Introducción a la inferencia estadística

Armando Aguilar Márquez
Jorge Altamira Ibarra
Omar García León

PEARSON

Introducción a la inferencia estadística

Introducción a la inferencia estadística

Armando Aguilar Marquéz
Jorge Altamira Ibarra
Omar García León

Profesores del departamento de Matemáticas
Facultad de Estudios Superiores Cuautitlán
UNAM

Pearson Custom Publishing

México • Argentina • Brasil • Colombia • Costa Rica • Chile • Ecuador
España • Guatemala • Panamá • Perú • Puerto Rico • Uruguay • Venezuela

Datos de catalogación bibliográfica

Introducción a la inferencia estadística

PEARSON EDUCACIÓN, México, 2010

ISBN: 978-607-442-737-0

Formato: 20 × 25.5 cm Páginas: 216

Todos los derechos reservados.

Editor: Carlos Mario Ramirez Torres
e-mail: carlosmario.ramirez@pearsoned.com
Editor de desarrollo: Alejandro Agustín Gómez Ruiz
Supervisor de producción: Juan José García Guzmán

PRIMERA EDICIÓN, 2010

D.R. © 2010 por Pearson Educación de México, S.A. de C.V.
Atlacomulco 500-5º piso,
Col. Industrial Atoto
C.P. 53519, Naucalpan de Juárez, Estado de México

Cámara Nacional de la Industria Editorial Mexicana. Reg. número 1031.

PEARSON CUSTOM PUBLISHING es una marca registrada de Pearson Educación de México, S.A. de C.V.

Reservados todos los derechos. Ni la totalidad ni parte de esta publicación pueden reproducirse, registrarse o transmitirse, por un sistema de recuperación de información, en ninguna forma ni por ningún medio, sea electrónico, mecánico, fotoquímico, magnético o electroóptico, por fotocopia, grabación o cualquier otro, sin permiso previo por escrito del editor.

El préstamo, alquiler o cualquier otra forma de cesión de uso de este ejemplar requerirá también la autorización del editor o de sus representantes.

Impreso en México. *Printed in Mexico*.
1 2 3 4 5 6 7 8 9 0 - 13 12 11 10

Pearson Custom Publishing
es una marca de

PEARSON

www.pearsonededucacion.net

ISBN: 978-607-442-737-0

Contenido

Primera parte

Teoría

1

Introducción a la inferencia estadística

El objetivo de la inferencia estadística es construir estimaciones y pruebas de hipótesis acerca de las características de una población mediante la información contenida en una muestra. En este capítulo revisaremos cómo se pueden usar diferentes tipos de muestreo para obtener las muestras de las poblaciones en estudio.

Poblaciones y muestras

En estadística, por lo general confiamos en una *muestra* para realizar inferencias acerca del grupo de donde fue seleccionada. Al grupo mayor de datos se le denomina *población*.

EJEMPLO 1.1 Eres contratado por el Instituto Federal Electoral para que investigues la percepción que tienen los ciudadanos acerca de la honestidad de los procedimientos de elección en México. ¿A qué ciudadanos les preguntarías su percepción?

En el ejemplo 1.1 no resultaría práctico preguntarle a cada uno de los ciudadanos mexicanos su percepción acerca de esta cuestión. En lugar de esto, sería más fácil entrevistar a un número pequeño de ciudadanos y, a partir de sus respuestas, realizar inferencias acerca de lo que piensan todos los ciudadanos del país. Las personas a las que realmente se les preguntó constituyen nuestra *muestra*, la cual se seleccionó de la población formada por todos los ciudadanos de la República en edad de votar. Los procedimientos matemáticos mediante los cuales convertimos la información proporcionada por una muestra en suposiciones inteligentes acerca de la población es el campo de estudio de la inferencia estadística.

Una muestra es un subconjunto de la población. En el caso de la percepción ante los procedimientos de elección, seleccionaríamos una muestra de unos cuantos miles de mexicanos, de los millones que hay en el país en edad de votar. La selección de la muestra es crucial, ya que debe representar a la población, y no sólo a un segmento de ésta en detrimento de otro. Por ejemplo, un error sería seleccionar en la muestra sólo ciudadanos del estado de Michoacán. Si la muestra está constituida sólo por michoacanos, entonces no puede usarse para inferir la percepción de todos los votantes del país. El mismo problema se presentaría si en la muestra sólo hubiera miembros de un partido en particular (por ejemplo, panistas). La inferencia estadística se basa en la suposición de que la muestra es aleatoria. Para nuestro caso, tendremos confianza en una muestra que represente a los diferentes segmentos de la sociedad en las proporciones que más se aproximen a las reales (a condición que la muestra sea lo suficientemente grande).

EJEMPLO 1.2 Un profesor de inferencia estadística quiere saber las calificaciones que obtuvieron sus alumnos en estadística descriptiva. En la primera clase del curso les pregunta a los diez estudiantes sentados en la primera fila sus calificaciones en esa asignatura. Concluye, con base en las respuestas recibidas, que el grupo obtuvo muy buenas calificaciones. ¿Cuál es la muestra? ¿Cuál es la población? ¿Puedes identificar cualquier problema relacionado con la forma en que el profesor seleccionó la muestra?

En el ejemplo 1.2, la población está formada por todos los estudiantes de la clase. Y la muestra se conformó con los diez estudiantes sentados al frente. Esta muestra probablemente no es representativa por la tendencia que existe a que los más aplicados se sienten al frente y éstos alcancen las más altas calificaciones. Por tanto, la muestra puede proporcionar una calificación promedio más alta que la que realmente le corresponde al grupo.

EJEMPLO 1.3 Un profesor de deportes está interesado en determinar el rendimiento promedio de los estudiantes en una carrera con obstáculos. Ocho estudiantes de su clase se apuntan como voluntarios. Después de observar su desempeño, el profesor concluye que sus estudiantes pueden realizar exitosamente la prueba.

En el ejemplo 1.3, la población son todos los estudiantes del grupo. La muestra se conformó con ocho voluntarios. La selección de la muestra fue deficiente porque los voluntarios son probablemente más hábiles en realizar la prueba que el resto de los estudiantes. Los estudiantes sin habilidad casi con seguridad no se anotaron como voluntarios. Además, nada se dice del género de los voluntarios. Por ejemplo, ¿cuántas voluntarias hubo? Esto puede afectar el resultado, adicionalmente al hecho de que la muestra no es representativa.

Muestreo aleatorio simple

Los investigadores adoptan una variedad de estrategias de muestreo. El más sencillo es el *muestreo aleatorio simple*. Este tipo de muestreo requiere que cada elemento de la población tenga la misma posibilidad de ser seleccionado en la muestra. Además, la selección de un elemento debe ser independiente de la selección de cualquier otro. Es decir, seleccionar un elemento de la población no debe aumentar o disminuir la probabilidad de escoger a cualquier otro. En este sentido, se puede decir que en el muestreo aleatorio simple se selecciona la muestra totalmente al azar. Para comprobar tu comprensión acerca del muestreo aleatorio simple, considera el siguiente ejemplo. ¿Cuál es la población? ¿Cuál es la muestra? ¿La muestra fue seleccionada mediante un muestreo aleatorio simple? ¿La muestra es sesgada?

EJEMPLO 1.4 Una investigadora está interesada en estudiar las experiencias de adolescentes hijos de familia en comparación con los hijos de padres divorciados. Obtiene una lista de adolescentes de 16 años de edad, de las oficinas del Registro Civil de la ciudad de Monterrey y selecciona dos subconjuntos de individuos para su estudio. Primero, elige a todos aquellos que su apellido empieza por *Z*. Luego, a todos los adolescentes cuyo apellido empieza con *A*. Debido a que muchos registros muestran apellidos que empiezan con *A*, selecciona a uno sí y a otro no, y así sucesivamente. Por último, envía una encuesta por correo a las personas seleccionadas y compara las características de los adolescentes hijos de familia con los hijos de padres divorciados.

En el ejemplo 1.4, la población está formada por todos los adolescentes de 16 años de edad, registrados civilmente en la ciudad de Monterrey. Es importante que la investigadora realice sólo generalizaciones sobre los adolescentes de esta lista, no sobre todos los del país o del mundo. Esto quiere decir que las listas obtenidas en las oficinas del Registro Civil de Monterrey no pueden representar a todos los adolescentes de esta edad. Aunque las inferencias se limiten a esta lista, se observan varias deficiencias en el procedimiento de muestreo descrito. Por ejemplo, seleccionar a aquellos jóvenes cuyo apellido empieza con *Z* no da igual probabilidad a cada individuo de ser seleccionado. Además, con este procedimiento se corre el riesgo de que una oficina quede sobrerrepresentada, ya que en algunas de éstas puede ser muy común un apellido que inicia con *Z*. Hay otras razones por las que

seleccionar sólo apellidos que empiezan con *Z* puede dar como resultado una muestra sesgada. Tal vez esas personas suelen ser más pacientes en promedio, debido a que están acostumbradas a ser nombradas al final de una lista. También pueden observarse deficiencias al seleccionar a las personas con apellidos que inician con *A*. Una deficiencia adicional para estos últimos es que el procedimiento de "uno sí, uno no", rechaza personas por el simple hecho de estar junto a las personas seleccionadas. Sólo esta deficiencia nos dice que la muestra no fue seleccionada de acuerdo con un muestreo aleatorio simple.

Tamaño de la muestra

Recuerda que la definición de una muestra seleccionada al azar es aquella en la cual cada elemento de la población tiene la misma probabilidad de ser seleccionado. Esto significa que el procedimiento de muestreo, más que los resultados obtenidos a partir de la muestra, definen lo que es necesario para obtener una muestra al azar. Las muestras escogidas al azar, especialmente si el tamaño de la muestra es pequeño, no son representativas de la población. Por ejemplo, si una muestra escogida al azar de 20 elementos fuera seleccionada de una población con un número igual de chicos y chicas, hay una probabilidad de 0.06 de que el 70% o más de las personas que forman la muestra sean mujeres (recuerda que estos resultados se calculan mediante la distribución binomial). Entonces, la muestra no sería representativa, aunque fuera seleccionada aleatoriamente. Sólo una muestra grande es probable que sea representativa de la población. Por esta razón, la inferencia estadística toma en cuenta el tamaño de muestra cuando se generalizan los resultados encontrados en las muestras. En próximos capítulos revisaremos qué clases de técnicas matemáticas aseguran la sensibilidad del tamaño de la muestra.

Muestreos más sofisticados

Algunas veces no es posible hacer un muestreo utilizando un muestreo aleatorio simple. Supón que las ciudades de Durango y Gómez Palacio compiten para que se construya en ellas un Sistema de Centros de Rehabilitación Infantil Teletón. Imagina que te contratan para evaluar si la mayoría de los duranguenses prefieren Durango o Gómez Palacio como ciudad seleccionada. Dado lo impráctico de obtener la opinión de todos y cada uno de los duranguenses, debes determinar una muestra de esta población. Pero en este momento te das cuenta de la dificultad de realizar un muestreo aleatorio simple. Por ejemplo, si te basas en el Registro Federal de Electores, cómo vas a contactar a aquellos individuos que no votan; si es con base en los directorios telefónicos, cómo vas a localizar a los que no tienen teléfono. Incluso entre las personas que aparecen en los directorios telefónicos, ¿cómo puedes identificar a los que acaban de mudarse del estado (y no tenían ninguna razón para informar de su cambio de residencia)? ¿Qué haces con el hecho de que desde el inicio del estudio otras personas han establecido su residencia en el estado? Como puedes ver, algunas veces es muy difícil desarrollar un procedimiento aleatorio. Por esta razón se han desarrollado otras clases de técnicas de muestreo. Discutamos sólo dos de ellas.

Asignación aleatoria

En la investigación experimental, las poblaciones por lo general son hipotéticas. Por ejemplo, para comparar la efectividad de un nuevo analgésico que se planea introducir en el mercado, no existe una población real de individuos que tomen la medicina. En este caso, la población se define con gente con algún grado de dolor y se toma una muestra aleatoria de esta población. La muestra en forma aleatoria

se divide en dos grupos; a un grupo se le asigna la condición de tratamiento (analgésico) y al otro grupo el control (placebo). La división aleatoria en dos grupos se denomina asignación aleatoria. La asignación aleatoria es crítica para la validez de un experimento. Por ejemplo, considera el sesgo que se produce si los primeros 20 individuos seleccionados se asignan al grupo experimental y los siguientes 20 al grupo control. Es posible que los sujetos que se seleccionaron al final tiendan a tener mayor nivel de dolor que los primeros, sin ninguna otra razón más que el azar. En este caso, el resultado sería que el grupo experimental tendrá menor dolor que el grupo control, aun antes de que se administre la medicina.

En una investigación experimental de este tipo, la falta de asignación aleatoria de los sujetos a los grupos es generalmente más grave que tener una muestra no aleatoria. La falta de aleatorización invalida los resultados experimentales; las muestras no aleatorias sólo restringen la generalización de los resultados.

Muestreo estratificado

Ya que el muestreo aleatorio simple por lo general no asegura una muestra representativa, un método de muestreo conocido como muestreo aleatorio estratificado se usa algunas veces para lograr que la muestra represente mejor a la población. El método puede ser utilizado siempre y cuando la población tenga distintos estratos o grupos. En el muestreo estratificado necesitas identificar a los miembros de la muestra, que pertenecen a cada grupo, entonces muestreas aleatoriamente cada uno de estos subgrupos, de tal manera que los tamaños de éstos, en la muestra, sean proporcionales a sus tamaños en la población.

Veamos un ejemplo: supón que estás interesado en la opinión de los estudiantes de la FESC, acerca del servicio de transporte. Tienes el tiempo y los recursos para entrevistar a 200 estudiantes. Éstos se pueden dividir respecto al turno. Es posible que los estudiantes de la tarde tengan un punto de vista muy diferente sobre el servicio de transporte. Si el 70% de los estudiantes son del turno matutino, tiene sentido asegurarse que el 70% de la muestra esté constituida por estudiantes de ese turno. Entonces, tu muestra de 200 estudiantes tendrá 140 estudiantes del turno matutino y 60 estudiantes del turno vespertino.

La proporción de estudiantes de los dos turnos en la muestra y en la población (total de estudiantes en la universidad) sería igual.

PREGUNTAS

1. Nuestros datos proceden de _____ pero nosotros realmente queremos estudiar _____

a) teorías y modelos matemáticos.

b) muestras y poblaciones.

c) poblaciones y muestras.

d) métodos subjetivos; métodos objetivos.

2. Una muestra aleatoria:

a) es la muestra que tiene mayor probabilidad de ser representativa de la población.

b) es siempre representativa de la población.

c) permite que puedan calcularse directamente los parámetros de la población.

d) todo lo anterior es verdadero.

e) todo lo anterior es falso.

3. Cuando algún participante en un estudio de investigación se integra a un grupo de tratamiento al azar:

a) se lleva a cabo un muestreo aleatorio.

b) se lleva a cabo una asignación aleatoria.

c) las conclusiones estadísticas serían absolutamente correctas.

d) los resultados de la investigación pueden estar comprometidos porque nunca debes asignar aleatoriamente los individuos a los grupos.

4. La incertidumbre respecto a las conclusiones sobre la población se pueden eliminar si:

a) usas una muestra aleatoria grande.

b) obtienes datos de todos los miembros de la población.

c) depende de la distribución *t* de student.

d) *a* y *b*.

5. ¿Cuál de las siguientes aseveraciones es verdadera?, usando muestreo aleatorio:

a) se acepta alguna incertidumbre acerca de las conclusiones.

b) hay condiciones para realizar cálculos estadísticos.

c) hay riesgo de realizar conclusiones erróneas acerca de la población.

d) los resultados son sesgados.

6. Una muestra aleatoria es una:

a) muestra que es casual.

b) muestra que no es planeada.

c) es aquella en la que cada muestra de un tamaño particular tiene la misma probabilidad de ser seleccionada.

d) es aquella que asegura que no habrá incertidumbre en las conclusiones.

7. ¿Cuál de las siguientes es una muestra aleator?

a) cada 5 personas que entran a la FESC campo IV entre las 8:30 y 10:00 horas de la mañana.

b) Juana López, Daniel Cantera y Luis Pérez, cuyos números de cuenta fueron seleccionados mediante una rifa.

c) cada 20 personas de un directorio estudiantil.

d) todos los anteriores.

