multivariable. Por ejemplo, en los modelos anteriores podemos hacer los cambios de variables:

1. En  $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_m x^m + \varepsilon$  se puede hacer el cambio de variables:

$$x_1 = x, x_2 = x^2, \dots, x_m = x^m$$

Entonces, el modelo se puede trabajar como el modelo de regresión múltiple:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m + \varepsilon$$

2. En  $y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 x^2 + \beta_4 z^2 + \beta_3 xz + \varepsilon$  se puede hacer el cambio de variables:

$$x_1 = x, x_2 = z, x_3 = x^2, x_4 = z^2, x_5 = xz$$

Entonces, el modelo se puede trabajar como uno de regresión múltiple:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon$$



3. En  $y = \beta_0 + \beta_1 x^{q_1} + \beta_2 x^{q_2} + \dots + \beta_m x^{q_p} + \varepsilon$  se puede hacer el cambio de variables:

$$x_1 = x^{q_1}, x_2 = x^{q_2}, \dots, x_m = x^{q_p}$$

Entonces, el modelo se puede trabajar como uno de regresión múltiple:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m + \varepsilon$$

4. En  $y = \beta_0 + \beta_1 \log(x_1) + \beta_2 \log(x_2) + \varepsilon$  se puede hacer el cambio de variables:

$$z_1 = \log(x_1), z_2 = \log(x_2)$$

Entonces, el modelo se puede trabajar como uno de regresión múltiple:

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \varepsilon$$

En el ejemplo 6.10 se presentó una situación de este tipo, donde probamos que el ajuste lineal no era correcto, pero en la figura 6.5 se observa que sería recomendable un modelo de regresión cuadrático.

### Ejemplo 6.19

Con los datos de la tabla 6.15 construya el modelo de regresión cuadrático que mejor ajuste. Muestre los resultados de las salidas de Excel, junto con sus gráficas e interprete los resultados.

#### Solución

En este caso proponemos un modelo de regresión  $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$ , que tiene errores aditivos, entonces mediante el cambio de variable  $x_1 = x, x_2 = x^2$  lo convertimos en un modelo de regresión múltiple  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ .

Primero, realizamos los cálculos en Excel, para lo cual agregamos una nueva columna de variables independientes que se forma con los cuadrados de la variable independiente original. Los cálculos se muestran en la tabla 6.32.

**Tabla 6.32** Observaciones muestrales para el modelo de regresión múltiple.

$Y$	$x_1 = x$	$x_2 = x^2$
12.75	1.50	2.25
31.75	2.50	6.25
34.75	3.50	12.25
56.00	5.00	25.00
57.00	7.00	49.00
60.00	9.00	81.00
50.00	11.00	121.00
46.00	12.00	144.00
31.00	13.00	169.00
16.00	15.00	225.00

A continuación se muestran las tablas de resultados de Excel.

1. El resumen de resultados se muestra en la tabla 6.33.



**Tabla 6.33** Resumen de los coeficientes de la regresión

Estadísticas de la regresión	
Coefficiente de correlación múltiple	0.9771
Coefficiente de determinación $R^2$	0.9548
$R^2$ ajustado	0.9419
Error típico	4.0763
Observaciones	10

El coeficiente de correlación ajustado es alto, lo que indica fuertes evidencias de que el modelo de regresión múltiple es adecuado, no obstante esto se debe corroborar con el análisis de varianza.

2. El análisis de varianza para la prueba  $F$  se muestra en la tabla 6.34.

**Tabla 6.34** Tabla ANOVA de la prueba  $F$ 

$FV$	$GL$	$SC$	$CM$	$F$	valor $p$
Regresión	2	2457.620	1228.810	73.954	0.00002
Residuos	7	116.311	16.616		
Total	9	2573.931			

Como se puede comprobar, para un nivel de significancia  $\alpha > 0.00002$  se tiene que rechazar la hipótesis nula  $H_0: \beta_1 = 0, \beta_2 = 0$ . Con esto se concluye que las variables independientes sí explican el modelo.

3. El modelo de regresión, la prueba  $t$  y los intervalos de confianza para los parámetros del modelo se muestran en la tabla 6.35.

**Tabla 6.35** Tabla de la prueba  $t$ 

	Coefficientes	Error típico	Estadístico $t$	Probabilidad	Inferior 95%	Superior 95%
Intersección	-6.2716	4.4799	-1.3999	0.2043	-16.8649	4.3216
Beta 1	16.2692	1.3537	12.0183	0.0000	13.0682	19.4702
Beta 2	-1.0008	0.0824	-12.1403	0.0000	-1.1958	-0.8059

El modelo de regresión estimado está dado por:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 = -6.2716 + 16.2692x - 1.0008x^2$$

A 95% de confianza, los parámetros del modelo se encuentran en:

$$\beta_1 \in (13.0682, 19.4702)$$

$$\beta_2 \in (-1.1958, -0.8059)$$

4. Los valores de  $y$  predichos y residuales con residuales estándares se muestran en la tabla 6.36.

Tabla 6.36 Residuales

Observación	Pronóstico $\hat{Y}$	Residuos	Residuos estándares
1	15.8804	-3.1304	-0.8708
2	28.1463	3.6037	1.0024
3	38.4106	-3.6606	-1.0183
4	50.0539	5.9461	1.6540
5	58.5726	-1.5726	-0.4375
6	59.0847	0.9153	0.2546
7	51.5902	-1.5902	-0.4424
8	44.8405	1.1595	0.3225
9	36.0892	-5.0892	-1.4157
10	12.5815	3.4185	0.9509
<b>Promedio =</b>	<b>39.5250</b>	<b>0.0000</b>	<b>0.0000</b>
<b>Varianza =</b>	<b>273.0689</b>	<b>12.9235</b>	<b>1.0000</b>

Los residuales justifican  $E(\varepsilon) = 0$ .

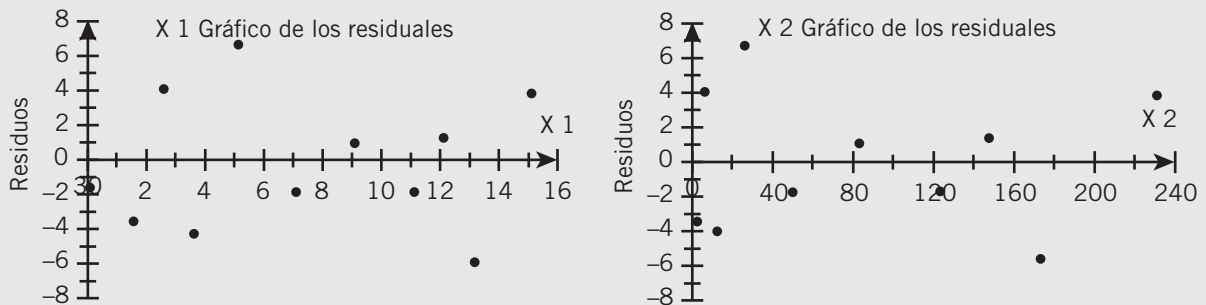


Figura 6.24 Residuales de ambas variables.

Los residuales no muestran tendencias.

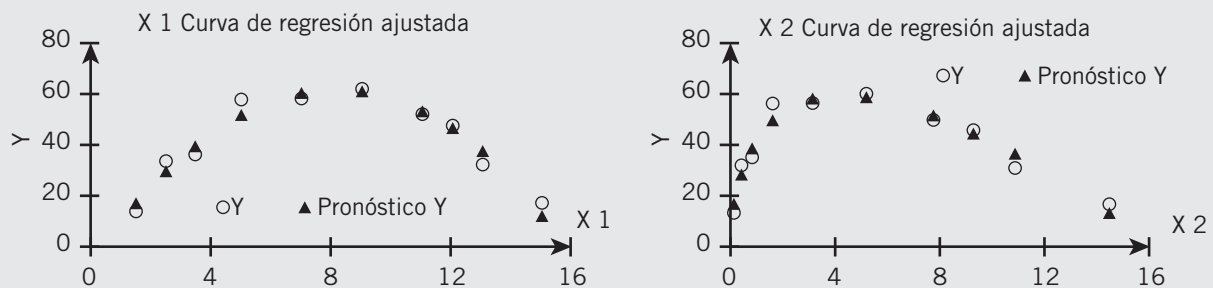


Figura 6.25 Curvas de regresión ajustadas para ambas variables.

Las gráficas muestran el ajuste cuadrático.