8. Una muestra sesgada es aquella que:

a) es muy pequeña.

b) siempre conduce a conclusiones erróneas.

c) seguramente tiene grupos de la población sobrerrepresentados o subrepresentados debido a factores aleatorios.

d) seguramente tiene grupos de la población sobrerrepresentados o subrepresentados debido a factores de muestreo.

e) es siempre una muestra buena y útil.

2 Distribuciones muestrales

El concepto de una distribución muestral es quizás el concepto más importante en la inferencia estadística. También es un concepto difícil de enseñar debido al hecho de que una distribución muestral, más que una distribución empírica, es una distribución teórica.

En la introducción se define el concepto de distribución muestral y se desarrolla un ejemplo para una distribución continua y una distribución para la proporción, también se discute el uso de las distribuciones muestrales en la inferencia estadística.

Las otras secciones de este capítulo se refieren a la importancia de las distribuciones muestrales: la distribución muestral de la media, la distribución muestral de la diferencia de medias, la distribución muestral del coeficiente de correlación y la distribución muestral de una proporción.

Imagina que obtienes una muestra aleatoria de diez personas de la población de mujeres de la FES-Cuautitlán y calculas el promedio de estatura de la muestra obtenida. No esperarás que la media de tu muestra sea igual a la media de estatura de todas las mujeres de la FESC. El valor obtenido en tu muestra es algo más grande o algo más pequeño, pero es muy probable que no será exactamente igual al promedio de estatura de la población. Supón que tomas una segunda muestra de diez chicas de la misma población, no esperarás que la media de esta segunda muestra sea igual a la media de la primera.

Ahora supongamos que tomamos muchas muestras de diez muchachas cada una y calculamos la media de cada una de las muestras. En el estudio de la inferencia estadística es de suma importancia determinar la variabilidad de las medias de las muestras y qué tan lejos se encuentran ellas de la estatura promedio de la población, es decir del parámetro. Para determinar esta variabilidad es necesario utilizar las distribuciones muestrales.

Distribuciones discretas

Ilustremos el concepto de distribución muestral con un ejemplo sencillo. Se tienen tres empleados con 1, 2 y 3 años de antigüedad (considera años completos, a fin de considerar la variable como discreta). Selecciona dos empleados aleatoriamente con reemplazo y calcula el promedio de sus antigüedades. Todos los posibles resultados se muestran en la tabla 2.1.

Resultado	Antigüedad del primer empleado	Antigüedad del segundo empleado	Media
1	1	1	1.0
2	1	2	1.5
3	1	3	2.0
4	2	1	1.5
5	2	2	2.0
6	2	3	2.5
7	3	1	2.0
8	3	2	2.5
9	3	3	3.0

Tabla 2.1 Resultados posibles cuando se seleccionan muestras de las antigüedades de dos empleados.

Observa que los valores posibles de la media son 1, 1.5, 2.0, 2.5 y 3.0. Las frecuencias de estas medias se presentan en la tabla 2.2. Las frecuencias relativas se calculan dividiendo las frecuencias entre 9, ya que son 9 los resultados posibles.

Media	Frecuencia	Frecuencia relativa
1.0	1.0	0.1111
1.5	2.0	0.2222
2.0	3.0	0.3333
2.5	2.0	0.2222
3.0	1.0	0.1111

Tabla 2.2 Frecuencias de las medias para $n = 2$.

En la figura 2.1 se muestra la distribución de las frecuencias relativas de las medias correspondientes a la tabla 2.2. Esta distribución es también de probabilidad, ya que las frecuencias relativas son una medida de la probabilidad de obtener una media dada en una muestra de dos personas.

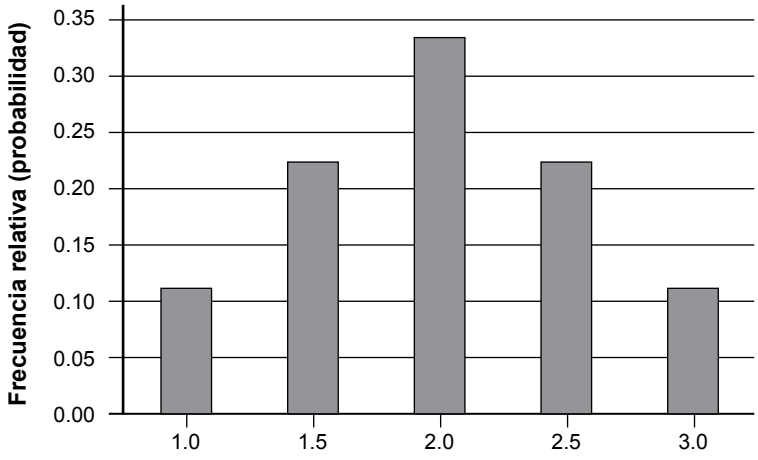


Figura 2.1 Distribución de las frecuencias relativas de las medias correspondientes a la tabla 2.2.

A la distribución que se muestra en la figura 2.1 se le denomina distribución muestral de la media. Para este ejemplo sencillo, la distribución de las antigüedades y la distribución de muestreo son discretas. Las antigüedades únicamente pueden ser de 1, 2 y 3 años, y la media de la muestra puede tomar sólo uno de cinco resultados posibles.

Hay una forma alternativa de conceptualizar la distribución muestral que puede ser útil para distribuciones más complejas. Imagina que se seleccionan dos empleados (con reemplazo) y se calcula la media de las antigüedades de las dos personas: este proceso se repite para una segunda muestra, una tercera muestra, y para miles de muestras. Después de seleccionar miles de muestras, se calculan las medias para cada una de ellas, y se construye una distribución de frecuencias relativas. Cuantas más muestras se hayan tomado, la distribución de frecuencias relativas se aproxima más a una distribución muestral. Esto significa que se puede conceptualizar una distribución muestral como una distribución de frecuencias construida con un número muy grande de muestras, aunque lo estrictamente correcto es que la distribución muestral sea igual a la distribución de frecuencias sólo cuando hay un número infinito de muestras.

Resultado	Antigüedad del primer empleado	Antigüedad del segundo empleado	Rango
1	1	1	0
2	1	2	1
3	1	3	2
4	2	1	1
5	2	2	0
6	2	3	1
7	3	1	2
8	3	2	1
9	3	3	0

Tabla 2.3 Resultados posibles cuando se seleccionan dos empleados.

Es importante tener presente que para cada estadístico, no sólo para la media, hay una distribución muestral. Por ejemplo, en la tabla 2.3 se muestran todos los posibles resultados para el rango, para el ejemplo de las antigüedades de los empleados. La tabla 2.4 muestra las frecuencias y las frecuencias relativas para cada valor posible de los rangos y en la figura 2.2 se muestra la distribución muestral para el rango.

Rango	Frecuencia	Frecuencia relativa
0	3	0.3333
1	4	0.4444
2	2	0.2222

Tabla 2.4 Frecuencias de los rangos para $n = 2$.

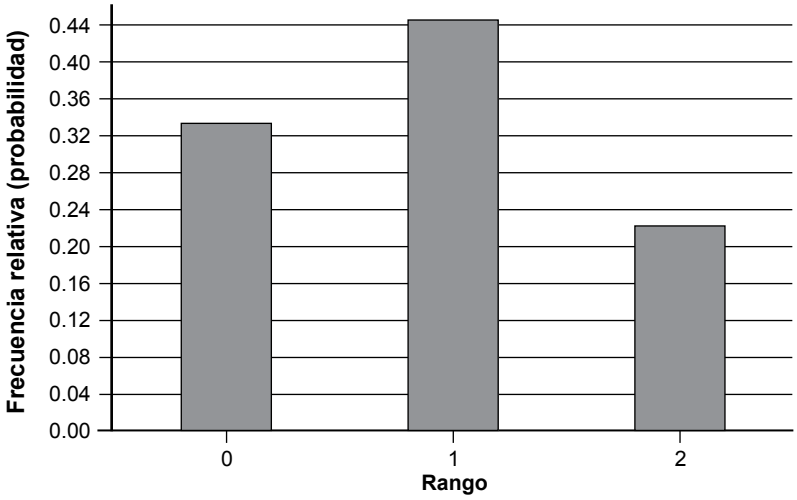


Figura 2.2 Distribución muestral para el rango.

Es también importante tener en cuenta que hay una distribución muestral para cada tamaño de muestra.

Distribuciones continuas

En la sección anterior, la población estaba formada por las antigüedades de tres empleados. Considera ahora una distribución muestral cuando la distribución de la población es continua. ¿Qué pasaría si tuviéramos datos de mil esferas que están numeradas del 0001 al 1 000 en incrementos iguales? (Aunque la distribución no es realmente continua, se aproxima lo suficiente y, por tanto, se puede considerar para fines prácticos como una distribución continua). Así, como en el ejemplo anterior, estamos interesados en la distribución de las medias que se obtendrían si seleccionamos dos esferas y calculamos la media. En el ejemplo anterior, empezamos por calcular la media para cada uno de los resultados posibles. En este caso no sería práctico, porque el número de resultados posibles es de 1 000 000 (1 000 de la primera esfera $\times$ 1 000 para la segunda). Por tanto, es más conveniente usar nuestra segunda conceptualización de las distribuciones muestrales, que concibe la distribución muestral en términos de una distribución de frecuencias relativas. Específica-

mente, la distribución de frecuencias relativas que resultaría si se tomaran muestras de tamaño dos, repetidamente y calculamos la media para cada una de ellas.

Cuando tenemos una verdadera distribución continua, no sólo es impráctico, sino que es imposible enumerar todos los resultados posibles. Por otro lado, recuerda que en las distribuciones continuas la probabilidad de obtener un valor es cero.

Distribuciones muestrales e inferencia estadística

Al inicio del capítulo se mencionó que las distribuciones muestrales son muy importantes en el estudio de la inferencia estadística. En los ejemplos desarrollados anteriormente, se especificó una población y se obtuvo la distribución muestral de la media y el rango. En la práctica, el procedimiento se realiza de otra forma: obtienes los datos de una sola muestra y, a partir de estos datos, estimas los parámetros de la distribución muestral. El conocimiento de la distribución muestral es de suma utilidad. Por ejemplo, conocer el grado en el que las medias de diferentes muestras difieren unas de otras, y el grado en que difieren o se alejan de la media de la población, te permite detectar qué tan probable es que la media de tu muestra particular esté cerca de la media de la población. Por fortuna, esta información se consigue directamente con la distribución muestral. La medida más común de cuánto difieren la media una de otra es la desviación estándar de la distribución muestral de la media. Esta desviación estándar se conoce como error estándar de la media. Si todas las medias de las muestras están muy cercanas a la media de la población, entonces el error estándar de la media es muy pequeño. Por otro lado, si las medias de las muestras varían considerablemente, entonces el error estándar de la media es muy grande.

Para ser específicos, supón que la media de tu muestra fue 125 y estimas que el error estándar de la media fue 5 (utilizando un método indicado en una sección posterior). Si tienes una distribución normal, es probable que tu media se aleje 10 unidades a partir de la media de la población (en cualquier dirección), ya que la densidad de la distribución normal se encuentra entre ± 2 desviaciones estándar a partir de la media.

PREGUNTAS

1. Una distribución muestral es igual a una distribución de frecuencias, cuando:

a) la distribución es discreta.

b) la distribución es continua.

c) hay al menos veinte muestras.

d) existe un número infinito de muestras.

e) es la distribución muestral de la media.

2. Selecciona todo lo que aplica. ¿Cuál de estos estadísticos tiene una distribución muestral?

a) media.

b) mediana.

c) rango.

d) desviación estándar.

e) coeficiente de correlación de Pearson (r).

3. ¿Cuál es el error estándar de la media?

a) la desviación estándar de la distribución muestral de la media.

b) la desviación estándar de la distribución normal estándar.

c) la variación entre la media y la suma de los datos de la población.

d) la diferencia entre la media de una primera muestra y la media de una segunda muestra.

Distribución muestral de la media

La distribución muestral de la media se definió en la introducción de este capítulo. En esta sección revisaremos algunas propiedades importantes de la distribución muestral de la media que debes haber conocido en la sección del laboratorio correspondiente a este capítulo.

Media

La media de la distribución muestral de la media es igual a la media de la población de la cual se obtuvieron las muestras. Por tanto, si la población tiene una media, μ, entonces la distribución muestral de la media es también μ. El símbolo $\mu_{\bar{x}}$ se utiliza para referirse a la media de la distribución muestral de la media. Entonces, la fórmula para la media de la distribución muestral de la media se puede expresar de la siguiente manera:

$$\mu = \mu_{\bar{x}}$$

Varianza

La varianza de la distribución muestral de la media se calcula con la siguiente fórmula:

$$\sigma_{\bar{x}}^2 = \frac{\sigma}{n}$$

Esto es, la varianza de la distribución muestral de la media es la varianza de la población dividida entre n, el tamaño de la muestra (el número de datos que se utilizaron para calcular la media). Entonces, cuanto mayor sea el tamaño de la muestra, la varianza de la distribución muestral de la media será menor.

Esta expresión puede derivarse fácilmente a partir de la Ley de la suma de las varianzas. La varianza de la distribución muestral de la suma de tres números obtenidos por muestreo de la población con varianza σ^2, sería $\sigma^2 + \sigma^2 + \sigma^2$. En forma general, para n números la varianza sería $n\sigma^2$. Ya que la media es la suma por $1/n$, la varianza de la distribución muestral de la media sería la varianza de la suma por $1/n^2$, lo cual es igual a σ^2/n.

El error estándar de la media es la desviación estándar de la distribución muestral de la media. Es, por tanto, la raíz cuadrada de la varianza de la distribución muestral de la media y se puede escribir como:

$$\sigma_{\bar{x}} = \frac{\sigma^2}{\sqrt{n}}$$

El error estándar se representa con σ porque es una desviación estándar. El subíndice ($\bar{x}$) indica que el error estándar es el error estándar de la media.

Teorema del límite central

Las declaraciones del Teorema del límite central son: *Dada una población con una media finita μ, y una varianza finita diferente de cero σ^2, la distribución muestral de la media se aproxima a una distribución normal con una media μ y una varianza σ^2/n a medida que n, el tamaño de la muestra, aumenta.*

Las expresiones de la media y varianza de la distribución muestral de la media no son nuevas, tampoco lo que queremos remarcar en este momento del teorema. En lo que debemos poner atención es en que, sin importar la forma de la distribución de la población, la distribución muestral de la media se aproxima a una distribución normal a medida que n, el tamaño de la muestra, aumenta. Observa

la figura 2.3 y los resultados para $n = 2$ y $n = 10$. La distribución de la población es *uniforme*. Puedes ver que la distribución para $n = 2$ está muy lejos de ser una distribución normal. Sin embargo, observamos que hay más densidad de datos en la parte media en comparación con la densidad en los extremos.

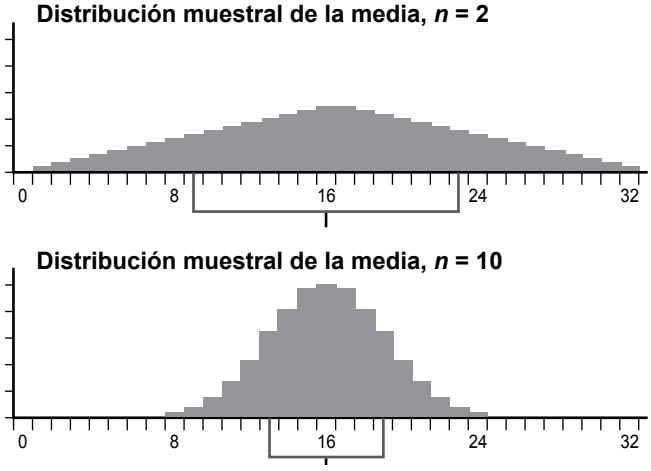


Figura 2.3 Una simulación de una distribución muestral.

Para $n = 10$, la distribución se aproxima más a una distribución normal. Observa que las medias de las dos distribuciones son iguales, pero que la dispersión de la distribución para $n = 10$ es más pequeña.

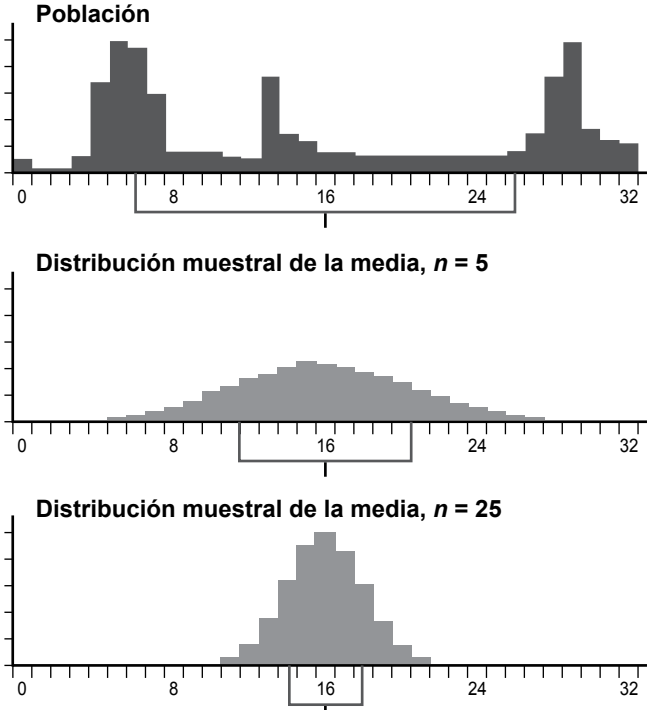


Figura 2.4 Una simulación de una distribución muestral. La población es muy diferente de la normal.

La figura 2.4 muestra lo cerca que se encuentra la distribución muestral de la media de una distribución normal, aun cuando la distribución de la población no es normal. Si observas con cuidado las gráficas, verás que las distribuciones muestrales tienen un ligero sesgo positivo. Cuanto más grande sea la muestra, la distribución muestral de la media se aproximará más a una distribución normal.

Distribución muestral de la diferencia entre medias

Los análisis estadísticos muy frecuentemente se refieren a diferencias entre medias. Un ejemplo típico es un experimento diseñado para comparar la media de un grupo control con la media de un grupo experimental. La estadística inferencial que se utiliza en el análisis de este tipo de experimentos depende de la distribución muestral de la diferencia de medias.

La distribución muestral de la diferencia de medias puede conceptualizarse como la distribución que resultaría si repetimos los siguientes tres pasos una y otra vez: (1) obtener una muestra de tamaño n_1 de la población 1 y obtener una muestra de tamaño n_2 de la población 2; (2) calcular las medias de las dos muestras ($\bar{x}_1$ y $\bar{x}_2$), y (3) calcular la diferencia entre la medias $\bar{x}_1 - \bar{x}_2$. La distribución de las diferencias entre las medias es la distribución muestral de la diferencia de medias.

Se puede esperar que la media de la distribución muestral de la diferencia de medias sea:

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$$

Lo cual indica que la media de la distribución muestral de la diferencia de medias es igual a la diferencia entre las medias de las poblaciones. Por ejemplo, supongamos que la media de calificaciones de todos los estudiantes de primer semestre de la carrera de la licenciatura en Administración es 8, y la de los estudiantes de segundo semestre es 7, entonces si obtenemos numerosas muestras de cada grupo de estudiantes y calculamos las diferencias de las muestras cada vez, la media de estas numerosas diferencias de medias muestrales debería ser $8 - 7 = 1$.

De acuerdo con la Ley de la suma de las varianzas, sabemos que:

$$\sigma^2_{\bar{x}_1 - \bar{x}_2} = \sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2}$$

Esta fórmula expresa que la varianza de la distribución muestral de la diferencia de medias es igual a la varianza de la distribución muestral de la media para la población 1, más la varianza de la distribución muestral de la media para la población 2. Sabemos que la varianza de la distribución muestral de la media está dada por:

$$\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$$

Debido a que se tienen dos poblaciones, es necesario distinguir sus varianzas y sus tamaños de muestra. Usaremos los subíndices 1 y 2. Usando este convencionalismo podemos escribir la fórmula para la varianza de la diferencia de medias como:

$$\sigma^2_{\bar{x}_1 - \bar{x}_2} = \sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2}$$

Ya que el error estándar de la distribución muestral es la desviación estándar de la misma, entonces el error estándar de la diferencia de medias viene dado por:

$$\sigma^2_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2}}$$

A la izquierda tenemos a sigma (σ), lo que significa que es una desviación estándar. Los subíndices $\bar{x}_1 - \bar{x}_2$ indican que es la desviación estándar de la distribución muestral de $\bar{x}_1 - \bar{x}_2$

Veamos una aplicación de esta fórmula. Supón que existen dos especies de plantas. La altura promedio de la especie 1 es 32 y la de la especie 2 es 22. Las varianzas para las dos especies son 60 y 70, respectivamente, y las alturas de ambas especies se distribuyen en forma normal. Seleccionas una muestra de 10 plantas de la especie 1 y una muestra de 14 plantas de la especie 2: ¿cuál es la probabilidad de que la media de la muestra de la especie 1 exceda a la de la dos en 5 o más unidades de altura? Sin necesidad de hacer ningún cálculo, tal vez te diste cuenta de que esta probabilidad es alta, debido a que la diferencia entre las medias de la población es de 10 unidades. Pero, ¿cuál es exactamente esta probabilidad?

Primero determinemos la distribución muestral de la diferencia de medias. Utilizando las fórmulas anteriores, tenemos que la media es:

$$\mu_{\bar{x}_1-\bar{x}_2} = 32 - 22 = 10$$

El error estándar es:

$$\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{60}{10} + \frac{70}{14}} = 3.317$$

En la figura 2.5 se puede ver la distribución muestral. Observa que se distribuye en forma normal con una media de 10 y una desviación estándar de 3.317. El área mayor a 5 se encuentra sombreada.

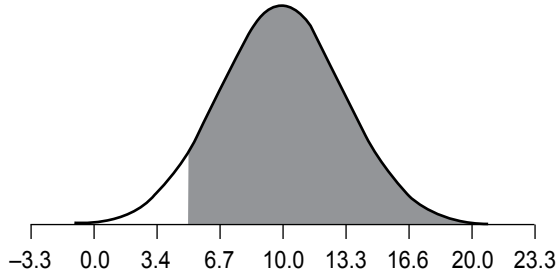


Figura 2.5 Distribución muestral de la diferencia de medias.

El último paso será determinar el área "sombreada". Usando ya sea una tabla de *z* o la calculadora convencional, se determina que esta área es 0.934. Ésta es la probabilidad de que la media de la muestra de la especie 2 sea mayor que la media de la muestra de la especie 1, en 5 unidades o más.

La fórmula para el error estándar de la diferencia de medias se simplifica si los tamaños de las muestras y las varianzas de las poblaciones son iguales. Ya que por un lado las varianzas de las poblaciones son iguales y por otro lado los tamaños de las muestras son los mismos, no es necesario utilizar subíndices para diferenciar estos términos. Entonces:

$$\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}} = \sqrt{\frac{2\sigma^2}{n}}$$

Esta versión simplificada de la fórmula la usaremos para resolver el siguiente problema: en una determinada población, la estatura promedio de los varones de 15 años (en cm) es 175 y la varianza es 64. Para las mujeres de 15 años, la media es 165 y la varianza es 64. Si se toma una muestra de 8 varones y de 8 chicas, ¿cuál es la probabilidad de que la altura media de la muestra de las chicas sea mayor que la de los chicos? En otras palabras, ¿cuál es la probabilidad de que la altura media de las chicas menos la altura media de los chicos sea mayor que 0?

Así como en el ejemplo anterior, el problema se resuelve en términos de la distribución muestral de la diferencia de medias (mujeres−varones). La media de la distribución es 165 − 175 = −10. El error estándar de la diferencia de medias es:

$$\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{2\sigma^2}{n}} = \sqrt{\frac{(2)(64)}{8}} = 4$$

En la figura 2.6 se muestra la gráfica de esta distribución. Es claro que es poco probable que la estatura promedio de las mujeres sea mayor que la de los varones, ya que la población de varones tiene mayor estatura. No obstante, no es imposible que la media muestral de las mujeres pueda ser mayor que la media muestral de los varones.

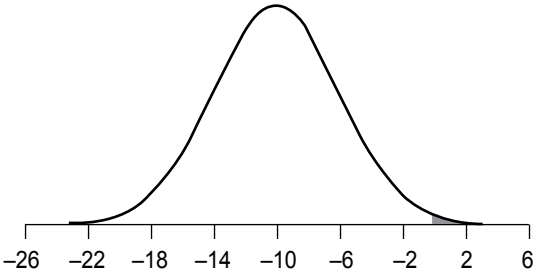


Figura 2.6 Distribución muestral de la diferencia de medias de las estaturas.

Una diferencia entre las medias de 0 o más es una diferencia de 10/4 = 2.5 desviaciones estándar por arriba de la media. La probabilidad de estar en el área delimitada por 2.5 o más desviaciones estándar por arriba de la media es 0.0062.

PREGUNTAS

1. La población 1 tiene una media de 20 y una varianza de 100; la población 2 tiene una media de 15 y una varianza de 64. Se seleccionan 20 elementos de la población 1, y 16 de la población 2. ¿Cuál es la media de la distribución muestral de la diferencia entre medias?

Respuesta _____

2. La población 1 tiene una media de 20 y una varianza de 100; la población 2 tiene una media de 15 y una varianza de 64. Se seleccionan 20 elementos de la población 1, y 16 de la población 2. ¿Cuál es la varianza de la distribución muestral de la diferencia entre medias?

Respuesta _____

3. La estatura media de los adolescentes varones de 15 años es de 175 cm con una varianza de 64. Para las chicas de la misma edad, la estatura media es de 165 cm con una varianza de 64. Si se seleccionan 8 chicas y 8 chicos, ¿cuál es la probabilidad de que la estatura promedio de la muestra de los chicos sea al menos 6 cm mas que la media en la muestra de las chicas?

Respuesta _____

4. El puntaje promedio en un examen de la asignatura "A" es de 32 puntos, mientras que para la asignatura "B" es de 28 puntos. Las varianzas son 60 y 50, respectivamente, y el puntaje para ambos exámenes se distribuye en forma normal. Se seleccionan al azar 14 exámenes de la asignatura "A" y 12 de la asignatura "B". ¿Cuál es la probabilidad de que la media de la muestra de los exámenes de la asignatura "B" sea mayor que la media de los exámenes de la asignatura "A" por 2 puntos o más?

Respuesta _____

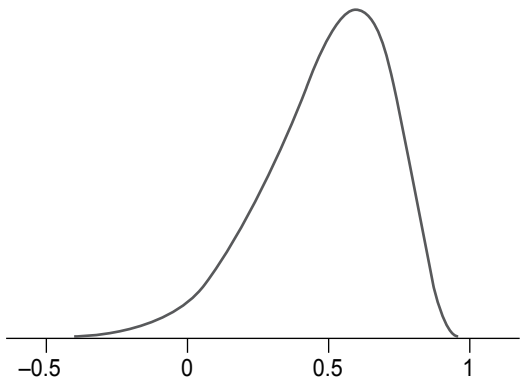


Figura 2.7 La distribución muestral de r para $n = 12$ y $\rho = 0.60$.

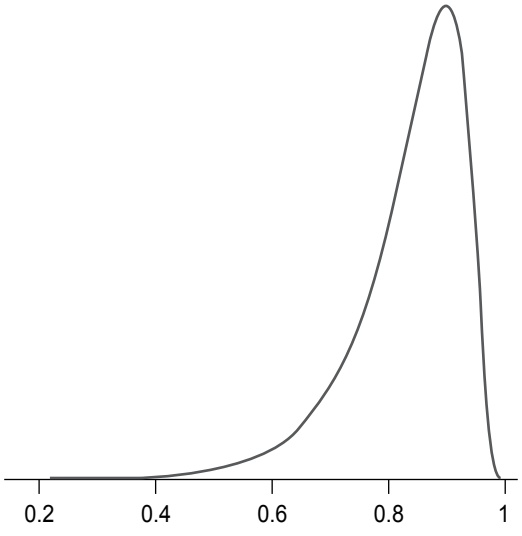


Figura 2.8 La distribución muestral de r para $n = 12$ y $\rho = 0.90$.

Distribución muestral del coeficiente de correlación de Pearson, *r*

Supón que la correlación entre un examen verbal y uno escrito en una población de estudiantes dada es 0.60. En otras palabras, $\rho = 0.60$. Si se seleccionan aleatoriamente a 12 estudiantes, el coeficiente de correlación de la muestra no será exactamente igual a 0.60. Por supuesto, diferentes muestras de 12 estudiantes proporcionarán diferentes valores de r. La distribución de los valores de r, después de tomar muestras repetidas de 12 estudiantes cada una, es la distribución muestral de r.

En la figura 2.7 se muestra la forma de la distribución muestral de r para nuestro ejemplo. Puedes observar que la distribución muestral no es simétrica. Presenta un sesgo negativo. La razón de este sesgo es debido a que r no puede tomar valores mayores a 1 y, por tanto, se extenderá de 0.6 a 1 en el lado positivo y de 0.6 a –1 en el negativo. Cuanto más grande sea el valor de ρ, más pronunciado será el sesgo.

En la figura 2.8 se muestra la distribución muestral para $\rho = 0.90$. Tiene una cola corta del lado positivo y larga en el negativo.

Para nuestro ejemplo, supón que quieres conocer la probabilidad de que en una muestra de 12 estudiantes, el valor muestral de r sea igual o mayor a 0.75. Piensas que todo lo que necesitas para calcular esta probabilidad es la media y el error estándar de la distribución muestral de r. Sin embargo, debido a que la distribución muestral no es normal, no estamos en condiciones de resolver el problema.

Fisher desarrolló una forma de transformar r en una variable distribuida normalmente con un error estándar conocido. A la nueva variable se le llama z', y la transformación se realiza de acuerdo con la siguiente fórmula:

$$z' = 0.5 \ln[(1 + r)/(1 - r)]$$

Para este nivel, los detalles y la derivación de la fórmula no son importantes; lo que es importante es tener en cuenta que z' se distribuye en forma normal y tiene un error estándar dado por:

$$\frac{1}{\sqrt{n-3}}$$

donde n es el número de pares de observaciones.

Para realizar la transformación de r en z' puedes usar tablas o la calculadora. Regresemos al problema de calcular la probabilidad de obtener un coeficiente de correlación muestral r igual o mayor a 0.75 en una muestra de tamaño 12, seleccionada de una población con un coeficiente de correlación de 0.60. El primer paso es convertir tanto 0.60, como 0.75 a valores de z'. Los valores son 0.693 y 0.973, respectivamente. El error estándar de z' para $n = 12$ es 0.333. Por tanto, el problema se reduce a calcular la probabilidad de obtener un valor igual o mayor a 0.973, en una distribución normal con media 0.693 y desviación estándar 0.333. La respuesta la puedes obtener directamente usando la calculadora de área dado un valor X. Usando la fórmula:

$$z = (X - \mu)/\sigma = (0.9730 - 0.693)/0.333 = 0.841$$

Utilizando la tabla de la distribución normal se determina que el área buscada ($X > 0.841$) es 0.20.

PREGUNTAS

1. ¿Cuál es la forma de la distribución muestral de r?

a) bimodal.

b) normal.

c) sesgada.

2. ¿Cuál es la z' que corresponde a un $r = -0.65$?

a) Respuesta _____.

3. ¿Cuál de los siguientes valores de r difiere más de su correspondiente z'?

a) $r = 0$.

b) $r = 0.6$.

c) $r = 0.2$.

d) $r = -0.8$.

4. La población tiene un coeficiente de correlación de 0.6, ¿cuál es la probabilidad de obtener en la muestra un coeficiente de correlación de por lo menos 0.5, si la muestra fue de 19 pares?

a) Respuesta _____

Distribución muestral de la proporción

Supón que en una elección entre el candidato A y el candidato B, 0.60 de los votantes prefieren al candidato A. Si seleccionas aleatoriamente una muestra de diez

Votante	Preferencia
1	1
2	0
3	1
4	1
5	1
6	0
7	1
8	0
9	1
10	1

Tabla 2.5 Muestra de votantes.

personas, entre todas las que pueden votar, es muy poco probable que exactamente el 60% de ellas prefieran al candidato *A*. La proporción de personas en la muestra que prefieren al candidato *A* con facilidad puede ser un poco menor o un poco mayor a 0.60. La distribución muestral de p es la distribución que resulta si se toman muestras repetidas de diez personas y se determina la proporción p que favorece al candidato *A*.