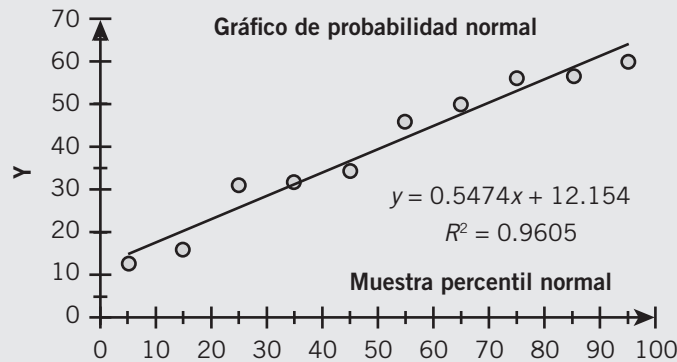


Figura 6.26 Gráfica para la normalidad.

La distribución de la variable de respuesta justifica la normalidad.

### Conclusión

Con los resultados anteriores podemos concluir que los datos iniciales muestran un ajuste cuadrático y se justifican los supuestos del modelo.

## Modelos de regresión con errores multiplicativos

Dentro de los modelos de regresión tenemos una clase muy particular que se denomina modelos de Cobb-Douglas y se caracterizan por tener errores multiplicativos, una clase de éstos se muestra en la fórmula (6.31):

$$z_i = \alpha Q_{i_1}^{\beta_1} Q_{i_2}^{\beta_2} \eta_i, \quad (6.31)$$

donde  $z_i$  es la variable de respuesta,  $\beta_1$  y  $\beta_2$  los parámetros del modelo,  $Q_{i_1}$  y  $Q_{i_2}$  las variables independientes y  $\eta_i$  los errores aleatorios caracterizados por ser multiplicativos y no aditivos como en el modelo de regresión lineal. Los modelos (6.31) se pueden transformar en un modelo de regresión lineal múltiple al extraer el logaritmo. Por ejemplo, si se extrae el logaritmo natural en (6.31) resulta:

$$\ln(z_i) = \ln(\alpha) + \beta_1 \ln(Q_{i_1}) + \beta_2 \ln(Q_{i_2}) + \ln(\eta_i)$$

Al denotar  $y_i = \ln(z_i)$ ,  $\beta_0 = \ln(\alpha)$ ,  $x_{1i} = \ln(Q_{i_1})$ ,  $x_{2i} = \ln(Q_{i_2})$  y  $\varepsilon_i = \ln(\eta_i)$  tendremos el modelo clásico de regresión lineal (6.12) con  $m = 2$  y errores aditivos, de los que se han mostrado los principales resultados; pruebas ANOVA y  $t$ , para estimar al mejor modelo y estimadores  $\hat{\beta}_0$ ,  $\hat{\beta}_1$  y  $\hat{\beta}_2$ . Por último, el mejor modelo para (6.31) se puede obtener con las operaciones inversas  $\hat{\alpha} = e^{\hat{\beta}_0}$ ,  $\hat{z}_i = e^{y_i}$ ,  $\hat{Q}_{i_1} = e^{x_{1i}}$ ,  $\hat{Q}_{i_2} = e^{x_{2i}}$  y  $\hat{\eta}_i = e^{\varepsilon_i}$ .

### Ejemplo 6.20

Un problema clásico del modelo (6.31) fue sugerido por Henry Theil, quien estudió los modelos tipo Cobb-Douglas para optimizar recursos dentro de la administración de las empresas. Para su estudio utilizó bases de datos anuales de 1923 a 1939, a las que ajustó el modelo (6.31), donde  $z$  representa el consumo de textiles,  $Q_1$  el ingreso real (per cápita) y  $Q_2$  precio relativo de textiles. Verifique que los errores son multiplicativos, encuentre un modelo que represente al problema y estime los parámetros. Además, verifique con anterioridad el cumplimiento de los supuestos del modelo de la regresión lineal.

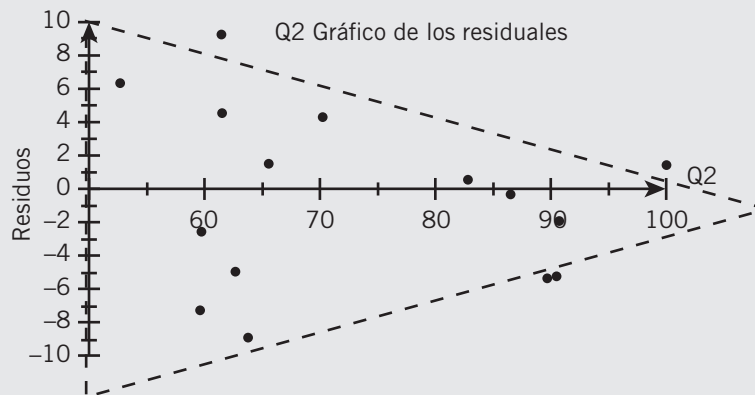
**Tabla 6.37** Datos anuales para el problema de Henry Theil

Año	$z$	$q_1$	$q_2$
1923	99.2	96.7	101.0
1924	99.0	98.1	100.1
1925	100.0	100.0	100.0
1926	111.6	104.9	90.6
1927	122.2	104.9	86.5
1928	117.6	109.5	89.7
1929	121.1	110.8	90.6
1930	136.0	112.3	82.8
1931	154.2	109.3	70.1

Año	$z$	$q_1$	$q_2$
1932	153.6	105.3	65.4
1933	158.5	101.7	61.3
1934	140.6	95.4	62.5
1935	136.2	96.4	63.6
1936	168.0	97.6	52.6
1937	154.3	102.4	59.7
1938	149.0	101.6	59.5
1939	165.5	103.8	61.3

**Solución**

Al emplear el modelo (6.12) de la regresión lineal, para los datos originales, después de hacer la corrida en Excel obtenemos los residuales en la figura 6.27.

**Figura 6.27** Gráfica de residuos para la variable  $Q_2$ .

En la figura 6.27 se observa que los residuales indican que los errores son multiplicativos, ya que decrecen cuando  $Q_2$  aumenta, y forman un cono. Por tanto, no sería adecuado emplear un modelo de regresión (6.12), aunque se cumplieran los otros supuestos.

Por tanto, en este problema aplicamos el modelo (6.31), donde de manera previa extraemos el logaritmo natural a ambos lados de la igualdad y obtenemos el modelo:

$$\ln(z) = \ln(\alpha) + \beta_1 \ln(Q_1) + \beta_2 \ln(Q_2) + \ln(\eta)$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Donde:

$$y = \ln(z), \beta_0 = \ln(\alpha), x_1 = \ln(Q_1), x_2 = \ln(Q_2) \text{ y } \varepsilon = \ln(\eta)$$

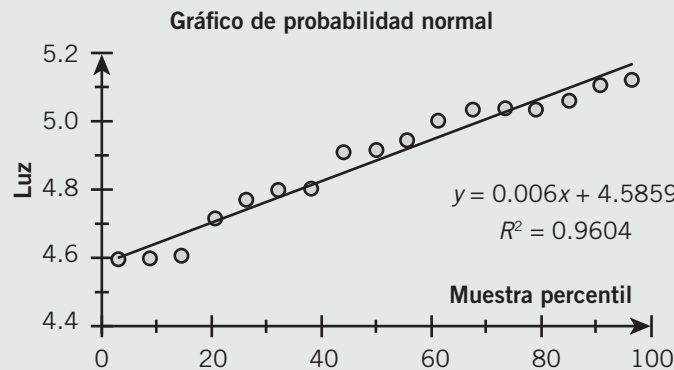
Para el segundo modelo, calculamos los residuales (véase tabla 6.38,) junto con las gráficas de normalidad (véase figura 6.28) y residuales (véase figura 6.29).

En la figura 6.28 se observa que las variables de respuesta después de aplicar el logaritmo sí cumplen con la normalidad. De igual manera, en las gráficas de la figura 6.29 se muestran los residuales que cumplen las condiciones de los supuestos para estimar los errores.