La distribución muestral de p es un caso especial de la distribución muestral de la media. En la tabla 2.5 se observan las preferencias de una muestra hipotética de diez votantes. A los que prefieren al candidato *A* se les da 1 punto, y a los que prefieren al candidato *B* se les da 0 puntos. Observa que siete de los votantes prefieren al candidato *A*, por lo que la proporción en la muestra (p) es

$$p = 7/10 = 0.70$$

Como puedes observar, p es la media de los puntajes de preferencia de los votantes.

La distribución de p se aproxima a una distribución binomial. La distribución binomial es la distribución del número total de eventos favorables (por ejemplo, los que favorecen al candidato *A*), mientras que la distribución de $\bar{x}$ es la distribución de la media del número de eventos. Por supuesto, la media es el total dividido entre el tamaño de la muestra n. Por tanto, la distribución muestral de p y la distribución binomial difieren en que en la distribución muestral p es la media de los puntajes (0.70) y en la binomial se trabaja con el número total de eventos favorables (7).

La distribución binomial tiene una media:

$$\mu = np$$

Dividiendo entre n, ajustamos al hecho de que en la distribución muestral de p se trabaja con medias en lugar de totales. Entonces la media de la distribución muestral de p es:

$$\mu_p = p$$

La desviación estándar de la distribución binomial es:

$$\sigma = \sqrt{n\pi(1 - \pi)}$$

Dividiendo entre n debido a que p es una media y no un total, encontramos que el error estándar de p es:

$$\sigma_p = \sqrt{\frac{n\pi(1 - \pi)}{n}} = \sqrt{\frac{\pi(1 - \pi)}{n}}$$

Volviendo al ejemplo de los votantes, $\pi = 0.60$ (no debes confundir $\pi = 0.60$, que es la proporción de la población y $p = 0.70$, la proporción de la muestra) y $n = 10$. Por tanto, la media de la distribución muestral de p es 0.60. La desviación estándar es:

$$\sigma_p = \sqrt{\frac{0.60(1 - 0.60)}{10}} = 0.155$$

La distribución muestral de p es una distribución discreta. Por ejemplo, con $n = 10$, es posible tener una p de 0.50 o una p de 0.60, pero no es posible una p de 0.55.

La distribución muestral de p se aproxima a la distribución normal si n es grande y π no está próximo a 0 o a 1. Una regla general es que la aproximación es buena si $n\,\pi$ y $n(1 - \pi)$ son mayores que 10. La distribución muestral para nuestro ejemplo se muestra en la figura 2.9. Observa que $n(1 - \pi)$ es únicamente 4 y sin embargo la aproximación es muy buena.

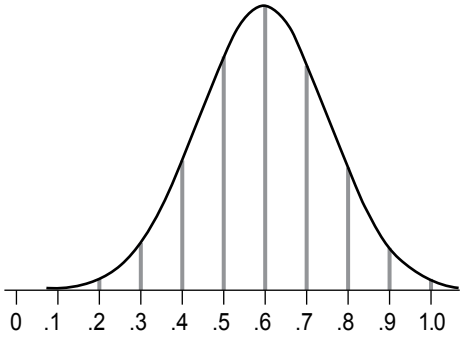


Figura 2.9 Distribución muestral de p. Las barras verticales son las probabilidades; la curva suavizada es la aproximación a la normal.

PREGUNTAS

1. La distribución binomial es la distribución del número total de eventos favorables, mientras que la distribución muestral de p es:

a) la distribución de la media del número de eventos favorables.

b) la distribución del número total de eventos no favorables.

c) la distribución de la relación del número de eventos al número de eventos no favorables.

d) una distribución con una media de 0.5.

2. De 300 estudiantes que presentaron examen de admisión en una escuela, 225 pasaron el examen. ¿Cuál es la media de la distribución muestral de la proporción de estudiantes que pasaron el examen de admisión?

a) Respuesta _____.

3. De 300 estudiantes que presentaron examen de admisión en una escuela, 225 pasaron el examen. Si seleccionas de todos los estudiantes una muestra de 10 ¿cuál es el error estándar de p?

a) Respuesta _____.

4. Un amigo tuyo te asegura que tiene percepción extrasensorial. Te dice que al lanzar una moneda puede adivinar si cae águila o sol. Decides aceptar que él tiene percepción extrasensorial, si acierta cuando menos el 80% de las veces. Suponiendo que la moneda es legal, ¿cuál es la probabilidad de que en 15 lanzamientos él pueda adivinar correctamente el resultado cuando menos el 80% de las veces?

a) Respuesta _____.

Actividades

I. Contesta y resuelve los siguientes ejercicios para reafirmar los conceptos.

1. Una población tiene una media de 50 y una desviación estándar de 6. *a*) ¿Cuál es la media y la desviación estándar de la distribución muestral de la media para $n = 16$? *b*) ¿Cuál es la media y la desviación estándar de la distribución muestral de la media para $n = 20$?

2. Los puntos de un examen se distribuyen en forma normal con una media de 100 y una desviación estándar de 12, encuentra:

a) la probabilidad de que un examen seleccionado al azar tenga un puntaje mayor que 110.

b) la probabilidad de que una muestra de 25 exámenes tenga una media mayor que 105.

c) la probabilidad de que una muestra de 64 exámenes tenga una media mayor que 105.

d) la probabilidad de que la media de una muestra de 16 exámenes se encuentre entre 95 y 105 puntos.

3. ¿Con que término nos referimos a la desviación estándar de la distribución muestral?

4. Si el error estándar de la media es 10 para $n = 12$, *a*) ¿cuál es el error estándar de la media para $n = 22$? *b*) Si el error estándar de la media es 50 para $n = 25$, ¿cuál es para $n = 64$?

5. Un cuestionario se diseñó para evaluar las actitudes de los hombres y las mujeres acerca del uso de los animales en la investigación. Una pregunta es si la investigación con animales no es adecuada, y se da una escala de 7 puntos para contestarla. Supón que, en la población, la media para las mujeres fue de 5 y para los hombres de 4, y la desviación estándar para ambos grupos fue de 1.5. Supón que las respuestas se distribuyen de manera normal. Si seleccionan aleatoriamente 12 hombres y 12 mujeres, ¿cuál es la probabilidad de que la media de las mujeres sea más de 1.5 puntos que la media de los hombres?

6. Si un número grande de muestras de tamaño $n = 15$ se seleccionan de una distribución uniforme y se grafica la distribución de las frecuencias relativas de las medias, ¿cuál sería la forma de esta distribución de frecuencias?

7. Una distribución normal tiene una media de 20 y una desviación estándar de 10. Si dos muestras se seleccionan aleatoriamente ¿cuál es la probabilidad que la diferencia entre las medias de las muestras sea más de 5?

8. Si seleccionas una muestra de tamaño 1 de una distribución estándar normal, ¿cuál es la probabilidad de que sea menor a 0.5?

9. Una variable se distribuye de manera normal con una media de 120 y una desviación estándar de 5. Cuatro elementos se seleccionan aleatoriamente. ¿Cuál es la probabilidad de que su media esté por arriba de 127?

10. La media del rendimiento escolar de los estudiantes de la escuela *A* es 3.0; la media para los estudiantes del colegio *B* es 2.8. La desviación estándar para ambas escuelas es 0.25. El rendimiento se distribuye de manera normal. Si nueve estudiantes se seleccionan de cada escuela, cuál es la probabilidad de que: *a*) la media muestral para la escuela *A* sea mayor en 0.5 o más puntos; *b*) la media muestral de la escuela *B* sea mayor que la de la escuela *A*.

11. En una ciudad, el 70% de la gente prefiere al candidato *A*. Supón que se seleccionan 30 personas de la ciudad. *a*) ¿Cuál es la media de la distribución muestral de p? *b*) ¿Cuál es el error estándar de p? *c*) ¿Cuál es la probabilidad de que el 80% o más de las personas de esta muestra prefieran al candidato *A*? *d*) ¿Cuál

es la probabilidad de que el 45% de las personas de esta muestra prefieran a otro candidato?

12. Una población tiene una media de 1 000, ¿tendría mayor probabilidad (o igual probabilidad) de obtener una media muestral de 1 200 si muestrea aleatoriamente a 10 o a 30 estudiantes? Explica.

13. El error estándar de la media es menor cuando $n = 20$ que cuando $n = 10$.

Verdadero_____ Falso_____.

14. Se seleccionan 20 estudiantes de una población y se calcula la media de sus estaturas. Se repite este proceso 100 veces y se gráfica la distribución de las medias. En este caso, el tamaño de muestra es 100.

Verdadero_____ Falso_____.

15. El 40% de los estudiantes de tu escuela ven televisión por la noche. Les preguntas a cinco estudiantes seleccionados cada día si vieron televisión por la noche. Todos los días debes encontrar que 2 de los 5 estudiantes seleccionados contestan afirmativamente.

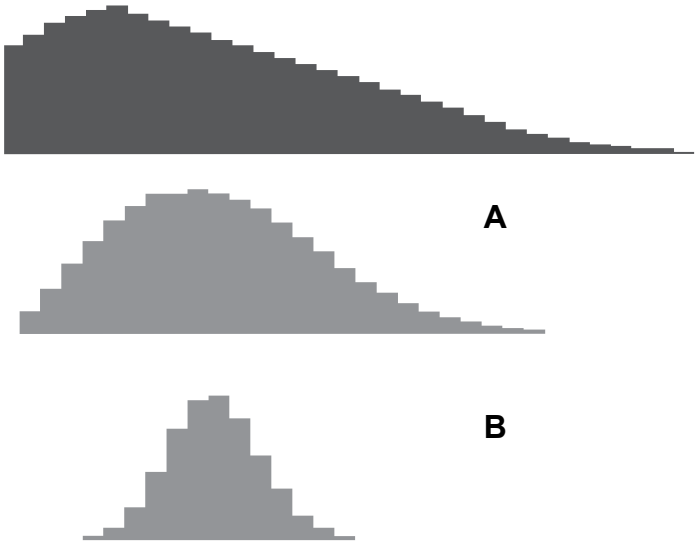
Verdadero_____ Falso_____.

16. La mediana tiene distribución muestral.

Verdadero_____ Falso_____.

17. La distribución de la población es la que se muestra primero y su correspondiente distribución muestral de la media para $n = 10$ es la señalada con la letra *A*.

Verdadero_____ Falso_____.



II. Resuelve los siguientes ejercicios de aplicación.

1. El valor de las cuatro marcas principales (Coca-Cola, Microsoft, IBM y General Electric)[1], en miles de millones de dólares es: 67.5, 59.9, 53.4 y 47.0

[1] El valor de una marca es independiente del valor real de la compañía que representa.

respectivamente. (*Fuente:* "Interbrand", en revista *Muy interesante*, julio de 2006). Considera estos datos como una población, realiza muestreos de tamaño 2 con reemplazo y comprueba que la media poblacional es igual a la media de medias.

2. Los salarios mínimos de los últimos 6 años (2003-2008) en México para el área geográfica A son 43.65, 45.24, 46.80, 48.67, 50.57 y 52.59. (*Fuente*: Comisión de los Salarios Mínimos). Considera estos datos como una población:

a) determina el número de muestras posibles de tamaño 4 sin reemplazo.

b) elabora la distribución teórica de la media en el muestreo de los salarios mínimos.

c) encuentra la media de medias.

d) calcula el error estándar.

3. La siguiente población de datos se refiere al tipo de cambio del dólar en pesos (*Fuente*: Banco de México, del 7 al 13 de octubre de 2008): 11.77, 12.12, 13.04, 12.44, 13.09, encuentra:

a) el número de muestras posibles de tamaño 3 sin reemplazo.

b) elabora la distribución muestral de la media del precio del dólar.

c) encuentra la media de medias.

d) calcula el error estándar.

4. La estatura de los estudiantes al ingresar a la universidad es una variable normalmente distribuida, con una media de 1.65 m, con una desviación estándar de 16 cm. Si se toma una muestra aleatoria de 25, ¿cuál es la probabilidad de que la muestra revele una media muestral de por lo menos 1.70 m? ¿Cuál es la probabilidad de que la muestra revele una media muestral entre 1.58 y 1.73 m? ¿Cuál será la estatura mínima del 8% de los estudiantes más altos?

5. La población de los salarios de los obreros calificados de la fábrica La Favorita está uniformemente distribuida. Se toma una muestra aleatoria de $n = 25$:

a) ¿cuál es la probabilidad de que una $\bar{x}$ se encuentre entre 400 y 500 pesos si una información independiente nos dice que la $\mu_{\bar{x}} = 600$ y que $\sigma_{\bar{x}} = 75$?

b) ¿cuál sería la respuesta si $n = 100$?

6. Una población de seis estudiantes es considerada para determinar la proporción de fumadores, y el resultado de la encuesta mostró que la primera y las últimas tres personas entrevistadas no fumaban. Realiza muestreos de tamaño 4 sin reemplazo y comprueba el teorema 3.

7. Una población de estudiantes tiene los siguientes gastos diarios por concepto de transporte, en pesos: 20, 35, 40, 13, 10. Calcula:

a) el número de muestras posibles de tamaño 3 sin reemplazo, para la proporción de estudiantes que tienen gastos en transporte por arriba de los 30 pesos.

b) elabora la distribución muestral de la proporción en el muestreo del inciso anterior.

c) calcula la proporción de proporciones.

d) calcula el error estándar.

8. Una población conformada por Pedro, María, Eugenia, José, Daniel y Blanca fue entrevistada con la finalidad de determinar la proporción de personas que tienen salarios por arriba de los 8 mil pesos mensuales. El resultado de la encuesta reveló que la 1a, 2a, 4a y 5a personas entrevistadas tenían sueldos menores a los 8 mil pesos mensuales:

a) determina el número de muestras posibles de tamaño 4 sin reemplazo.

b) elabora la distribución teórica de la proporción de personas que tienen sueldos superiores a los 8 mil pesos mensuales.

c) determina la proporción de proporciones.

d) calcula el error estándar.

9. Las edades de los miembros de un club de tabaco están normalmente distribuidas; una distribución muestral queda resumida por $\mu_{\bar{x}} = 50$ años y $\sigma_{\bar{x}} = 15$ años. ¿Cuál es la probabilidad de que una muestra aleatoria simple revele una media muestral de:

a) al menos 60 años?

b) entre 50 y 60 años?

c) entre 45 y 65 años?

d) cuando mucho 30 años?

e) entre 30 y 45 años?

10. La encuesta nacional sobre disponibilidad y uso de las tecnologías de la información 2008, realizada por el INEGI, reveló que sólo el 13.5% de los hogares en México tienen la disponibilidad de acceso a Internet. Si se toma una muestra aleatoria simple de 100 hogares, ¿cuál es la probabilidad de encontrar?

a) entre 10 y 15 hogares con acceso a Internet?

b) más de 12 hogares con acceso a Internet?

c) que más de 85 hogares no tengan acceso a Internet?

d) que entre 85 y 90 hogares no tengan acceso a Internet.

11. Se desea tomar una muestra aleatoria de tamaño $n = 200$ de la población estudiantil de la FES-C, que asciende a $N = 13\,032$ estudiantes, según el informe de 2007. (*Fuente*: Unidad de Administración Escolar, Secretaría General), con el objeto de conocer su opinión respecto al nuevo Reglamento de Exámenes Profesionales. Describe el procedimiento con cada uno de los siguientes métodos:

a) muestreo aleatorio simple, usando tablas de números aleatorios.

b) muestreo estratificado.

c) de los dos muestreos anteriores menciona cuál recomendarías y por qué.

12. Un estudiante que concluyó sus estudios de Psicología eligió para su trabajo de tesis el siguiente problema: ¿Cuáles son los factores fundamentales que incidieron durante el ciclo escolar 2006-2007, en el bajo rendimiento escolar de los estudiantes del segundo grado de primaria de las escuelas públicas de la zona centro del estado de Oaxaca, en el aprendizaje de las cuatro operaciones matemáticas básicas? Determina:

a) la población objetivo.

b) el diseño muestral más adecuado.