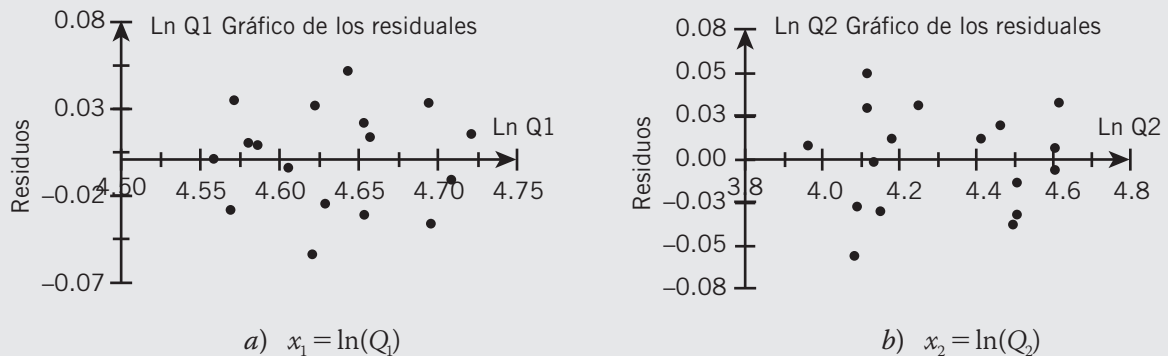
**Tabla 6.38** Logaritmo de las variables del problema de Henry Theil

Año	$y = \ln(z)$	$x_1 = \ln(Q_1)$	$x_2 = \ln(Q_2)$
1923	4.597	4.572	4.615
1924	4.595	4.586	4.606
1925	4.605	4.605	4.605
1926	4.715	4.653	4.506
1927	4.806	4.653	4.460
1928	4.767	4.696	4.496
1929	4.797	4.708	4.506
1930	4.913	4.721	4.416
1931	5.038	4.694	4.250

Año	$y = \ln(z)$	$x_1 = \ln(Q_1)$	$x_2 = \ln(Q_2)$
1932	5.034	4.657	4.181
1933	5.066	4.622	4.116
1934	4.946	4.558	4.135
1935	4.914	4.569	4.153
1936	5.124	4.581	3.963
1937	5.039	4.629	4.089
1938	5.004	4.621	4.086
1939	5.109	4.642	4.116



**Figura 6.28** Gráfica para la normalidad.



**Figura 6.29** Gráfica de los residuales.

A continuación se muestran las tablas de resultados de Excel.

1. El resumen de resultados se muestra en la tabla 6.39.

**Tabla 6.39** Resumen de los coeficientes de la regresión

Estadísticos de la regresión	
Coefficiente de correlación múltiple	0.98710
Coefficiente de determinación $R^2$	0.97436
$R^2$ ajustado	0.97070
Error típico	0.03118
Observaciones	17

El coeficiente de correlación ajustado es alto, esto indica fuertes evidencias de que el modelo de regresión múltiple es adecuado, pero esto se debe corroborar con el análisis de varianza.

2. El análisis de varianza para la prueba  $F$  se muestra en la tabla 6.40.

**Tabla 6.40** Tabla ANOVA de la prueba  $F$ .

<i>FV</i>	<i>GL</i>	<i>SC</i>	<i>CM</i>	<i>F</i>	<i>valor p</i>
Regresión	2	0.5173	0.2587	266.0179	7.2836E-12
Residuos	14	0.0136	0.0010		
Total	16	0.5309			

Para un nivel de significancia  $\alpha > 7.2836 \times 10^{-12}$  se concluye que debemos rechazar la hipótesis nula  $H_0: \beta_1 = 0, \beta_2 = 0$ . Entonces, las variables independientes sí explican el modelo.

3. El modelo de regresión, la prueba  $t$  y los intervalos de confianza para los parámetros del modelo se muestran en la tabla 6.41.

**Tabla 6.41** Tabla de la prueba  $t$ 

	Coefficientes	Error típico	Estadístico $t$	Probabilidad	Inferior 95%	Superior 95%
Intersección	3.1636	0.7048	4.4886	0.00051	1.6519	4.6752
Beta 1	1.1432	0.1560	7.3279	3.74E-06	0.8086	1.4777
Beta 2	-0.8288	0.0361	-22.9523	1.65E-12	-0.9063	-0.7514

El modelo de regresión lineal estimado está dado por:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 = 3.1636 + 1.1432x_1 - 0.8288x_2$$

Por tanto, después de aplicar la operación inversa del logaritmo, obtenemos el modelo Cobb-Douglas de mejor ajuste:

$$\hat{z} = \hat{\alpha} Q_1^{\hat{\beta}_1} Q_2^{\hat{\beta}_2} = e^{3.1636} Q_1^{1.1432} Q_2^{-0.8288} = 23.4565 Q_1^{1.1432} Q_2^{-0.8288}$$

A 95% de confianza los parámetros del modelo se encuentran en:

$$\beta_1 \in (0.8086, 1.4777)$$

$$\beta_2 \in (-0.9063, -0.7514)$$

## Modelos de regresión con variables de respuesta transformadas

Cuando tenemos una base de datos para ajustar un modelo de regresión lineal (6.12) puede ocurrir que las variables de respuesta no cumplan con la normalidad y, por ende, los errores tampoco cumplan este supuesto. En algunas ocasiones es posible que después de realizar una transformación,  $f$ , en la variable de respuesta el modelo (6.32) sí cumpla con los supuestos de un modelo de regresión lineal múltiple

$$f(y) = z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \varepsilon \quad (6.32)$$

En este caso, en lugar de ajustar el modelo:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \varepsilon,$$

lo hacemos con (6.32), pero debemos tener cuidado con la interpretación de los parámetros, ya que debido a la transformación utilizada para el ajuste, los parámetros  $\beta_i$  se interpretan como hemos visto en las secciones previas para la variable de respuesta  $z$ , pero no así para la variable de respuesta original,  $y$ .

Para la interpretación de los parámetros en la variable de respuesta original supongamos el modelo  $f(y) = z = \beta_0 + \beta_1 x + \varepsilon$ , estimado por  $z = \hat{\beta}_0 + \hat{\beta}_1 x$ . Ahora queremos saber cómo influye un cambio en una unidad de la variable predictiva,  $x$ , en la variable de respuesta,  $y$ . Después de ajustar el modelo en  $z$ , para los valores en  $y$ , requerimos que la transformación  $f$  sea invertible en el rango de valores de  $y$ , si cumple con esta condición, entonces  $y = f^{-1}(\hat{\beta}_0 + \hat{\beta}_1 x)$ . Por tanto, un incremento en  $\delta$  unidades de  $x$  dentro del rango permitido del modelo genera un incremento en  $y$  dado por (6.33). Sean los valores del incremento en  $x$ ,  $x = x_0$  y  $x_1 = x_0 + \delta$ , entonces  $y_1 = f^{-1}(\hat{\beta}_0 + \hat{\beta}_1 x_1) = f^{-1}(\hat{\beta}_0 + \hat{\beta}_1 x_0 + \hat{\beta}_1 \delta)$  y  $y_0 = f^{-1}(\hat{\beta}_0 + \hat{\beta}_1 x_0)$  son los valores de la variable de respuesta que tiene un incremento por unidad de incremento en  $x$ :

$$y_1 - y_0 = f^{-1}(\hat{\beta}_0 + \hat{\beta}_1 x_0 + \hat{\beta}_1 \delta) - f^{-1}(\hat{\beta}_0 + \hat{\beta}_1 x_0) \quad (6.33)$$

Otra forma de comparar el incremento de las variables de respuesta por  $\delta$  unidades de incremento de la variable predictiva está dada en (6.34) y se refiere al porcentaje de incremento o (decremento) de la variable de respuesta dado por:

$$\frac{y_1 - y_0}{y_0} = \frac{f^{-1}(\hat{\beta}_0 + \hat{\beta}_1 x_0 + \hat{\beta}_1 \delta) - f^{-1}(\hat{\beta}_0 + \hat{\beta}_1 x_0)}{f^{-1}(\hat{\beta}_0 + \hat{\beta}_1 x_0)} \quad (6.34)$$

Entonces, el problema a resolver consiste en determinar la transformación más adecuada para cada problema particular, entre las más utilizadas están las propuestas por los estadísticos George E. P. Box y David Cox, mejor conocidas como transformaciones Box-Cox.

- $z = \ln(y)$ . Con esta transformación es preferible utilizar el porcentaje de incremento de  $y$ , puesto que la inversa existe, es la exponencial que cumple la propiedad multiplicativa:

$$\exp(\hat{\beta}_0 + \hat{\beta}_1 x_0 + \hat{\beta}_1 \delta) = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_0) \exp(\hat{\beta}_1 \delta)$$

donde:

$$\begin{aligned} \frac{y_1 - y_0}{y_0} &= \frac{f^{-1}(\hat{\beta}_0 + \hat{\beta}_1 x_0 + \hat{\beta}_1 \delta) - f^{-1}(\hat{\beta}_0 + \hat{\beta}_1 x_0)}{f^{-1}(\hat{\beta}_0 + \hat{\beta}_1 x_0)} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_0) \exp(\hat{\beta}_1 \delta) - \exp(\hat{\beta}_0 + \hat{\beta}_1 x_0)}{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_0)} \\ &= \exp(\hat{\beta}_1 \delta) - 1 \end{aligned}$$

- $z = y^\alpha$ . La existencia de la inversa de esta transformación depende tanto del exponente como del rango de valores de la variable de respuesta. El incremento de la variable de respuesta es preferible calcularlo con la fórmula (6.33), ya que en general depende de ambos estimadores de los parámetros.

## Ejemplo 6.21

En una empresa se lleva a cabo un estudio para predecir el volumen de ventas, el responsable del modelo establece que influyen las siguientes variables:

- A: gastos de propaganda de la empresa en el periodo  $t$
- E: gastos de ventas de la empresa en el periodo  $t$
- S: volumen de ventas de la empresa en millones de pesos durante el periodo  $t$

Los resultados se muestran en la tabla 6.42.

Tabla 6.42 Datos para el volumen de ventas

Observación	S	A	E	ln(S)	Observación	S	A	E	ln(S)
1	7.474	1.988	0.31	2.0114	12	11.772	1.875	0.60	2.4658
2	4.529	1.944	0.30	1.5104	13	17.630	2.271	0.75	2.8696
3	6.478	2.200	0.38	1.8684	14	13.246	1.112	0.62	2.5837
4	4.979	2.001	0.33	1.6052	15	13.765	1.774	0.68	2.6221
5	8.416	1.693	0.43	2.1301	16	11.235	0.959	0.65	2.4190
6	5.960	1.743	0.32	1.7850	17	16.282	1.989	0.72	2.7900
7	8.150	2.069	0.39	2.0980	18	9.249	1.971	0.54	2.2245
8	8.386	1.017	0.41	2.1266	19	11.941	2.266	0.62	2.4799
9	7.960	2.019	0.48	2.0745	20	8.324	1.983	0.41	2.1191
10	7.591	1.061	0.45	2.0270	21	13.051	2.101	0.60	2.5689
11	13.714	1.460	0.59	2.6184	22	13.761	1.068	0.66	2.6219

Se pide proponer un modelo para ajustar el volumen de ventas de la empresa en el periodo  $t$ .

### Solución

Con los valores de la variable de respuesta de la tabla 6.42, trazamos un histograma de frecuencias, de lo que resulta la gráfica a) de la figura 6.30, donde podemos apreciar que la variable de respuesta no es normal y es creciente, entonces no es posible aplicar los modelos vistos en esta sección, ya que la variable de respuesta no cumple con la normalidad.

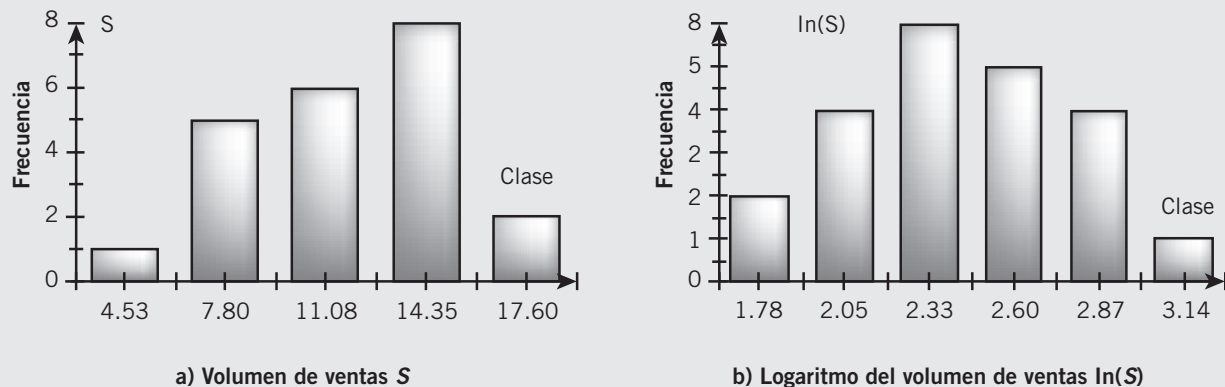


Figura 6.30 Histogramas de las distribuciones de frecuencias de la variable de respuesta.



Para poder utilizar los modelos vistos en esta sección, transformamos los valores de la variable de respuesta por medio de la función logaritmo, lo que da como resultado los valores de la derecha de la tabla 6.42, trazamos el histograma y resulta la gráfica *b*) de la figura 6.30, que muestra normalidad en el logaritmo de la variable de respuesta.

Proponemos el modelo de  $\ln(S) = \beta_0 + \beta_1 A + \beta_2 E + \varepsilon$ . Si se realizan las pruebas ANOVA y *t*, tenemos.

1. El resumen de resultados se muestra en la tabla 6.43.

**Tabla 6.43** Resumen de los coeficientes de la regresión

Estadísticos de la regresión	
Coefficiente de correlación múltiple	0.9448
Coefficiente de determinación $R^2$	0.8927
$R^2$ ajustado	0.8814
Error típico	0.1293
Observaciones	22

El coeficiente de correlación ajustado es alto, esto indica fuertes evidencias de que el modelo de regresión múltiple es adecuado, pero esto se debe corroborar con el análisis de varianza.

2. El análisis de varianza para la prueba *F* se muestra en la tabla 6.44.

**Tabla 6.44** Tabla ANOVA de la prueba *F*

<i>FV</i>	<i>GL</i>	<i>SC</i>	<i>CM</i>	<i>F</i>	<i>valor p</i>
Regresión	2	2.6432	1.3216	79.0033	6.198E-10
Residuos	19	0.3178	0.0167		
Total	21	2.9610			

Para un nivel de significancia  $\alpha > 6.198 \times 10^{-10}$  se concluye que debemos rechazar la hipótesis nula  $H_0: \beta_1 = 0, \beta_2 = 0$ . Entonces las variables independientes sí explican el modelo.

3. El modelo de regresión, la prueba *t* y los intervalos de confianza para los parámetros del modelo se muestran en la tabla 6.45.

**Tabla 6.45** Tabla de la prueba *t*

	Coefficientes	Error típico	Estadístico <i>t</i>	Probabilidad	Inferior 95%	Superior 95%
Intesección	0.9969	0.1642	6.0712	7.723E-06	0.6532	1.3406
Beta 1	-0.0003	0.0654	-0.0047	0.9963	-0.1373	0.1366
Beta 2	2.4644	0.1976	12.4734	1.344E-10	2.0509	2.8779

El valor *p* para la variable *A* es muy grande, esto significa que no hay evidencias para rechazar  $H_0: \beta_1 = 0$ . Entonces, tenemos que revisar el modelo  $\ln(S) = \beta_0 + \beta_2 E + \varepsilon$ .

4. El resumen de resultados se muestra en la tabla 6.46.

**Tabla 6.46** Resumen de los coeficientes de la regresión

Estadísticos de la regresión	
Coefficiente de correlación múltiple	0.9448
Coefficiente de determinación $R^2$	0.8927
$R^2$ ajustado	0.8873
Error típico	0.1261
Observaciones	22

El coeficiente de correlación ajustado es alto y como estamos en el caso de una variable predictora no es tan necesario el análisis ANOVA.

5. El análisis de varianza para la prueba  $F$  se muestra en la tabla 6.47.

**Tabla 6.47** Tabla ANOVA de la prueba  $F$ 

$FV$	$GL$	$SC$	$CM$	$F$	valor $p$
Regresión	1	2.6432	2.6432	166.3224	3.767E-11
Residuos	20	0.3178	0.0159		
Total	21	2.9610			

Para un nivel de significancia  $\alpha > 3.767 \times 10^{-11}$  se concluye que tenemos que rechazar la hipótesis nula  $H_0: \beta_2 = 0$ . Entonces la variable independiente sí explica el modelo.

6. El modelo de regresión, la prueba  $K$  y los intervalos de confianza para los parámetros del modelo se muestran en la tabla 6.48.

**Tabla 6.48** Tabla de la prueba  $T$ 

	Coefficientes	Error típico	Estadístico $t$	Probabilidad	Inferior 95%	Superior 95%
Intersección	0.9963	0.1013	9.8385	4.158E-09	0.7851	1.2075
Beta2	2.4645	0.1911	12.8966	3.767E-11	2.0659	2.8631

El modelo de regresión lineal estimado está dado por:

$$\ln(\hat{S}) = \hat{\beta}_0 + \hat{\beta}_2 E = 0.9963 + 2.4645E$$

A 95% de confianza el parámetro del modelo se encuentra en:

$$\beta_2 \in (2.0659, 2.8631)$$

Por tanto, el porcentaje de incremento en  $S$  para un incremento  $\delta = 0.1$  en  $E$  es igual a:

$$\frac{S_1 - S_0}{S_0} = \exp(\hat{\beta}_1 \delta) - 1 = \exp(2.4645 \times 0.1) - 1 = \exp(0.24645) - 1 = 0.2795 = 27.95\%$$

Note que no se consideran incrementos en  $E$  enteros, ya que quedarían fuera del rango de valores.