13. Supón que se desea hacer un estudio sobre la procedencia (zona de residencia) de los estudiantes de la FES-C, con el objeto de estudiar la posibilidad de resolver el problema del transporte de sus hogares a la facultad. Explica cómo tomarías una muestra representativa.

14. El número total de estudiantes que se titularon en la FES-C, en el periodo de octubre de 2006 a septiembre de 2007 (*Fuente*: Unidad de Administración Escolar, Secretaría General), se presenta en el siguiente cuadro:

Carrera	Número de titulados
Ingeniería Química	27
Química	9
Químico Farmacéutico Bitólogo	97
Ingeniería en Alimentos	44
Lic. en Contaduría	244
Lic. en Administración	137
Informática	8
Ingeniería Mecánica Eléctrica	95
Medicina Veterinaria y Zootecnia	88
Ingeniería Agrícola	22
Química Industrial	13
Diseño y Comunicación Visual	25

Se desea hacer una investigación para conocer si los titulados continuaron con estudios de posgrado. Para tal fin se requiere tomar una muestra que constituya el 5% de los titulados en el periodo indicado. Diseña un plan de muestreo de manera que los titulados de las carreras mencionadas queden representados proporcionalmente en la muestra.

15. A continuación se presenta la estatura de los estudiantes del grupo: 1 301 del semestre 2009-I de las licenciaturas en Administración y en Contaduría:

Alumno	Altura (m)	Alumno	Altura (m)	Alumno	Altura (m)
1	1.50	21	1.63	41	1.78
2	1.54	22	1.54	42	1.52
3	1.60	23	1.59	43	1.64
4	1.72	24	1.67	44	1.70
5	1.80	25	1.62	45	1.66
6	1.50	26	1.51	46	1.54
7	1.60	27	1.63	47	1.63
8	1.54	28	1.50	48	1.76
9	1.63	29	1.72	49	1.56
10	1.70	30	1.73	50	1.65
11	1.60	31	1.57	51	1.67
12	1.52	32	1.70	52	1.71
13	1.54	33	1.50	53	1.63
14	1.66	34	1.78	54	1.68
15	1.57	35	1.77	55	1.67
16	1.56	36	1.58	56	1.58
17	1.60	37	1.59	57	1.69
18	1.65	38	1.55	58	1.56
19	1.85	39	1.54	59	1.64
20	1.57	40	1.50	60	1.60

16. Toma de esta población una muestra de $n = 15$, usando la tabla de números aleatorios y calcula la media aritmética y la desviación estándar.

a) compara tus resultados con los de tus compañeros y discute las causas de las diferencias.

17. La población de los precios promedio en pesos, de la gasolina tipo Magna de los últimos 6 meses, excluyendo la frontera norte del país (*Fuente*: Pemex, indicadores petroleros) es: 7.33, 7.24, 7.17, 7.13, 7.10, 7.07:

a) calcula la media μ y la varianza σ^2.

b) determina las muestras posibles de tamaño 3, sin reemplazo.

c) elabora la distribución muestral de medias.

d) calcula la media de las medias.

e) calcula la desviación estándar de la distribución muestral de medias.

18. ¿Cuántas muestras aleatorias simples de tamaño 3 sin reemplazo se pueden obtener de las siguientes poblaciones?

a) una población de 5 estudiantes.

b) una población de 10 cuentas por pagar.

c) una población de 20 trabajadores del departamento de planeación de una empresa.

19. Se tiene una población de 5 obreros calificados, los cuales tienen los siguientes ingresos por laborar horas extras a la semana: \$758, \$618, \$550, \$589, \$720:

a) determina el número de muestras posibles de tamaño 3 y 4 sin reemplazo.

b) elabora las dos distribuciones muestrales para cada tamaño de muestra.

c) calcula la media de medias para ambos casos.

d) calcula el error estándar de las dos distribuciones.

e) observa los resultados obtenidos y discute el efecto del tamaño de muestra en el valor del error estándar.

20. Una población de las seis golosinas más consumidas en la facultad tienen los siguientes precios en pesos: 3.50, 6.00, 5.50, 4.00, 7.20 y 4.50. ¿Cómo quedaría resumida una distribución muestral para el promedio del costo de las golosinas, para una muestra aleatoria simple de $n = 3$?

21. Se sacan varias muestras de poblaciones normalmente distribuidas, con medias y varianzas, como se muestra a continuación:

a) $n = 10$; $\mu = 30$; $\sigma^2 = 9$.

b) $n = 15$; $\mu = 50$; $\sigma^2 = 4$.

c) $n = 30$; $\mu = 100$; $\sigma^2 = 100$.

d) $n = 100$; $\mu = 400$; $\sigma^2 = 64$.

e) en cada caso, determina la media y la desviación estándar de la distribución muestral de $\bar{x}$.

22. La población del precio promedio del diesel por litro de los últimos 6 meses en pesos (*Fuente*: Pemex, indicadores petroleros) es: 6.48, 6.18, 6.10, 6.05, 6.02, 5.99:

a) calcula el número de muestras posibles de tamaño 4 sin reemplazo.

b) elabora la distribución muestral del precio del diesel.

c) calcula la media de medias y el error estándar de la distribución.

23. Una población de 6 ejecutivos es considerada para determinar la proporción de personas que tienen automóvil utilitario. El resultado de la encuesta mostró que únicamente los dos primeros y el último ejecutivos entrevistados tenían este tipo de automóvil:

a) determina el número de muestras posibles de tamaño 5 sin reemplazo.

b) elabora la distribución muestral de la proporción de ejecutivos con automóvil utilitario.

c) encuentra el error estándar.

24. Una población de ocho golosinas muy consumidas en la facultad tienen los siguientes precios en pesos: 3.50, 6.00, 12.50, 5.80, 4.00, 7.20, 8.00 y 4.50. Establece la distribución muestral para la proporción de golosinas que tienen un costo superior a 7.00, para una muestra aleatoria simple de $n = 5$.

25. Un consumidor tiene un total de cuatro tarjetas de crédito cuyas tasas de interés mensual son 2.9, 3.4, 3.1 y 1.9. Para una muestra aleatoria simple de $n = 2$, Elabora:

a) la distribución muestral de la proporción de tarjetas que cobran interés por arriba del 15%.

b) calcula la proporción de proporciones.

c) calcula el error estándar.

26. El "Aviso oportuno" del periódico *El Universal* publicó el precio de venta de 5 inmuebles de diferentes características, el 14 de octubre de 2008, en miles de pesos:

a) determina el número de muestras posibles de tamaño 3 sin reemplazo.

b) elabora la distribución muestral de la proporción de los inmuebles que tienen un precio de venta superior a los 800.

c) encuentra la proporción de proporciones y el error estándar de la distribución.

Casa	Precio de venta (miles de pesos)
1	780
2	850
3	450
4	950
5	610

27. Una población de seis tornillos es considerada para determinar la proporción de tornillos defectuosos. El resultado de la prueba de calidad reveló que los tres primeros y el último tornillos revisados no presentaron defecto:

a) determina el número de muestras posibles de $n = 4$.

b) elabora la distribución muestral de la proporción de tornillos defectuosos.

c) calcula la proporción de proporciones y el error estándar de la distribución.

28. Se sabe que la media aritmética de la estatura de los estudiantes de la carrera de Informática es $\mu = 1.66$ m, con desviación típica $\sigma = 0.07$ m. Si se toma una muestra de tamaño 40, y considerando que está distribuida normalmente, determina la probabilidad de que la media obtenida en la muestra:

a) exceda de 1.65 m.

b) esté entre 1.63 m y 1.68 m.

c) sea inferior a 1.64 m.

d) sea inferior a 1.63 m o superior a 1.69 m.

e) ¿cuál será la estatura mínima del 10% de los estudiantes más altos?

29. Una máquina envasadora de azúcar sirve el producto en bolsas de plástico y está ajustada para verter 2 kg netos. La desviación estándar del proceso es $\sigma = 18$ gramos. Si se toma una muestra de 40 bolsas:

a) ¿cuál será el peso máximo que tendrán el 8% de las bolsas más vacías?

b) ¿cuál será el peso mínimo que tendrán el 10% de las bolsas más llenas?

30. Las cuentas de gastos en gasolina por día de los vendedores de una empresa de cigarros tiene una media de 65 pesos y una desviación estándar de $\sigma = 7$ pesos. Si se selecciona una muestra aleatoria de 25 cuentas, y considerando que estos gastos tienen una distribución normal, ¿cuál es la probabilidad de que la muestra revele una media:

a) al menos de 68 pesos?

b) entre 64 pesos y 66 pesos?

c) ¿cuál es el gasto máximo reportado del 20% de los vendedores que gastan menos en este rubro?

31. Con referencia al problema anterior, si la empresa tiene un total de 190 vendedores, contesta las mismas preguntas.

32. Fabricantes Muebleros, S. A. (FAMSA) vende, entre otros artículos, televisores marca Sony. Para poder lograr el máximo descuento por volumen, todas las tiendas de la cadena deben hacer un nuevo pedido de estos aparatos al mismo tiempo. La decisión de realizar los nuevos pedidos para inventario es hacerlos cuando las existencias, en una muestra aleatoria, de tiendas, sean a lo sumo de 15 televisores. Según registros, la desviación típica es 5. Si se selecciona una muestra aleatoria de 16 tiendas, ¿cuál es la probabilidad de que se vuelvan a pedir televisores de esa marca cuando el inventario promedio real de todas las tiendas sea 12? Considera una distribución normal.

33. Al examinar los registros de facturación mensual de una empresa editora con ventas por Internet, el auditor informa a la gerencia que el promedio de ventas de las facturas ascienden a la cantidad de $\mu = 1\,213$ pesos, con una desviación típica $\sigma = 250$ pesos. Si se toma una muestra de 50 facturas, ¿cuál es la probabilidad de que la muestra revele una media:

a) superior a los 1 300?

b) entre 1 150 y 1 200?

c) ¿cuál es la venta mínima del 10% de las facturas más altas?

d) ¿cuáles son las ventas del 95% de las facturas?

34. El departamento de impuestos de una empresa está conformado por los siguientes contadores, así como su título actual:

Contador	Título
Martínez	Licenciado
Pérez	Maestría
Hernández	Licenciado
López	Licenciado
Castro	Maestría

a) considera a estos 5 contadores como una población.

b) calcula el número de muestras posibles de tamaño 3 sin reemplazo.

c) elabora la distribución muestral de proporciones de los contadores con grado de maestría.

d) calcula la proporción de proporciones y el error estándar de la distribución muestral.

35. El 82% de las personas prefieren la pasta dental Colgate. Si se selecciona una muestra aleatoria de 80 personas, ¿cuál es la probabilidad de que la proporción de personas que prefieren esta marca:

a) esté entre 80 y 85%?

b) sea menor que 86%?

c) sea igual o mayor que 86%?

d) sea mayor que 90%?

36. De las 420 empresas manufactureras en cierta zona de León Guanajuato, 14% de ellas se dedican a la producción de calzado. Si se toma una muestra aleatoria de 80 empresas, ¿cuál es la probabilidad de que, de esa muestra, 10% o más se dediquen a la producción de calzado?

37. Se sabe que el 7% de los focos que llegan a las tiendas distribuidoras Home Mart presentan algún tipo de defecto. Si de un pedido de 5 000 focos se extrae una muestra aleatoria de tamaño 100, sin reemplazo:

a) determina el valor esperado de la distribución muestral de proporciones.

b) determina el error estándar de la distribución muestral de proporciones.

c) ¿cuál es la probabilidad de que 10 focos o más de la muestra estén defectuosos?

d) ¿cuál es la probabilidad de que la proporción de focos defectuosos esté entre 8 y 9%?

38. Una de las tiendas distribuidoras de focos Home Mart utiliza el siguiente criterio para aceptar o rechazar lotes de 500 focos que recibe bimestralmente: seleccionar una muestra aleatoria de 80 focos; si 3% o más presentan algún tipo de defecto, rechaza el lote; en caso contrario, lo acepta. ¿Cuál es la probabilidad de rechazar el lote que contiene 2% de focos defectuosos?

39. Según registros que lleva cierta cadena de tintorerías, 20% de los clientes pagan con tarjeta de crédito. Si se selecciona una muestra de 200 órdenes:

a) ¿cuál es el valor esperado del porcentaje de órdenes pagadas con tarjeta de crédito?

b) ¿cuál es la probabilidad de que la muestra revele una proporción de órdenes pagadas con tarjeta de crédito entre 23 y 26%?

40. Un auditor del Servicio de Administración Tributaria México, SAT, utiliza la siguiente regla de decisión para examinar o no todas las declaraciones de impuestos sobre la renta que presenta un despacho contable: toma una muestra aleatoria de 60 declaraciones; si 5% o más indican deducciones no autorizadas, se examinan todas las declaraciones:

a) ¿cuál es la probabilidad de examinar todas las declaraciones, si realmente 3% de estás indican deducciones no autorizadas?

b) ¿cuál es la probabilidad de no examinar todas las declaraciones, si realmente 7% de ellas indican deducciones no autorizadas?

41. El análisis estadístico de las llamadas telefónicas de larga distancia (Telmex), indica que 35% de ellas duran por lo menos 480 segundos. Si se toma una

muestra aleatoria de 50 llamadas, ¿cuál es la probabilidad de que el porcentaje de llamadas cuya duración sea de por lo menos 480 segundos:

a) esté entre 25 y 50%?

b) sea menor que 30%?

c) sea a lo sumo 45%?

42. Los registros que lleva el departamento de servicios de una agencia automotriz de Chrysler indican que 18% de todos los automóviles nuevos de la marca Stratus han requerido cierto tipo de reparación durante el periodo de su garantía. Si se toma una muestra de 64 automóviles nuevos de esta marca, ¿cuál es la probabilidad de que:

a) el porcentaje de autos que necesiten reparación esté entre 12 y 16%?

b) a lo sumo 20% necesiten reparación?

c) si se toma una muestra al azar de 80 automóviles nuevos, 8 o más necesiten algún tipo de reparación?

43. El peso promedio de las personas en el corporativo de Banco Azteca es de 68.3 kg, con una desviación estándar de 7.9 kg. La población de los pesos está normalmente distribuida. Si la capacidad máxima del elevador del corporativo es de 1 135 kg y se indica como "16 personas", ¿cuál es la probabilidad de que el elevador siempre sea seguro?

44. En una gran ciudad, una ambulancia tarda un promedio de 12 minutos en llegar después de una llamada de emergencia. La desviación estándar es de 4 minutos:

a) describe la distribución muestral del tiempo medio de respuesta para una $n = 36$ llamadas.

b) calcula la probabilidad de hallar una media muestral de más de 11 minutos.

c) calcula la probabilidad de entre 13 y 14 minutos.

d) ¿cuál es la probabilidad de hallar una media muestral que se encuentre a 30 segundos de la media poblacional?

45. Si el 38.4% de todos los profesores de carrera de la FES-C tienen doctorado (*Fuente*: Informe 2007, Secretaría de Planeación) y se toma una muestra aleatoria simple de 60 profesores de carrera, ¿cuál es la probabilidad de encontrar:

a) más de 25 con ese grado?

b) entre 18 y 28 con ese grado?

c) más de la mitad con ese grado?

46. Si bien la mayoría de las personas creen que el desayuno es el alimento más importante del día, 25% de los adultos no desayunan. Si se toma una muestra aleatoria de 200 adultos:

a) ¿cuál es la probabilidad de que la proporción muestral quede a ±0.03 de la proporción poblacional?

47. El salario inicial promedio mensual para auxiliares contables con un año de experiencia es de 5 500 pesos con una desviación estándar de 1 100 pesos.

a) ¿cuál es la probabilidad de que una muestra aleatoria de 50 auxiliares contables tengan una media muestral dentro de ±300 pesos de la media poblacional?

48. La revista *Muy interesante* (agosto de 2006) publicó que 7 de cada 10 mexicanos padece de sobrepeso. Si se toma una muestra aleatoria de 80 personas.

a) ¿cuál es la probabilidad de que por lo menos 65 padezcan de sobrepeso?

b) ¿cuál es la probabilidad de que ¾ partes de la muestra padezcan de sobrepeso?