## Ejercicios 6.5

1. En el ejemplo 6.21 realice una transformación cúbica en las variables de respuesta, ajuste el modelo y compare la suma de valores absolutos de los residuales. ¿Qué le indica este resultado?

2. Con los datos del ejemplo 6.20 ajuste un modelo de regresión lineal múltiple y compare la suma de valores absolutos de los residuales. ¿Qué le indica este resultado?
3. En el ejemplo 6.21 ajuste los datos originales a un modelo de regresión lineal múltiple y compare la suma de valores absolutos de los residuales. ¿Qué le indica este resultado?
4. En una empresa se lleva a cabo un estudio para predecir el volumen de ventas, el responsable del modelo establece que en las ventas influyen las siguientes variables:
  - A: gastos de propaganda de la empresa al tiempo  $t$
  - P: gastos de promoción de la empresa al tiempo  $t$
  - E: gastos de ventas de la empresa al tiempo  $t$
  - AR y PR: valores de A y P, respectivamente, retrasados en un periodo  $t$
  - S: volumen de ventas de la empresa

Los resultados se muestran en la tabla 6.49.

**Tabla 6.49** Datos para el volumen de ventas

Observación	A	P	E	AR	PR	S
1	1.98786	1.0	0.30	2.01722	0.0	20.1137
2	1.94418	0.0	0.30	1.98786	1.0	15.1044
3	2.19954	0.8	0.35	1.94418	0.0	18.6837
4	2.00107	0.0	0.35	2.19954	0.8	16.0517
5	1.69292	1.3	0.30	2.00107	0.0	21.3010
6	1.74334	0.3	0.32	1.69292	1.3	17.8500
7	2.06907	1.0	0.31	1.74334	0.3	18.8756
8	1.01709	1.0	0.41	2.06907	1.0	21.2660
9	2.01906	0.9	0.45	1.01709	1.0	20.4847
10	1.06139	1.0	0.45	2.01906	0.9	20.5403
11	1.45999	1.5	0.50	1.06139	1.0	26.1844
12	1.87511	0.0	0.60	1.45999	1.5	21.7161
13	2.27109	0.8	0.65	1.87511	0.0	28.6959
14	1.11191	1.0	0.65	2.27109	0.8	25.8372
15	1.77407	1.2	0.65	1.11191	1.0	29.3199
16	0.95878	1.0	0.65	1.77407	1.2	24.1904
17	1.98930	1.0	0.62	0.95878	1.0	26.5897
18	1.97111	0.0	0.60	1.98930	1.0	22.2447
19	2.26603	0.7	0.60	1.97111	0.0	24.7994
20	1.98343	0.1	0.61	2.26603	0.7	21.1910
21	2.10054	1.0	0.60	1.98343	0.1	26.0344
22	1.06815	1.0	0.58	2.10054	1.0	27.3930

- a) Realice una prueba de bondad de ajuste Q-Q para verificar la normalidad de los valores de la variable de respuesta.
- b) Calcule una matriz de correlaciones de las variables predictivas. ¿Qué puede concluir de los resultados obtenidos?

- c) Ajuste el modelo de regresión lineal múltiple a las cinco variables predictivas. ¿Cuáles son los valores de los tres coeficientes de correlación y cómo se interpretan?
5. Con las observaciones del ejercicio 4:
- Formule el contraste de hipótesis para el análisis de varianza del modelo de regresión lineal múltiple y concluya sobre el contraste de hipótesis.
  - Proporcione el mejor modelo de regresión lineal que ajusta a las cinco variables independientes. ¿Cómo se interpreta cada uno de los  $\hat{\beta}_i$  para  $i = 1, 2, 3, 4, 5$ ?
  - Realice la prueba  $t$  a 5% de significancia para los contrastes de hipótesis y concluya sobre los resultados:

$$\begin{cases} H_0: \beta_i = 0 \\ H_1: \beta_i \neq 0 \end{cases} \text{ con } i = 1, 2, 3, 4, 5$$

6. Con las observaciones del ejercicio 4 y los resultados de los ejercicios 4 y 5:
- Calcule un intervalo de 95 y 99% de confianza para los parámetros  $\hat{\beta}_i$  del modelo con  $i = 1, 2, 3, 4, 5$ .
  - Explique si el modelo es adecuado.
  - Determine la posible multicolinealidad entre las variables.
  - Con base en las respuestas anteriores, ¿qué puede concluir del modelo?
7. Con las observaciones del ejercicio 4 y los resultados de los ejercicios 4, 5 y 6:
- Calcule un intervalo a 95% de confianza para la estimación del volumen de ventas medias cuando  $A_0 = 2.15$ ,  $P_0 = 1.21$ ,  $E_0 = 0.40$ ,  $AR_0 = 2.05$  y  $PR_0 = 1.10$ .
  - Calcule un intervalo a 95% de confianza para la estimación del volumen de ventas en un periodo  $t$ , cuando  $A_0 = 2.15$ ,  $P_0 = 1.21$ ,  $E_0 = 0.40$ ,  $AR_0 = 2.05$  y  $PR_0 = 1.10$ .
8. Con las observaciones del ejercicio 4 y los resultados de los ejercicios 4, 5, 6 y 7, proponga un modelo parsimonioso.
9. Los datos de la tabla 6.50 corresponden a la economía de un país, donde:
- $x_1$ : insumos laborales en millones de años-hombre
  - $x_2$ : capital real en miles de millones de dólares medidos a partir de un origen arbitrario
  - $x_3$ : años medidos a partir de 1991, tomado como año 1
  - $y$ : producto real bruto nacional en miles de millones de dólares

Tabla 6.50 Datos para la economía del país

Observación	$x_1$	$x_2$	$x_3$	$Y$
1	47	54	1	142
2	43	59	2	127
3	39	57	3	113
4	34	48	4	98
5	34	36	5	94
6	36	24	6	102
7	38	19	7	116
8	41	18	8	128
9	42	22	9	140
10	37	24	10	131
11	40	23	11	143
12	42	27	12	157

Observación	$x_1$	$x_2$	$x_3$	$Y$
13	47	36	13	182
14	51	9	18	209
15	53	25	19	214
16	53	39	20	225
17	50	51	21	221
18	52	62	22	243
19	54	75	23	257
20	54	94	24	265
21	55	108	25	276
22	52	118	26	271
23	54	124	27	291

- a) Realice una prueba de bondad de ajuste Q-Q para verificar la normalidad de los valores de la variable del producto real bruto.
- b) Calcule una matriz de correlaciones de las variables predictivas. ¿Qué puede concluir de los resultados obtenidos?
- c) Ajuste el modelo de regresión lineal múltiple a las tres variables predictivas. ¿Cuáles son los valores de los tres coeficientes de correlación y cómo se interpretan?

10. Con las observaciones del ejercicio 9.

- a) Formule el contraste de hipótesis para el análisis de varianza del modelo de regresión lineal múltiple y concluya sobre el contraste de hipótesis.
- b) Proporcione el mejor modelo de regresión lineal que ajusta a las tres variables predictivas. ¿Cómo se interpreta cada uno de los  $\hat{\beta}_i$  para  $i = 1, 2, 3$ ?

Realice la prueba  $t$ .

- a) Realícelo a 5% de significancia para los contrastes de hipótesis y concluya sobre los resultados:

$$\begin{cases} H_0: \beta_i = 0 \\ H_1: \beta_i \neq 0 \end{cases} \text{ con } 1, 2, 3$$

11. Con las observaciones del ejercicio 9 y los resultados de los ejercicios 9 y 10:

- a) Calcule un intervalo de 95 y 99% de confianza para los parámetros  $\beta_i$  del modelo con  $i = 1, 2, 3$ .
- b) Explique si el modelo es adecuado.
- c) Determine la posible multicolinealidad entre las variables.
- d) Con base en las respuestas anteriores, ¿qué puede concluir del modelo?

12. Con las observaciones del ejercicio 9 y los resultados de los ejercicios 9, 10 y 11:

- a) Calcule un intervalo a 95% de confianza para la estimación del producto real bruto medio cuando  $x_{01} = 51$ ,  $x_{02} = 80$  y  $x_{03} = 22$ .
- b) Calcule un intervalo a 95% de confianza para la estimación del producto real bruto en un año cuando  $x_{01} = 51$ ,  $x_{02} = 80$  y  $x_{03} = 22$ .

13. Con las observaciones del ejercicio 9 y los resultados de los ejercicios 9, 10, 11 y 12, proponga un modelo parsimonioso.

## Ejercicios de repaso

## Preguntas de autoevaluación

- 6.1 ¿Qué es un modelo de regresión lineal? ¿Para qué se puede utilizar?
- 6.2 ¿Cuáles son los supuestos de un modelo de regresión lineal?
- 6.3 ¿Qué es el coeficiente de determinación ajustado?
- 6.4 ¿En qué consiste el problema de la multicolinealidad?
- 6.5 ¿Qué se puede hacer para identificar el problema de multicolinealidad?
- 6.6 ¿Cuándo se dice que los errores son multiplicativos y cuándo son aditivos?
- 6.7 ¿Qué es una transformación Box-Cox?
- 6.8 ¿Qué información proporciona la prueba  $F$ ?
- 6.9 ¿Para qué se utiliza la prueba  $t$ ?
- 6.10 ¿Cómo se puede eliminar la multicolinealidad en un modelo de regresión?
- 6.11 ¿Para qué se utiliza una transformación Box-Cox en un modelo de regresión?
- 6.12 ¿Cuál es la interpretación del parámetro  $\beta_1$  en un modelo de regresión lineal simple?
- 6.13 ¿Cómo se puede interpretar el parámetro  $\beta_1$  en un modelo de regresión lineal múltiple?
- 6.14 ¿Por qué no es de interés inferir sobre el parámetro  $\beta_0$  en un modelo de regresión lineal?
- 6.15 ¿Cómo se puede probar que los errores son multiplicativos?
- 6.16 Con el material estudiado en el texto, ¿cómo se puede probar el supuesto de normalidad?
- 6.17 ¿Cuáles son las condiciones para poder aplicar una transformación en un conjunto de observaciones para llevar a cabo un ajuste de datos?
- 6.18 ¿En qué situaciones se recomienda utilizar un modelo de regresión curvilíneo?
- 6.19 ¿Qué es un modelo parsimonioso?
- 6.20 ¿Qué recomienda hacer ante la presencia de observaciones aberrantes?

## Ejercicios complementarios con grado de dificultad uno

- 6.21 Un comerciante al menudeo realizó un estudio para determinar la relación entre los gastos de publicidad semanal y las ventas (véase tabla 6.51).

Tabla 6.51 Gastos de publicidad semanal y ventas

Costos de publicidad	40	20	25	20	30	50	40	20	50	40	25	50
Ventas	385	400	395	365	475	440	490	420	560	525	480	510

- a) Encuentre la ecuación de la recta de regresión para pronosticar las ventas semanales resultantes de los

gastos de publicidad. Pruebe  $H_0: \beta_1 \leq 2$  contra  $H_1: \beta_1 > 2$  con  $\alpha = 0.10$ .

- b) Estime las ventas semanales cuando los gastos de publicidad sean de \$35 y obtenga un intervalo a 90% de confianza para la predicción.

- 6.22 Con la información del ejercicio anterior:

- a) Calcule los tres coeficientes de correlación.

- b) Con un  $\alpha = 0.05$  pruebe  $H_0: \rho \leq 0.50$  contra  $H_1: \rho > 0.50$ .

- 6.23 Los siguientes datos muestran la relación entre el peso y el volumen de una mezcla.

Tabla 6.52 Peso y volumen de una mezcla

Peso en kilogramos	2.8	2.1	3.4	4.1	1.9
Volumen en $\text{cm}^3$	7.9	4.8	11.1	15.8	3.8

- a) Encuentre la ecuación de la recta de regresión para pronosticar el volumen. Pruebe  $H_0: \beta_1 \leq 3$  contra  $H_1: \beta_1 > 3$  con  $\alpha = 0.05$ .

- b) Estime el volumen de la mezcla, para un peso de 3.85 kg y obtenga un intervalo a 95% de confianza para la predicción.

- 6.24 Con la información del ejercicio anterior:

- a) Calcule los tres coeficientes de correlación.

- b) Con un  $\alpha = 0.05$  pruebe  $H_0: \rho \leq 0.50$  contra  $H_1: \rho > 0.50$ .

- 6.25 En la tabla 6.53 se presentan datos muestrales sobre el número de horas de estudio invertidas por los estudiantes fuera de clase, durante un periodo de tres semanas, para un curso de estadística, junto con las calificaciones que obtuvieron en un examen aplicado al final del periodo.

Tabla 6.53 Horas de estudio y calificaciones

Horas de estudio ( $x$ )	20	16	34	23	27	32	18	22
Calificación ( $y$ )	64	61	84	70	88	92	72	77

- a) Encuentre la ecuación de la recta de regresión. Pruebe  $H_0: \beta_1 = 0$  contra  $H_1: \beta_1 \neq 0$  con  $\alpha = 0.05$ .

- b) Estime la calificación para 28 horas de estudio y obtenga un intervalo a 95% de confianza para la predicción.

- 6.26 Con la información del ejercicio anterior:

- a) Calcule los tres coeficientes de correlación.

- b) Con un  $\alpha = 0.05$  pruebe  $H_0: \rho \leq 0.50$  contra  $H_1: \rho > 0.50$ .

- 6.27 Un analista determinará si existe una relación lineal entre el consumo de electricidad y el número de cuartos en una vivienda unifamiliar. Como el consumo de electricidad varía de un mes a otro, decide estudiar el mes de enero. Para ello obtiene los datos que se muestran en la tabla 6.54.

**Tabla 6.54** Consumo de electricidad y cuartos en una vivienda unifamiliar

Números de cuartos ( $x$ )	13	10	15	7	9	5	8	9	4	6	14	5
Kilovatios-hora (miles), $y$	8	6	9	5	8	5	7	9	3	6	8	6

- a) Encuentre la ecuación de la recta de regresión. Pruebe  $H_0: \beta_1 = 0$  contra  $H_1: \beta_1 \neq 0$  con  $\alpha = 0.05$ .
- b) Estime el consumo de electricidad para ocho cuartos y obtenga un intervalo a 95% de confianza para la predicción.

**6.28** Con la información del ejercicio anterior:

- a) Calcule los tres coeficientes de correlación.
- b) Con un  $\alpha = 0.05$  pruebe  $H_0: \rho \leq 0.50$  contra  $H_1: \rho > 0.50$ .

**6.29** La tabla 6.55 muestra la edad  $x$  y la presión sanguínea  $y$  de 12 mujeres.

**Tabla 6.55** Edad y presión sanguínea de mujeres

Edad $x$	56	42	72	36	63	47	55	49	38	42	68	60
Presión sanguínea $y$	147	125	160	118	149	128	150	145	115	140	152	155

- a) Encuentre la ecuación de la recta de regresión. Pruebe  $H_0: \beta_1 \leq 1$  contra  $H_1: \beta_1 > 1$  con  $\alpha = 0.05$ .
- b) Estime la presión sanguínea en una mujer de 55 años y obtenga un intervalo a 95% de confianza para la predicción.

**6.30** Con la información del ejercicio anterior:

- a) Calcule los tres coeficientes de correlación.
- b) Con un  $\alpha = 0.05$  pruebe  $H_0: \rho \leq 0.50$  contra  $H_1: \rho > 0.50$ .

**6.31** El dueño de una compañía contrata un analista de tiempo parcial y le pide que analice la relación entre el número de permisos de construcción emitidos y la cantidad de trabajo disponible para su empresa. Se cuenta con datos como son las tasas de interés sobre hipotecas para pronosticar el número de permisos de construcción que se emitirán el mes próximo.

**Tabla 6.56** Permisos de construcción y cantidad de trabajo disponible

Tasa de interés ( $x$ )	10.2	12.6	13.5	9.7	10.8	9.5	10.9	9.2	14.2
Permisos emitidos, $y$	786	494	289	892	343	888	509	987	187

- a) Encuentre la ecuación de regresión de la recta. Pruebe  $H_0: \beta_1 \leq 30$  contra  $H_1: \beta_1 > 30$  con  $\alpha = 0.02$ .
- b) Estime los permisos emitidos para una tasa de interés de 12.5% y obtenga un intervalo a 98% de confianza para la predicción.

**6.32** Con la información del ejercicio anterior:

- a) Calcule los tres coeficientes de correlación.
- b) Con un  $\alpha = 0.05$  pruebe  $H_0: \rho \leq 0.50$  contra  $H_1: \rho > 0.50$ .

**6.33** Una empresa de calzado que realiza ventas por catálogo, tiene datos sobre 10 ciudades.

**Tabla 6.57** Ventas por catálogos

Número de catálogos distribuidos (miles), $x$	7	3	5	2	10	8	14	4	1
Número de órdenes por correo recibidas (miles), $y$	23	17	23	14	31	24	37	19	15

- a) Encuentre la ecuación de regresión de la recta. Pruebe  $H_0: \beta_1 \leq 2$  contra  $H_1: \beta_1 > 2$  con  $\alpha = 0.02$ .
- b) Estime el número de órdenes recibidas para 9 000 catálogos distribuidos y obtenga un intervalo a 98% de confianza para la predicción.