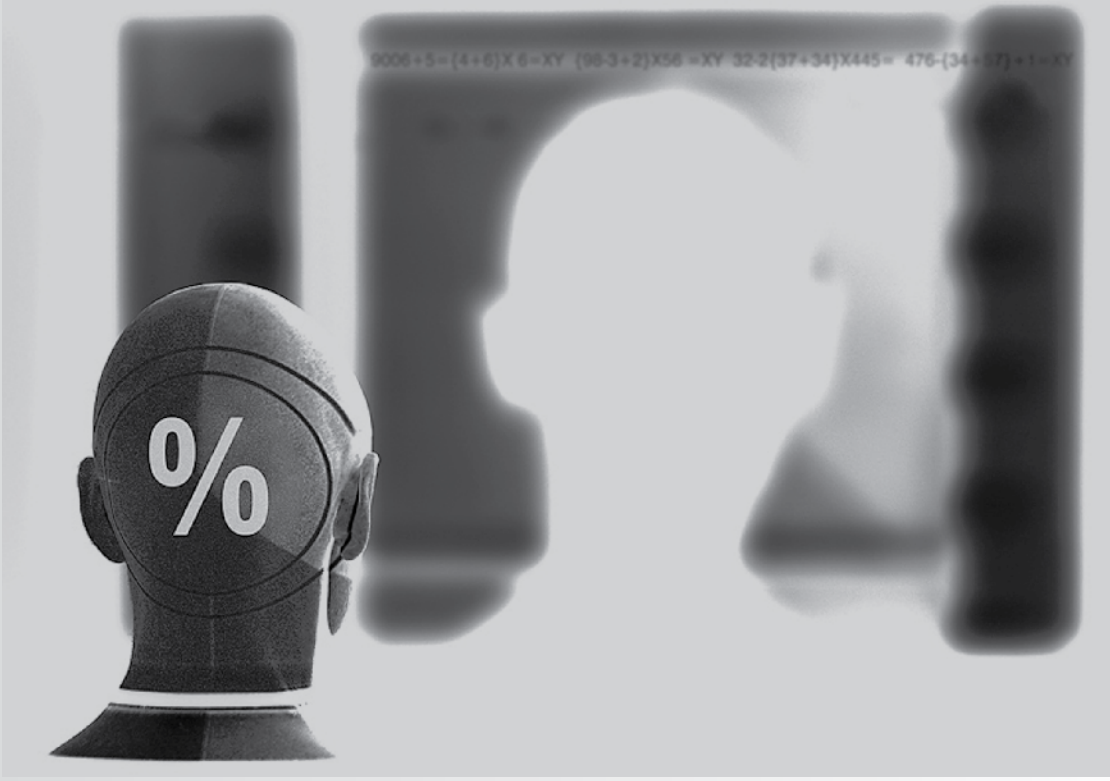
49. La encuesta nacional sobre disponibilidad y uso de las tecnologías de la información en los hogares en México 2008, realizada por el INEGI, reveló que el 29.1% de los usuarios de Internet que han realizado transacciones electrónicas han sido para realizar compras. Si se toma una muestra de 100 usuarios de Internet:

a) ¿cuál es la probabilidad de que por lo menos 35 hayan realizado alguna compra por Internet?

b) ¿cuál es la probabilidad de que cuando mucho 65 no hayan realizado alguna compra por Internet?

50. Encuestas de México sabe que el 38% de todas las personas que son encuestadas por teléfono contestan el cuestionario. ¿Cuál es la probabilidad de que por lo menos 200 personas contesten y respondan a una encuesta si se toma una muestra aleatoria simple de 500 personas?

51. Si la mitad de todos los pedidos de McDonald's incluyen un helado, ¿cuál es la probabilidad de que entre 40 y 55% de los siguientes 36 pedidos incluyan un jugo?



3 Estimación

Una de las aplicaciones más importantes de la estadística es la estimación de los parámetros de la población a partir de los estadísticos de una muestra. Por ejemplo, mediante una encuesta se desea estimar la proporción de adultos que está de acuerdo con la propuesta de la construcción de una autopista que pasará por una reserva ecológica. De un grupo de 200 personas, 106 seleccionadas aleatoriamente están de acuerdo con esta construcción. Entonces, podemos decir que el 0.53 de las personas que conforman la muestra están de acuerdo con la proposición. A este valor de 0.53 se le denomina estimación puntual de la proporción de la población. Es una estimación puntual porque la estimación proporciona un solo valor o punto.

Los estimadores puntuales se utilizan para estimar intervalos conocidos como intervalos de confianza. Éstos son intervalos que se construyen usando un método que permite que el parámetro de la población esté contenido en los intervalos en una proporción especificada de veces. Por ejemplo, si la compañía encuestadora usa un método que contiene al parámetro 95% de las veces, llegará al siguiente intervalo de confianza: $0.46 < \pi < 0.60$. La encuestadora concluirá que una proporción de personas adultas en la población entre 0.46 y 0.60 están de acuerdo con la propuesta. Los medios de comunicación, por lo general, dan a conocer este tipo de resultados diciendo que el 53% de los encuestados están a favor, con un margen de error del 7%. La sección sobre intervalos de confianza muestra cómo se calculan éstos para diferentes parámetros.

En una encuesta realizada a consumidores de dos tipos de jugo de naranja se les pidió que calificaran de 0 a 10 diferentes aspectos, como precio, sabor, dulzura y densidad. La calificación promedio para el jugo "Pura naranja" fue de 6.82, mientras que para el "Big orange" fue de 8.17. Por tanto, una estimación puntual de la diferencia de las medias poblacionales es 1.35. El intervalo de confianza al 95% para la diferencia de medias se extiende de 1.04 a 1.67 puntos.

PREGUNTAS

1. Estimamos el _____ con un _____ de la muestra.

a) parámetro; estadístico.

b) estadístico; parámetro.

2. Seleccionar todo lo que aplique. De las personas muestreadas en un estado, el 0.63 prefieren al senador Martínez. Este valor de 0.63 es:

a) parámetro.

b) estadístico.

c) estimación puntual.

d) estimación por intervalo.

e) intervalo de confianza.

Grados de libertad

Algunas estimaciones se basan en más información que otras. Por ejemplo, una estimación de la varianza a partir de una muestra de tamaño 100 se basa en mayor información que la misma estimación a partir de una muestra de tamaño 5. Los grados de libertad (gl) de un estimador es el número de elementos independientes de información con los cuales se calculó.

Como ejemplo, supongamos que conocemos el promedio de estatura de los marcianos adultos y éste es igual a 6 unidades marcianas, y que deseamos estimar la varianza de las estaturas. Seleccionamos a un marciano adulto al azar y encontramos que su estatura es 8. Recuerda que la varianza se define como la media de las desviaciones cuadradas de los valores individuales o datos con respecto a la media de la población. Entonces, calculamos la desviación cuadrada de nuestro valor 8 con respecto a la media de la población 6. La desviación al cuadrado de este valor individual con respecto a la media es $(8 - 6)^2 = 4$ y es un estimador de la desviación cuadrada media para todos los marcianos. Por tanto, con base en una muestra de un solo marciano estimamos que la varianza de la población es 4. Esta estimación se realizó con base en una sola pieza de información y, por tanto, tiene un grado de libertad. Si ahora seleccionamos al azar a otro marciano y éste mide 5 unidades, tendremos una segunda estimación de la varianza $(5 - 6)^2 = 1$. Con el promedio de nuestras dos estimaciones (4 y 1) obtenemos una mejor estimación, 2.5. Esta estimación se obtuvo con dos piezas independientes de información por lo que tiene dos grados de libertad. Las dos estimaciones son independientes porque se realizaron con base en las mediciones hechas a dos marcianos seleccionados al azar y en forma independiente. Una estimación no sería independiente, por ejemplo, si después de seleccionar al primer marciano decidimos seleccionar a su hermano como el segundo marciano.

Es probable que pienses que sería muy raro que se conozca el valor de la media de la población cuando el problema es estimar la varianza. En el caso de la

media poblacional desconocida (que es lo más frecuente), primero la tendríamos que estimar con la media de la muestra ($\bar{x}$). El hecho de haber estimado la media afecta a los grados de libertad como se muestra en seguida.

Volviendo a nuestro problema de la estimación de la varianza de las estaturas de los marcianos, supongamos que no se conoce la media de la población y, por tanto, la estimamos con la media de la muestra. Tenemos una muestra de dos marcianos con alturas de 8 y 5. Por tanto $\overline{X}$, nuestra estimación de la media de la población es:

$$\overline{X} = (8 + 5)/2 = 6.5$$

Ahora calculemos dos estimaciones de la varianza:

$$\text{estimación } 1 = (8 - 6.5)^2 = 2.25$$

$$\text{estimación } 2 = (5 - 6.5)^2 = 2.25$$

Y hagamos la pregunta clave: ¿las dos estimaciones son independientes? La respuesta es negativa, porque cada estatura contribuye al cálculo de $\overline{X}$. La estatura del primer marciano, 8, tiene influencia sobre $\overline{X}$, y además tiene influencia en la segunda estimación. Por ejemplo, si el primer marciano hubiera medido 10, entonces $\overline{X}$ hubiera sido 7.5 y la segunda estimación sería $(5 - 7.5)^2 = 6.25$ en lugar de 2.25. El punto importante es el hecho de que las dos estimaciones no son independientes y por consiguiente no tenemos dos grados de libertad. Otra forma de conceptualizar la falta de independencia es pensar, que si conoces la media y uno de los datos, entonces puedes determinar el valor del otro dato. Por ejemplo, si uno de los datos es 5 y la media es 6.5, fácilmente puedes calcular que la suma de los dos datos es 13 y entonces el otro dato es $13 - 5 = 8$.

En general, los grados de libertad para un estimador son iguales al número de datos menos el número de parámetros estimados para llevar a cabo el cálculo del estimador en cuestión. En el ejemplo de los marcianos, tenemos dos datos (8 y 5), y tenemos que estimar un parámetro, μ, en el proceso de estimar el parámetro de interés, σ^2. Por tanto, la estimación de la varianza tiene $2 - 1 = 1$ grado de libertad. Si hubiéramos seleccionado una muestra de 12 marcianos, tendríamos para estimar la varianza 11 grados de libertad. Entonces los grados de libertad para la estimación de la varianza son igual a $n - 1$, donde n es el número de observaciones o tamaño de la muestra.

La fórmula para la estimación de la varianza de la muestra es:

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n - 1}$$

El denominador de esta fórmula son los grados de libertad.

PREGUNTAS

1. Conoces la media de una población. Seleccionas 10 elementos al azar de esta población y estimas la desviación estándar. ¿Cuántos grados de libertad tiene esta estimación?

a) Respuesta _____.

2. No conoces la media de una población. Seleccionas 15 elementos al azar de esta población y estimas la media y la desviación estándar. ¿Cuántos grados de libertad tiene la estimación de la desviación estándar?

a) Respuesta _____

3. ¿Para cuál de estos grados de libertad el estadístico muestral será una estimación menos precisa del parámetro poblacional?

a) 21.

b) 5.

c) 2.

d) 100.

Características de los estimadores

En esta sección se discuten dos importantes características de los estadísticos usados como estimadores puntuales de los parámetros: el sesgo y la variabilidad del muestreo. El sesgo se refiere a la tendencia que muestra un estimador en sobreestimar o subestimar el parámetro. La variabilidad del muestreo se refiere a la variabilidad observada en el valor del estimador de muestra a muestra.

Supón que tienes que realizar un estudio de protección al consumidor y con este fin realizas una inspección a las básculas de los comercios de alimentos. Uno de tus inspectores te muestra los resultados de dos básculas: una, de un supermercado, y otra de una tienda de abarrotes, en las cuales se pesó una bolsa de verduras preparada para tal fin. La báscula 1 tenía una escala digital y proporcionó esencialmente el mismo resultado al pesar la bolsa repetidamente. La variación máxima entre pesada y pesada fue de 3.5 g. Esta báscula potencialmente es muy precisa, pero se calibró incorrectamente y en promedio subestima el peso en 50 g. La báscula 2 es muy barata y proporciona muy diferentes resultados entre pesadas. Algunas veces sobreestima mucho y otras veces subestima de la misma manera. Sin embargo, el promedio de un número grande de pesadas proporcionó el peso real. La báscula 1 tiene sesgo porque en promedio las mediciones proporcionan un peso 50 g por abajo del peso real. En comparación, la báscula 2 proporciona estimaciones sin sesgo en el peso, pero las mediciones tienen mucha variabilidad y frecuentemente están muy alejadas del valor real. La báscula 1, a pesar de proporcionar mediciones sesgadas, es muy precisa. Las mediciones nunca exceden de 3.5 g del peso real. Ahora proporcionemos las definiciones de exactitud y precisión (variabilidad). Las ideas básicas son las mismas que se desarrollaron para las básculas.

Sesgo

Un estadístico presenta sesgo si el promedio con muchos datos del mismo no es igual al parámetro que se está estimando. Más formalmente se dice que un estadístico es sesgado si la media de la distribución muestral es diferente al parámetro. A la media de la distribución muestral del estadístico se le llama algunas veces *valor esperado del estadístico*.

Como se vio en la sección de la "Distribución muestral de la media", la media de la distribución muestral de la media es igual a la media de la población, μ. Por tanto, la media de la distribución muestral de la media es un estimador insesgado de μ. Cualquier media de una muestra dada puede subestimar o sobreestimar a μ, pero no existe una tendencia sistemática de las medias muestrales, en subestimar o sobreestimar a μ. Un estimador *insesgado* se dice que es exacto.

Se ha visto que la fórmula para la varianza de una población es:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}$$

La fórmula para estimar la varianza a partir de una muestra es:

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n - 1}$$

Ten en cuenta que los denominadores de las fórmulas son diferentes: N para la población y $n - 1$ para la muestra. Si se usa n en la fórmula para s^2, entonces las estimaciones tienden a ser menores y, por tanto, el estimador es sesgado. La fórmula con $n - 1$ en el denominador proporciona un estimador insesgado de la varianza de la población. Nota que $n - 1$ son los grados de libertad.

Variabilidad muestral

La variabilidad de un estadístico se refiere a la variación del valor del estadístico de muestra a muestra, y se mide generalmente mediante el error estándar; un error estándar más pequeño indica una menor variabilidad muestral. Por ejemplo, el error estándar de la media es una medida de la variabilidad muestral de la media. Recuerda que la fórmula para el error estándar de la media es:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Cuanto más grande sea la muestra n, menor será el error estándar de la media y, por tanto, será menor la variabilidad muestral de la media. A la variabilidad muestral se le conoce también como precisión.

Diferentes estadísticos tienen distinta variabilidad muestral aun cuando se calculen con el mismo tamaño de muestra. Por ejemplo, para la distribución normal, el error estándar de la mediana es mayor que el error estándar de la media. Cuanto menor sea el error estándar de un estadístico, mayor eficiencia tendrá éste. La eficiencia relativa de dos estadísticos se define como la relación de sus errores estándar, aunque algunas veces se define como la relación de sus errores estándar cuadráticos.

PREGUNTAS

1. Estás jugando a "ponle la cola al burro", en una fiesta de cumpleaños. Mientras estás vendado tienes tres oportunidades de poner la cola en el lugar correcto; en tus tres intentos la pones en la parte baja de la pata y cerca del casco. Selecciona todos los términos que describen tu estimación:

a) insesgado.

b) sesgado.

c) preciso.

d) no preciso.

2. En una población, el valor de un parámetro es 15 con base en la media y el error estándar de la distribución muestral. ¿Cuál de los siguientes estimadores de este parámetro tiene mayor sesgo?

a) $\bar{x} = 14,$ $= 2$

b) $\bar{x} = 8,$ $\sigma_{\bar{x}} = 2$

c) $\bar{x} = 15,$ $\sigma_{\bar{x}} = 6$

d) $\bar{x} = 20,$ $\sigma_{\bar{x}} = 1$

3. En una población, el valor de un parámetro es 10; con base en la media y el error estándar de la distribución muestral, ¿cuál de los siguientes estimadores de este parámetro tiene la menor variabilidad muestral?

a) $\overline{X} = 10,$ $\sigma_{\bar{x}} = 5$

b) $\overline{X} = 9,$ $\sigma_{\bar{x}} = 4$

c) $\overline{X} = 11,$ $\sigma_{\bar{x}} = 2$

d) $\overline{X} = 13,$ $\sigma_{\bar{x}} = 3$

4. En una población, un parámetro llamado "Cuautitlán" tiene un valor de 9. Si el estimador de "Cuautitlán" es insesgado, ¿cuál es el valor esperado?

a) Respuesta _____.