**6.34** Con la información del ejercicio anterior:

- a) Calcule los tres coeficientes de correlación.
- b) Con un  $\alpha = 0.05$  pruebe  $H_0: \rho \leq 0.50$  contra  $H_1: \rho > 0.50$ .

**6.35** El gerente de una cadena de librerías desea pronosticar las ventas semanales de los libros de bolsillo, para ello se basa en la cantidad de espacio en las repisas (en pies) que se le proporciona. El gerente reúne una muestra de 11 semanas.

**Tabla 6.58** Ventas semanales de libros

Pies de espacio en repisa $x$	6.8	3.3	4.1	4.2	4.8	3.9	4.9	7.7	3.1	5.9	5
Número de libros vendidos $y$	275	142	168	197	215	188	241	295	125	266	200

- a) Encuentre la ecuación de regresión de la recta. Pruebe  $H_0: \beta_1 \leq 30$  contra  $H_1: \beta_1 > 30$  con  $\alpha = 0.10$ .
- b) Estime el número de libros de bolsillo para una semana en la que se proporcionan 4 pies de espacio en las repisas y obtenga un intervalo a 90% de confianza para la predicción.

**6.36** Con la información del ejercicio anterior:

- a) Calcule los tres coeficientes de correlación.
- b) Con un  $\alpha = 0.05$  pruebe  $H_0: \rho \leq 0.50$  contra  $H_1: \rho > 0.50$ .

**6.37** La directora de investigación y desarrollo de una empresa pretende defender su petición para aumentar los fondos de su presupuesto. Para su defensa obtuvo una muestra de ocho compañías farmacéuticas con los siguientes datos.



**Tabla 6.59** Ganancia e inversión para investigación de compañías farmacéuticas

$x =$ Inversión para investigación (millones de \$)	5	7	4	10	8	12	6	11
$y =$ Ganancia anual (millones de \$)	25	30	20	50	40	60	50	35

- a) Encuentre la ecuación de regresión de la recta. Pruebe  $H_0: \beta_1 \leq 2.5$  contra  $H_1: \beta_1 > 2.5$  con  $\alpha = 0.04$ .
- b) Estime la ganancia anual para una inversión de 9 millones y obtenga un intervalo a 96% de confianza para la predicción.

**6.38** Con la información del ejercicio anterior:

- a) Calcule los tres coeficientes de correlación.
- b) Con un  $\alpha = 0.05$  pruebe  $H_0: \rho \leq 0.50$  contra  $H_1: \rho > 0.50$ .

**6.39** El director de personal de una empresa cree que existe relación entre las edades de los operadores de computadoras y el número de días que faltaron a trabajar el mes pasado.

**Tabla 6.60** Edades de operadores y faltas

Edad ( $x$ )	25	30	62	33	45	27	55	41	22	58
Días ausentes ( $y$ )	5	0	1	7	8	12	6	11	2	3

- a) Encuentre la ecuación de regresión de la recta. Pruebe  $H_0: \beta_1 = 0$  contra  $H_1: \beta_1 \neq 0$  con  $\alpha = 0.08$ .
- b) Estime los días de ausencia en la empresa de los operadores de una edad de 23 años y obtenga un intervalo a 92% de confianza para la predicción.

**6.40** Con la información del ejercicio anterior:

- a) Calcule los tres coeficientes de correlación.
- b) Con un  $\alpha = 0.05$  pruebe  $H_0: \rho \leq 0.50$  contra  $H_1: \rho > 0.50$ .

## Ejercicios complementarios con grado de dificultad dos

**6.41** Se lleva a cabo un estudio acerca de la cantidad de azúcar refinada mediante un cierto proceso a varias temperaturas diferentes. Los datos para  $X =$  temperatura (entre 1.6 y 2.1) y  $Y =$  azúcar refinada, dieron los siguientes resultados:

$$\sum_{i=1}^{11} x_i = 19.33; \quad \sum_{i=1}^{11} x_i^2 = 34.1455; \quad \sum_{i=1}^{11} y_i = 110.3;$$

$$\sum_{i=1}^{11} y_i^2 = 1110.89; \quad \sum_{i=1}^{11} x_i y_i = 194.229$$

- a) Determine la línea de regresión estimada.
- b) Estime la cantidad de azúcar refinada que se produce cuando la temperatura codificada es de 1.75.

**6.42** Dada la siguiente información para las variables  $x$  (aleatoria) y  $y$  (no aleatoria)

$$\sum_{i=1}^9 x_i = 45; \quad \sum_{i=1}^9 x_i^2 = 285; \quad \sum_{i=1}^9 y_i = 45;$$

$$\sum_{i=1}^9 y_i^2 = 285; \quad \sum_{i=1}^9 x_i y_i = 168$$

determine la línea de regresión estimada.

**6.43** Los siguientes datos muestran la relación entre el peso ( $x$ ) y el tamaño de tórax ( $y$ ) de recién nacidos:

$$\sum_{i=1}^9 x_i = 32.68; \quad \sum_{i=1}^9 x_i^2 = 128.66; \quad \sum_{i=1}^9 y_i = 267.20;$$

$$\sum_{i=1}^9 y_i^2 = 8008.28; \quad \sum_{i=1}^9 x_i y_i = 991.87$$

- a) Encuentre la ecuación de la línea de regresión lineal para pronosticar el tamaño del tórax de los recién nacidos.
- b) Estime el tamaño del tórax de un recién nacido que pesó 3.80 kg.

**6.44** Los siguientes datos muestran,  $x$  la fuerza de tensión aplicada a una probeta de acero en miles de libras y  $y$  es la elongación resultante en milésimas de pulgadas:

$$\sum_{i=1}^6 x_i = 21; \quad \sum_{i=1}^6 x_i^2 = 91; \quad \sum_{i=1}^6 y_i = 311;$$

$$\sum_{i=1}^6 y_i^2 = 19855; \quad \sum_{i=1}^6 x_i y_i = 1342$$

Estime los parámetros de la recta de regresión y obtenga la expresión de la recta.

## Ejercicios complementarios con grado de dificultad tres

Los artículos a los que se hace referencia en el texto se encuentran en la página del libro en SALI, en la carpeta archivos de regresión.

**6.45** Si en la variable de respuesta de un conjunto de observaciones se aplica la transformación raíz cúbica y se considera que  $\beta_0 = 0$ , ¿cómo quedaría la interpretación de  $\beta_1 = 0$ , ante un incremento de la variable explicativa en  $\delta$  unidades?

**6.46** Con la información de la tabla 1, página 300 del artículo 3 (véase página electrónica del libro en SALI) considere a la variable de respuesta frecuencia máxima medida en Hertz (Average Peak Frequency) con variable predictiva edad en años (*age*). Calcule el modelo que mejor ajuste a los datos.

**6.47** Con la gráfica de la figura 4 de la página 2232 del artículo 4 (véase página electrónica del libro en SALI), ajuste los puntos.



- a) Obtenga el mejor modelo lineal que ajuste a las observaciones.
- b) Logre el mejor modelo cuadrático que ajuste a las observaciones.
- c) Compare ambos resultados y concluya.
- 6.48** Con la gráfica de la figura 3 de la página 785 del artículo 6 (véase página electrónica del libro en SALI), ajuste los puntos de las mujeres.
- a) Obtenga el mejor modelo lineal que ajuste a las observaciones.
- b) Logre el mejor modelo cuadrático que ajuste a las observaciones.
- c) Compare ambos resultados y concluya.
- d) ¿Qué puede decir de las observaciones aberrantes?
- 6.49** Con la gráfica de la figura 3 de la página 785 del artículo 6 (véase página electrónica del libro en SALI), ajuste los puntos de los jóvenes.
- a) Obtenga el mejor modelo lineal que ajuste a las observaciones.
- b) Logre el mejor modelo cuadrático que ajuste a las observaciones.
- c) Compare ambos resultados y concluya.
- d) ¿Qué puede decir de las observaciones aberrantes?
- 6.50** Con la gráfica de la figura 4 de la página 132 del artículo 5 (véase página electrónica del libro en SALI), ajuste los puntos de la calificación y la calificación del profesor.
- a) Obtenga el mejor modelo lineal que ajuste a las observaciones.
- b) Obtenga el mejor modelo cuadrático que ajuste a las observaciones.
- c) Compare ambos resultados y concluya.
- d) ¿Qué puede decir de las observaciones aberrantes?
- 6.51** Con la información de la tabla 2, página 830 del artículo 2 (véase página electrónica del libro en SALI) y considerando a la variable de respuesta tamaño del grano *equiaxed ingot* en mm (*Grain size equiaxed ingot*), con variables de predicción  $x_1$ , distancia de la cara fría en mm (*Distance from chill face*),  $x_2$  velocidad de enfriamiento y temperatura de solidificación en K/s (*Cooling rate at liquidus temperature*) y  $x_3$ , gradiente de temperatura en K/mm (*Temperature gradient at liquidus*). Obtenga el mejor modelo lineal que ajuste a  $y$  con  $x_1$ .
- a) Logre el mejor modelo cuadrático que ajuste a  $y$  con  $x_1$ .
- b) Consiga el mejor modelo lineal que ajuste a  $y$  con  $x_1$ ,  $x_2$  y  $x_3$ .
- c) Compare y concluya sobre los tres modelos.