Intervalo de confianza

Supón que estamos interesados en la duración media de un nuevo tipo de foco ahorrador de electricidad producido por la compañía José de la Luz, S. A. Ya que resultaría impráctico medir la duración de todos los focos producidos, se toma una muestra de 16 focos y se encuentra que la vida media de los focos es de 650 horas. La media de la muestra es igual a 650 horas y es un estimador puntual de la media de la población. Los estimadores puntuales tienen un uso limitado debido a que no muestran el grado de incertidumbre asociado con la estimación; no tienes idea de qué tan alejada está la media muestral de la media poblacional. Por ejemplo, ¿puedes tener confianza en que la media de la población se aleje de 650 ± 40 horas? Simplemente no lo sabes.

Los intervalos de confianza proporcionan mayor información que los estimadores puntuales. Los intervalos de confianza para la media son intervalos que se construyen de acuerdo con un procedimiento, que se revisa en la siguiente sección, de tal manera que contendrán a la media de la población una proporción especificada de veces, por lo general 90, 95 o 99% de las veces. Nos referiremos a estos intervalos como intervalos de confianza al 90, 95 y 99% de confianza, respectivamente. Un ejemplo de un intervalo de confianza al 95% sería:

$$630.4 < \mu < 669.6 \text{ horas}$$

Hay una buena razón para creer que la media de la población se encuentra dentro de estos dos límites, 630.4 y 669.6 horas, ya que en el 95% de las veces los intervalos de confianza construidos con este nivel de confianza contienen el verdadero valor de la media.

Si se toman muestras repetidamente y se calcula el intervalo de confianza al 95% para cada una de ellas, 95% de los intervalos contendrán la media de la población. Naturalmente, 5% de los intervalos no la contendrán.

Es natural que un intervalo de confianza al 95% se interprete como un intervalo que tiene una probabilidad de 0.95 de contener la media de la población. Sin embargo, esta interpretación no es la apropiada. Una vez que calculaste el intervalo de confianza, éste contiene el parámetro o no lo contiene. Es decir la probabilidad

es 0 o 1. Por tanto, lo más apropiado es expresarse en términos de confianza y no de probabilidad. Otro problema que aparece es que en el cálculo de los intervalos de confianza no se toma en cuenta cualquier otra información que se pudiera tener acerca del valor de la media poblacional. Por ejemplo, si muchos estudios previos han encontrado que todas las medias de las muestras se encuentran por arriba de 600, no tiene sentido concluir que tenemos una confianza de 0.95, de que la media de la población se encuentre entre 572.85 y 597.15 horas.

¿Qué podríamos decir de las situaciones en las que no existe información previa acerca del valor de la media de la población? Aun en este caso la interpretación es compleja. El problema es que puede haber más de un procedimiento para construir intervalos que contengan el parámetro poblacional 95% de las veces. ¿Cuál de los procedimientos proporciona el intervalo de confianza al 95% "verdadero"? Aunque los distintos métodos son iguales desde el punto de vista puramente matemático, el método estándar para calcular los intervalos de confianza tiene una propiedad deseable: cada intervalo es simétrico alrededor del estimador puntual.

Los intervalos de confianza pueden calcularse para varios parámetros, no sólo para la media. Por ejemplo, en una sección posterior veremos cómo se calcula el intervalo de confianza para ρ, coeficiente de correlación poblacional.

PREGUNTAS

1. Estrictamente hablando, ¿cuál es la mejor interpretación de un intervalo de confianza para la media?

a) si se realiza muestreo repetido y se calcula el intervalo de confianza al 95% para cada muestra, el 95% de los intervalos contienen la media de la población.

b) un intervalo de confianza al 95%, tiene una probabilidad de 0.95 de contener la media de la población.

c) el 95% de la distribución de la población está contenida en el intervalo de confianza.

2. Los intervalos de confianza sólo pueden calcularse para la media

a) falso.

b) verdadero.

Intervalo de confianza para la media

Cuando construyes un intervalo de confianza, calculas la media de la muestra con el fin de estimar la media de la población. Por supuesto, si conocieras el valor de la media de la población no tendrías ninguna necesidad de hacer un intervalo de confianza. Sin embargo, para explicar cómo se construyen los intervalos de confianza trabajaremos al revés, y empezaremos suponiendo que conocemos las características o los parámetros de la población. Entonces mostraremos cómo se pueden usar los datos de la muestra para construir un intervalo de confianza.

En el estudio del contenido de grasa en la leche light supón que el contenido medio de grasa en cada envase de un litro se distribuye en forma normal con una media de 10 g y una desviación estándar de 2.5 g. ¿Cuáles son los parámetros de la distribución muestral de la media para un tamaño de muestra de 9 envases? Recordemos que en la sección de la "Distribución muestral de la media" (cap. 2) vimos que la media de la distribución muestral es μ, y el error estándar de la media viene dado por:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Para nuestro ejemplo, la distribución muestral de la media tiene una media de 10 y una desviación estándar de 2.5/3 = 0.83. Observa que la desviación estándar de la distribución muestral es el error estándar. En la figura 3.1 se muestra esta distribución. El área sombreada representa el 95% del área central de la distribución y se extiende desde 8.4 a 11.6. Estos límites se calcularon sumando y restando 1.96 desviaciones estándar al valor de la media:

$$10 - (1.96)(0.83) = 8.4$$

$$10 + (1.96)(0.83) = 11.6$$

El valor de 1.96 es el valor de z, que le corresponde al 95% del área de una distribución normal con centro en la media de la distribución; el error estándar de la media es 0.83.

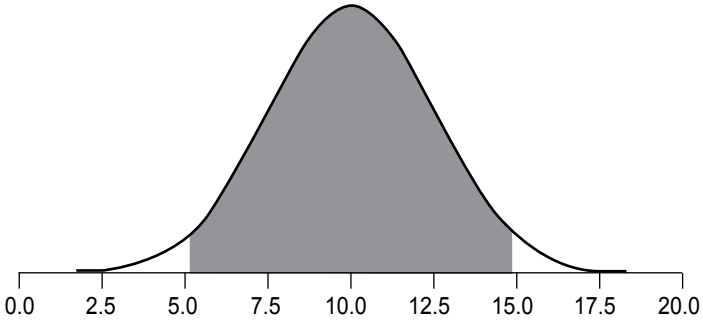


Figura 3.1 La distribución muestral de la media para $n = 9$.

La figura 3.1 muestra que el 95% de las medias no se alejan más de 1.6 unidades (1.96 desviaciones estándar) del valor de la media, que es 10. Ahora, consideremos la probabilidad de que la media de una muestra aleatoria se aleje de la media 10, cuando mucho 1.6 unidades. Ya que el 95% de la distribución se encuentra entre 10 ± 1.6 unidades, la probabilidad de que la media calculada a partir de una muestra dada no se aleje más de 1.6 de 10 es 0.95. Esto significa que si muestreamos repetidamente y para cada muestra se calcula la media y se construyen intervalos $\overline{X} - 2.61$ a $\overline{X} + 2.61$, los intervalos contendrán la media de la población, 95% de las veces. En general, se calcula un intervalo de confianza al 95% para la media con la fórmula siguiente:

$$\text{límite inferior} = \overline{X} - z_{0.95} \cdot \sigma_{\overline{x}}$$

$$\text{límite superior} = \overline{X} + z_{0.95} \cdot \sigma_{\overline{x}}$$

donde $z_{0.95}$ es el número de desviaciones estándar a partir de la media de una distribución normal que se requieren para incluir el 0.95 del área y $\sigma_{\overline{x}}$ es el error estándar de la media.

Si vemos con atención la fórmula para el intervalo de confianza, notamos que se necesita conocer el valor de la desviación estándar, σ, con el fin de estimar la media. Esto puede parecer poco real y de hecho así es. Sin embargo, desarrollemos un ejemplo, con fines didácticos, suponiendo σ conocida, debido a que el cálculo del intervalo de confianza es fácil. Más adelante se mostrará el cálculo de un intervalo de confianza para la media cuando σ es desconocida y hay que estimarla.

Supón que se obtuvo una muestra aleatoria de cinco datos de una población distribuida en forma normal con una desviación estándar de 2.5. Los cinco valores fueron: 2, 3, 5, 6, y 9. Para construir el intervalo de confianza al 95%, empezamos por calcular la media y el error estándar:

$$\overline{X} = \frac{2 + 3 + 5 + 6 + 9}{5} = 5$$

$$\sigma_{\overline{x}} = \frac{2.5}{\sqrt{5}} = 1.118$$

$Z_{0.95}$ se encuentra usando las tablas. Como se muestra en la figura 3.2, el valor es 1.96. Si deseas calcular el intervalo de confianza al 99%, debes indicar que el área sombreada es 0.99 y el resultado será 2.58.

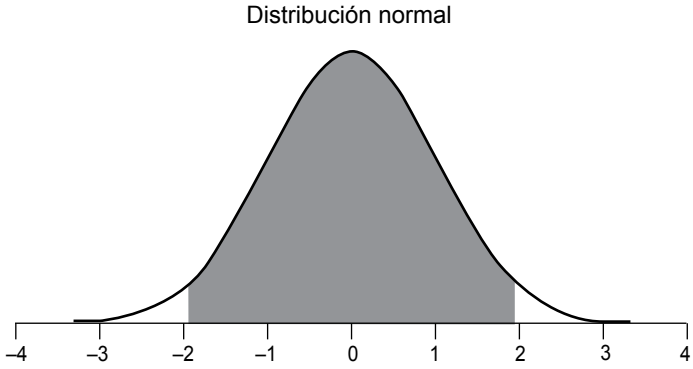


Figura 3.2 Distribución normal con un área sombreada de 0.95.

El intervalo de confianza se calcula como sigue:

$$\text{límite inferior} = 5 - (1.96)(1.118) = 2.81$$

$$\text{límite superior} = 5 + (1.96)(1.118) = 7.19$$

Cuando la varianza no se conoce, pero la estimamos a partir de los datos de la muestra, debemos usar la distribución *t* en lugar de la distribución normal. Cuando el tamaño de la muestra es muy grande, por ejemplo de 100 o más, la distribución *t* es prácticamente igual a la distribución normal estándar. Para muestras pequeñas, la distribución *t* es leptocúrtica, lo que significa que incluye más área en las colas en comparación con la distribución normal. Como resultado tienes que extender los límites más lejos de la media para que incluir la misma proporción del área. Recuerda que, para una distribución normal, el 95% de la distribución se encuentra entre ± 1.96 desviaciones estándar a partir de la media. Utilizando la distribución *t*, si tienes un tamaño de muestra de únicamente 5 elementos, el 95% del área se encuentra a ± 2.98 desviaciones estándar a partir de la media. Por tanto, el error estándar de la media será multiplicado por 2.98 en lugar de 1.96.

Los valores de *t* que se usan para construir intervalos de confianza se pueden encontrar en la tabla de la distribución *t*. Una versión reducida de ésta se muestra en la tabla 3.1. La primera columna, *gl*, muestra los grados de libertad. Los intervalos de confianza para la media tienen $n - 1$ grados de libertad, donde *n* es el tamaño de la muestra. Aprenderás más acerca de la distribución *t* en la próxima sección.

gl	0.95	0.99
2	4.303	9.925
3	3.182	5.841
4	2.776	4.604
5	2.571	4.032
8	2.306	3.355
10	2.228	3.169
20	2.086	2.845
50	2.009	2.678
100	1.984	2.626

Tabla 3.1 Versión reducida de la distribución *t*.

Supón que los siguientes 5 números fueron seleccionados aleatoriamente de una distribución normal: 2, 3, 5, 6 y 9, y que la desviación estándar es desconocida. El primer paso será calcular la media y la varianza de la muestra:

$$\overline{X} = 5$$

$$s^2 = 7.5$$

El siguiente paso es estimar el error estándar de la media. Si conociéramos la varianza de la población, usaríamos la fórmula siguiente:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

En lugar de esto, calculamos un estimador del error estándar ($s_{\bar{x}}$):

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = 1.225$$

El siguiente paso es encontrar el valor de *t*. Usando la tabla 3.1, el valor para un intervalo al 95% y $gl = n - 1 = 4$ es 2.776. El intervalo de confianza se calcula de la misma manera que lo hicimos cuando conocemos $\sigma_{\bar{x}}$. La única diferencia es que usamos $s_{\bar{x}}$ y *t* en lugar de $\sigma_{\bar{x}}$ y *z*, respectivamente.

$$\text{límite inferior} = 5 - (2.776)(1.225) = 1.60$$

$$\text{límite superior} = 5 + (2.776)(1.225) = 8.40$$

En forma general, la fórmula para un intervalo de confianza para la media viene dada por:

$$\text{límite inferior} = \overline{X} - (t_{\alpha/2,g.l.})(s_{\bar{x}})$$

$$\text{límite superior} = \overline{X} + (t_{\alpha/2,g.l.})(s_{\bar{x}})$$

Donde $\overline{X}$ es la media de la muestra, $t_{\alpha/2,g.l.}$ es el valor de *t* para el nivel de confianza deseado y los grados de libertad y $s_{\bar{x}}$ es la estimación del error estándar de la media.

Terminemos con un análisis de los datos para el caso de la valuación de inmuebles residenciales realizada del Banco del Centro, S. A. Una auditoría seleccionó al azar 10 valuaciones realizadas por el departamento de valuación de inmuebles y comparó cada una de éstas con el precio comercial real, sospechando que existe

una tendencia a sobrevaluar los bienes. La auditoría debe reportar si en realidad el banco sobrevalúa los bienes. En la tabla 3.2 se muestra la diferencia en pesos entre el precio valuado y el precio comercial real.

Inmueble	Valuado	Real
1	1 800 000	1 700 000
2	2 000 000	1 500 000
3	1 500 000	1 800 000
4	2 000 000	1 800 000
5	1 300 000	1 150 000
6	1 200 000	1 350 000
7	2 500 000	2 000 000
8	2 200 000	1 900 000
9	1 555 000	1 700 000
10	2 320 000	2 100 000

Tabla 3.2 Valuación de inmuebles residenciales.

La diferencia promedio para las 10 valuaciones es de 137 500 y la desviación estándar es de 269 848, el error estándar de la media es 85 333. En la tabla de t, encontramos que el valor de t, para $10 - 1 = 9$ grados de libertad, es igual a 2.262 para un intervalo de confianza al 95%. Por tanto, el intervalo de confianza se calcula como sigue:

$$\text{límite inferior} = 137\,500 - (2.262)(85\,333) = -55\,538$$

$$\text{límite superior} = 137\,500 + (2.262)(85\,333) = 330\,538$$

Por tanto, el intervalo se extiende de −55 538 a 330 538. ¿Qué significa el signo negativo? Recuerda que la diferencia que estamos estudiando es d = precio valuado − precio real, entonces un signo negativo indicará que el precio de valuación es menor que el real y uno positivo significa que el precio de valuación es mayor que el real. Por tanto, la valuación a veces subestima los bienes y a veces los sobreestima, y no podemos decir que exista una tendencia a sobreestimarlos.

PREGUNTAS

1. Conoces la media y la desviación estándar de una población. Seleccionas una muestra de esta población y calculas un intervalo de confianza al 90% para la media. ¿A cuántas desviaciones estándar a partir de la media se extiende este intervalo (en ambas direcciones)?

a) Respuesta _____.

2. Tienes una población de exámenes. Seleccionas una muestra de 11 para estimar la media y la desviación estándar de la población. Calculas un intervalo de confianza para la media. ¿A cuántas desviaciones estándar a partir de la media se extiende este intervalo (en ambas direcciones)?

a) Respuesta _____.

3. Tomas una muestra de 25 elementos de una población. La media de la muestra es de 38 y la desviación estándar de la población es de 6.5. Selecciona el intervalo de confianza para la media al 95 por ciento.

a) (37.49, 38.51).

b) (36.49, 39.51).

c) (35.45, 40.55).

d) (25.26, 50.74).

4. Obtienes una muestra de 9 elementos de una población. La media de la muestra es 49 y su desviación estándar es 4. Selecciona el intervalo de confianza para la media al 99%

a) (39.76, 58.24).

b) (44.53, 53.47).

c) (45.93, 52.07).

d) (47.51, 50.49).