## Proyectos de la unidad 6

I. En el archivo artículo 1, tabla 1 de la página 131 (véase página del libro en SALI) se encuentran los valores de las variables:

$x_1$ : valor estimado del terreno

$x_2$ : valor estimado de las mejoras y  $y$ : precio de venta

Referentes a propiedades escogidas al azar.

- a) Realice una prueba de bondad de ajuste Q-Q para verificar la normalidad de los precios de venta.
- b) Calcule matriz de correlaciones de las variables predictivas. ¿Qué puede concluir de los resultados obtenidos?
- c) Ajuste el modelo de regresión lineal múltiple a las dos variables predictivas. ¿Cuáles son los valores de los tres coeficientes de correlación y cómo se interpretan?
- d) Formule el contraste de hipótesis para el análisis de varianza del modelo de regresión lineal múltiple y concluya sobre el contraste de hipótesis.
- e) Proporcione el mejor modelo de regresión lineal que ajusta a las dos variables predictivas. ¿Cómo se interpreta cada uno de los  $\hat{\beta}_i$  para  $i = 1, 2$ ?
- f) Realice la prueba  $t$  a 5% de significancia para los contrastes de hipótesis y concluya sobre los resultados

$$\begin{cases} H_0: \beta_i = 0 \\ H_1: \beta_i \neq 0 \end{cases} \text{ con } i = 1, 2$$

- a) Calcule un intervalo de 95% de confianza para los parámetros  $\beta_i$  del modelo con  $i = 1, 2$ .
- b) Explique si éste es adecuado.
- c) Determine la posible multicolinealidad entre las variables.
- d) Con base en las respuestas anteriores, ¿qué puede concluir del modelo?

- e) Calcule un intervalo a 95% de confianza para la estimación del precio de venta media cuando  $x_{01} = 2.51$  y  $x_{02} = 8.31$ .
- f) Calcule un intervalo a 95% de confianza para la estimación del precio de venta de una propiedad cuando  $x_{01} = 2.51$  y  $x_{02} = 8.31$ .
- g) Proponga un modelo parsimonioso.

**II. Modelo Cobb-Douglas.** Los datos siguientes se refieren a un modelo tipo de Cobb-Douglas, en el que  $N$ : nitrógeno en kg/Ha,  $D$ -kg de semilla/Ha y  $Y$ : rendimiento de grano de maíz en kg/Ha. Ajuste los datos de la hoja de Excel, ejercicio 53, a un modelo Cobb-Douglas, de la forma:

$$y_i = \beta_0 + \beta_1 N_i^{0.8} + \beta_2 D_i^{0.7} + \beta_3 N_i^{1.6} + \beta_4 D_i^{1.4} + \beta_5 N_i^{0.8} D_i^{0.7} + e_i$$

- a) Acomode los datos anteriores a un modelo de regresión múltiple.
- b) Explique si el modelo es adecuado.
- c) Determine la posible multicolinealidad entre las variables.
- d) Con base en las respuestas anteriores, ¿qué puede concluir del modelo?

**III.** Con el modelo del ejercicio 53 se lleva a cabo un estudio económico de optimización del ingreso sobre la producción de cierto producto, sujeto a:

$$y_i = \beta_0 + \beta_1 N_i^{0.8} + \beta_2 D_i^{0.7} + \beta_3 N_i^{1.6} + \beta_4 D_i^{1.4} + \beta_5 N_i^{0.8} D_i^{0.7} + e_i$$

cuyas observaciones se encuentran arriba. Su función de ingreso es:

$$I = P_Y \hat{y} - C_0 - P_N N - P_D D$$

donde, se ha estimado que  $P_N = \$5.17$ ,  $P_D = \$3.84$ ,  $P_T = \$3.25$  y costo fijo de  $C_0 = \$2860$ .

Obtenga el ingreso máximo y los valores de las variables  $N$  y  $D$ .

**IV.** Con la tabla 1 de la página 6 del artículo 7 (la tabla también está en la hoja del ejercicio 55 de Excel), ajuste un modelo de regresión a la variable de respuesta tiempo en taladrar en húmedo vs. la variable predictiva profundidad en pies.

- a) Realice una prueba de bondad de ajuste Q-Q para verificar la normalidad de los tiempos, considere la segunda columna de la tabla.
- b) Ajuste el modelo de regresión lineal simple.
- c) Realice las pruebas  $F$  y  $t$  y obtenga los resultados.

**V.** Con la tabla 1 de la página 6 del artículo 7, ajuste un modelo de regresión a la variable de respuesta tiempo en taladrar en húmedo vs. la variable predictiva profundidad en pies.

- a) Realice una prueba de bondad de ajuste Q-Q para verificar la normalidad de los tiempos, considere la cuarta columna de la tabla.
- b) Ajuste el modelo de regresión lineal simple.
- c) Realice las pruebas  $F$  y  $t$  y obtenga los resultados.

**VI.** Con la tabla 1 de la página 6 del artículo 7, ajuste un modelo de regresión a la variable de respuesta tiempo en taladrar en seco vs. la variable predictiva profundidad en pies.

- a) Realice una prueba de bondad de ajuste Q-Q para verificar la normalidad de los tiempos, considere la quinta columna de la tabla.
- b) Ajuste el modelo de regresión lineal simple.
- c) Realice las pruebas  $F$  y  $t$  y obtenga los resultados.

**VII.** Con la tabla 1 de la página 6 del artículo 7, ajuste un modelo de regresión a la variable de respuesta tiempo en taladrar en seco vs. la variable predictiva profundidad en pies.

- a) Realice una prueba de bondad de ajuste Q-Q para verificar la normalidad de los tiempos, considere la séptima columna de la tabla.
- b) Ajuste el modelo de regresión lineal simple.
- c) Realice las pruebas  $F$  y  $t$  y obtenga los resultados.

VIII. En la hoja de Excel, ejercicio 60, están los valores de las variables para el rendimiento de azúcar del ingenio Aarón Sáenz, en el cual las variables se definen como:

$x_1$ : porcentaje de humedad en la sección 8-10 de tallos de caña de azúcar

$x_2$ : temperatura mínima media cuatro semanas antes de la cosecha de la caña de azúcar

$x_3$ : precipitación acumulada en el periodo de ocho semanas antes de la cosecha de la caña de azúcar

$y$ : rendimiento de fábrica (azúcar producido, % de caña)

Se pide buscar el modelo de regresión que mejor ajuste a las observaciones.

## Caso de estudio

I. El problema consiste en usar el procedimiento de selección de variables para generar modelos y elegir el mejor, en el cual las variables son:

- Elevación promedio, en pies ( $x_1$ )
- Temperatura promedio en grados Fahrenheit ( $x_2$ )
- Precipitación media anual en pulgadas ( $x_3$ )
- Densidad de vegetación (porcentaje de cobertura) ( $x_4$ )
- Área de drenaje en millas<sup>2</sup> ( $x_5$ )
- Latitud en grados ( $x_6$ )
- Longitud en grados ( $x_7$ )
- Elevación de la estación que registra la temperatura en pies ( $x_8$ )
- Intensidad de la precipitación pluvial en pulgadas/hora ( $x_9$ ). Datos recolectados durante 25 años. ( $x_9$ )
- Rendimiento anual de agua en pulgadas ( $y$ )

a) Realice una prueba de bondad de ajuste Q-Q para verificar la normalidad de  $y$ .

b) Calcule matriz de correlaciones de las variables predictivas. ¿Qué puede concluir de los resultados obtenidos?

c) Ajuste el modelo de regresión lineal múltiple a las nueve variables predictivas.

d) Calcule un intervalo de 95% de confianza para los parámetros  $\beta_i$  del modelo con  $i = 1, 2$ .

e) Explique si el modelo es adecuado.

f) Determine la posible multicolinealidad entre las variables.

g) Con base en las respuestas anteriores, ¿qué puede concluir del modelo?

h) Proponga un modelo parsimonioso.