Distribución *t*

Sabemos que en la distribución normal el 95% del área está comprendida entre ± 1.96 desviaciones estándar a partir de la media. Por tanto, si obtienes un valor a de una distribución normal con una media de 100, la probabilidad de que este valor a se encuentre dentro de la banda definida por $100 \pm 1.96\ \sigma$ es 0.95. Similarmente, si seleccionas n valores de la población, la probabilidad de que la media de la muestra ($\overline{X}$) se encuentre dentro de $\pm 1.96\ \sigma_{\overline{x}}$ a partir de la media es 0.95.

Ahora consideremos el caso en el que se tiene una distribución normal, pero no se conoce la desviación estándar. Seleccionas n valores, calculas la media de la muestra ($\overline{X}$) y estimas el error estándar de la media, $\sigma_{\overline{x}}$, con $s_{\overline{x}}$. ¿Cuál es la probabilidad de que $\overline{X}$ se encuentre dentro de la región delimitada por $\mu \pm 1.96\ s_{\overline{x}}$? Es un problema difícil, porque hay dos formas en las cuales $\overline{X}$ puede estar fuera de estos límites: (1) $\overline{X}$ puede ser muy grande o muy pequeña solamente debido al azar, y (2) $s_{\overline{x}}$ puede ser muy pequeño debido únicamente al azar. Intuitivamente, tiene sentido pensar que la probabilidad de que la media de la muestra se encuentre dentro de la banda comprendida entre $\mu \pm 1.96$ errores estándar de la media, debe ser menor que la que se tiene cuando la desviación estándar es conocida (y no es subestimada). Pero, ¿exactamente en cuánto es más pequeña? Afortunadamente, este problema fue resuelto a principios del siglo XX por W. S. Gossett, quien determinó la distribución de la media dividida por una estimación del error estándar. A esta distribución se le conoce como distribución *t* de student, o algunas veces sólo como distribución *t*. Gossett trabajó la distribución *t* y asoció pruebas estadísticas, mientras trabajaba en una cervecería en Irlanda. Debido a un acuerdo contractual con la cervecería, publicó el artículo bajo el seudónimo *"student*". Por eso, a la distribución se le conoce con el nombre de *t* de *student.*

La distribución *t* es muy similar a la distribución normal cuando la estimación de la varianza se calcula con muchos grados de libertad, pero tiene relativamente más área en las colas cuando se tienen pocos grados de libertad. En la figura 3.3 se muestra la distribución *t* para 4 grados de libertad y la distribución normal. Observa que la distribución normal tiene relativamente más área en el centro de la distribución y la *t* más en las colas. La distribución *t*, por tanto, es leptocúrtica.

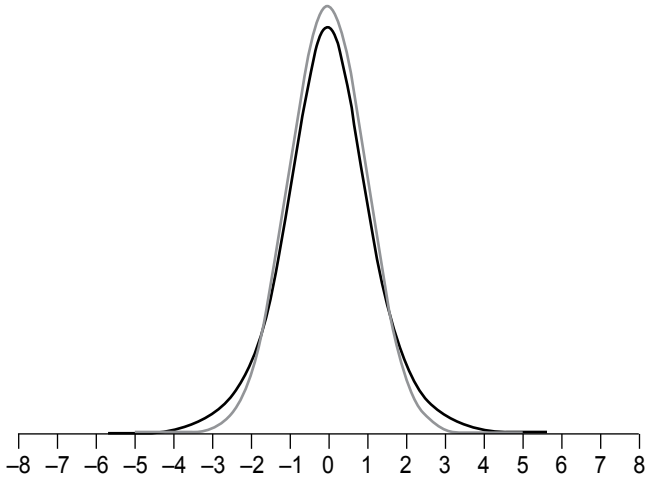


Figura 3.3 Una comparación de la distribución con 4 *gl* (en negro) y la distribución normal estándar (en gris).

gl	0.95	0.99
2	4.303	9.925
3	3.182	5.841
4	2.776	4.604
5	2.571	4.032
8	2.306	3.355
10	2.228	3.169
20	2.086	2.845
50	2.009	2.678
100	1.984	2.626

Tabla 3.3 Una tabla pequeña de *t*.

Ya que la distribución *t* es leptocúrtica, el porcentaje del área delimitada por la media ± 1.96 desviaciones estándar es menor que en la distribución normal (para la normal es 95%). La tabla 3.3 muestra el número de desviaciones estándar, a partir de la media, que se requieren para abarcar 95 y 99% del área de la distribución *t* para varios grados de libertad. Éstos son los valores de *t* que se usan para construir los intervalos de confianza. Los correspondientes valores para una distribución normal son 1.96 y 2.58, respectivamente. Observa que, cuando se tienen pocos grados de libertad, los valores de *t* son mucho más grandes que los correspondientes valores de la distribución normal, y esta diferencia disminuye al aumentar los grados de libertad.

Regresando al problema planteado al inicio de esta sección, en el que suponemos que se obtuvieron 9 valores de una población normal y se estimó el error estándar de la media, $\sigma_{\bar{x}}$ con $s_{\bar{x}}$, ¿cuál es la probabilidad de que $\overline{X}$ se encuentre en la región delimitada por $\mu \pm 1.96\ s_{\bar{x}}$? Ya que el tamaño de la muestra es 9, tenemos $n - 1 = 8$ *gl*. Utilizando la tabla 3.3 encontramos que para 8 *gl*, la probabilidad de que la media se encuentre entre $\mu \pm 2.306\ s_{\bar{x}}$ es 0.95. Por lo que resulta lógico que la probabilidad de que la media se encuentre entre $\mu \pm 1.96 s_{\overline{X}}$ será menor a 0.95. Utilizando la tabla de *t* para 8 grados de libertad al valor más aproximado a 1.96, le corresponde un área de 0.05 en una cola (0.01 en las dos colas) por lo que aproximadamente el área entre $\mu \pm 1.96 s_{\overline{X}}$ será igual a $1-0.1 = 0.9$

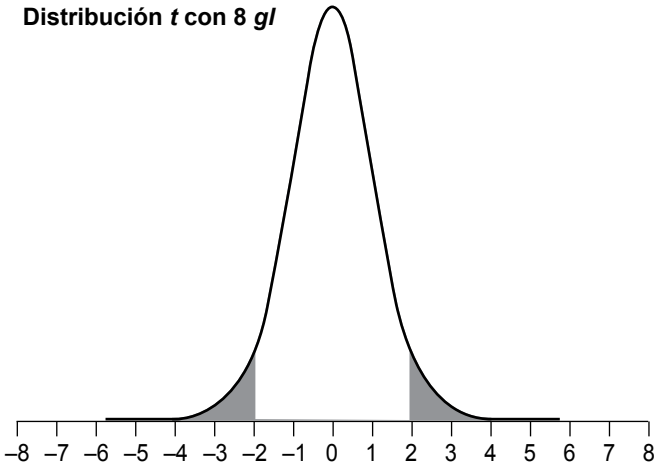


Figura 3.4 Área mayor a ± 1.96 desviaciones estándar a partir de la media para una distribución *t* con 8 *gl*.

Como se esperaba, esta probabilidad es menor que 0.95, que es el valor que se hubiera obtenido si $\sigma_{\bar{x}}$ hubiera sido conocida en lugar de estimada.

PREGUNTAS

1. Selecciona las descripciones correctas, acerca de la distribución *t:*

a) tiene mayor densidad en las colas, que una distribución normal.

b) tiene mayor densidad en el centro, que una distribución normal.

c) es leptocúrtica.

d) la usas cuando no conoces la desviación estándar de la población.

e) una distribución *t* con 20 grados de libertad, tiene 95% de su área entre la media ± 1.96 desviaciones estándar.

2. Una distribución *t*, ¿con cuáles de los siguientes grados de libertad se aproxima más a una distribución normal?

a) 0.

b) 2.

c) 12.

d) 50.

3. Para una distribución *t* con 15 grados de libertad, ¿cuántas desviaciones estándar, a partir de la media en ambas direcciones, se tiene que recorrer para abarcar el 90% de la distribución?

a) Respuesta _____.

4. En una distribución *t* con 10 grados de libertad, ¿cuál es la probabilidad de obtener un valor que se encuentre en la banda definida por la media ± 2 desviaciones estándar?

a) Respuesta _____.

5. Hay una población con desviación estándar desconocida. Seleccionas 21 elementos de esta población y calculas su media y su desviación estándar. Obtienes un valor para la media que es 1.5 veces el error estándar más grande de lo que supones que es el valor de la media de la población, ¿cuál es la probabilidad de obtener un valor que se aleje 1.5 desviaciones estándar o más de la media de esta distribución *t*?

a) Respuesta _____.

Intervalo de confianza para la diferencia de medias

Es más común que los investigadores estén interesados en la diferencia entre las medias que en sus valores específicos. Veamos un ejemplo: se hizo una encuesta para determinar la cantidad de cervezas (número de cervezas) que consumen los viernes los estudiantes de una universidad. En la tabla 3.4 se muestran los tamaños de muestra, medias y varianzas para una muestra de hombres y una de mujeres.

Género	*n*	Media	Varianza
Hombres	17	5.353	2.743
Mujeres	17	3.882	2.985

Tabla 3.4 Medias y varianzas de una encuesta referente al consumo de cerveza.

Como puedes ver, los hombres consumen mayor cantidad de cerveza en promedio. La diferencia entre la media de los hombres de 5.35 y la media de las mujeres de 3.88 es 1.47. La diferencia en el consumo de cerveza entre el género en esta muestra particular no es de nuestro interés. Lo que es importante es la diferencia en la población. La diferencia entre las medias muestrales se usa para estimar la diferencia entre las medias poblacionales. La precisión de la estimación se da a conocer mediante el intervalo de confianza.

Para calcular el intervalo de confianza, se hacen tres suposiciones:

1. Las dos poblaciones tienen la misma varianza. Esta suposición se conoce como el supuesto de homogeneidad u homocedasticidad de las varianzas.
2. Las poblaciones se distribuyen en forma normal.
3. Cada valor se obtiene por muestreo en forma independiente de cualquier otro valor.

Las consecuencias de violar estas suposiciones se discutirán más tarde. Por ahora, es suficiente decir que moderadas violaciones a los supuestos 1 y 2 no traen consecuencias graves en el análisis de los datos.

Se usan las siguientes fórmulas para calcular un intervalo de confianza para la diferencia entre medias:

$$\text{límite superior} = (\overline{X}_1 - \overline{X}_2) + t_{\%c,gl} s_{\overline{x}_1 - \overline{x}_2}$$

$$\text{límite inferior} = (\overline{X}_1 - \overline{X}_2) - t_{\%c,gl} s_{\overline{x}_1 - \overline{x}_2}$$

donde $\overline{X}_1 - \overline{X}_2$ es la diferencia entre las medias muestrales, $t_{\%c,gl}$ es el valor de t para un nivel de confianza deseado y los grados de libertad correspondientes, y $s_{\overline{x}_1 - \overline{x}_2}$ es la estimación del error estándar para la diferencia entre medias. Lo que significan estos términos se aclarará con los cálculos del ejemplo.

Utilizando los datos de la encuesta, calculamos un intervalo de confianza para la diferencia entre el consumo medio de los hombres y las mujeres. Para realizar este cálculo, suponemos que las varianzas en las dos poblaciones son iguales, es decir, que se cumple el supuesto de homogeneidad de las varianzas.

El primer paso consiste en calcular un estimador del error estándar de la diferencia entre las medias, $s_{\overline{x}_1 - \overline{x}_2}$. Recuerda que en el capítulo sobre distribuciones muestrales, vimos que la fórmula para el error estándar de la diferencia entre medias poblacionales viene dado por:

$$\sigma_{\overline{x}_1 - \overline{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}} = \sqrt{\frac{2\sigma^2}{n}}$$

Con el fin de estimar este valor, estimamos σ^2. Ya que suponemos que las varianzas de las poblaciones son iguales, estimamos esta varianza con el promedio ponderado de nuestras dos muestras. Entonces, la estimación de la varianza se calcula usando la siguiente fórmula:

$$CME = \frac{s_1^2 + s_2^2}{2}$$

donde *CME*, cuadrado medio del error, es nuestra estimación de σ^2. En este ejemplo,

$$CME = (2.743 + 2.985)/2 = 2.864.$$

Debido a que n (el tamaño de muestra para cada grupo) es 17,

$$s_{\overline{x}_1 - \overline{x}_2} = \sqrt{\frac{2CME}{n}} = \sqrt{\frac{(2)(2.864)}{17}} = 0.5805$$

El siguiente paso consiste en encontrar el valor de *t*, que se utilizará para calcular el intervalo de confianza ($t_{\%c,gl}$). Para encontrar el valor de $t_{\%c,gl}$, necesitamos conocer los grados de libertad. Los grados de libertad son el número de estimaciones independientes de la varianza con las que se calculó el valor del *CME*. Los grados de libertad son iguales a $(n_1 - 1) + (n_2 - 1)$, donde n_1 es el tamaño de la muestra del primer grupo y n_2 es el tamaño de la muestra del segundo grupo. Para nuestro ejemplo, $n_1 = n_2 = 17$. Cuando $n_1 = n_2$, en forma convencional se usa *n* para referirse al tamaño de la muestra de cada grupo. Los grados de libertad para nuestro ejemplo son $16 + 16 = 32$.

Usando la tabla de *t*, encontramos que el valor de *t* para un 95% de confianza y 32 gl es 2.0369.

Ahora tenemos todos los componentes necesarios para calcular el intervalo de confianza. Primero, calculamos la diferencia entre las medias:

$$\overline{X}_1 - \overline{X}_2 = 5.3523 - 3.8824 = 1.470$$

Calculamos el error estándar de la diferencia entre las medias:

$$S_{\overline{x}_1 - \overline{x}_2} = 0.5805$$

dado que el valor de *t* para un 95% de confianza y 32 gl es:

$$t_{\%c,gl} = 2.037$$

Entonces, el intervalo de confianza al 95% es

$$\text{límite inferior} = 1.470 - (2.037)(0.5805) = 0.29$$

$$\text{límite superior} = 1.470 + (2.037)(0.5805) = 2.65$$

Escribimos el intervalo de confianza como:

$$0.29 \leq \mu_{\text{mujeres}} - \mu_{\text{hombres}} \leq 2.65$$

Este análisis proporciona pruebas de que la media del consumo de los hombres es más alta que la de las mujeres, y que se tiene una confianza del 95% que la diferencia entre las medias de consumo en la población se encuentre entre 0.3 y 2.7 cervezas.

La mayoría de los programas estadísticos en computadora requieren que los datos se proporcionen en una forma específica a fin de procesar las pruebas de *t*. Considera los datos de la tabla 3.5.

Grupo 1	Grupo 2
3	5
4	6
5	7

Tabla 3.5 Datos para identificar.

Hay dos grupos, cada uno con tres observaciones. El formato de los datos para su uso en computadora, por lo general, consiste en identificar los datos del primer grupo con un 1 y los del segundo con un 2, como se muestra en la tabla 3.5.

Grupo 1	Grupo 2
1	3
1	4
1	5
2	5
2	6
2	7

Tabla 3.6 Datos identificados por grupo.

Los cálculos cambian cuando los tamaños de muestra son diferentes. Una suposición importante es considerar que en la estimación de la varianza tiene mayor influencia la muestra de mayor tamaño. Se calcula la suma de las desviaciones cuadradas o la suma de cuadrados del error como sigue:

$$SCE = \sum(X - \overline{X}_1)^2 + \sum(X - \overline{X}_2)^2$$

donde $\overline{X}_1$ es la media del grupo 1 y $\overline{X}_1$ es la media del grupo 2. Consideremos un ejemplo con muy pocos datos para realizar en forma rápida los cálculos (véase la tabla 3.7):

Grupo 1	Grupo 2
3	2
4	4
5	

Tabla 3.7 Datos de ejemplo.

$$\overline{X}_1 = 4 \text{ y } \overline{X}_1 = 3$$

$$SCE = (3-4)^2 + (4-4)^2 + (5-4)^2 + (2-3)^2 + (4-3)^2 = 4$$

Entonces, el cuadrado medio del error, *CME* se calcula de la forma siguiente: $CME = SCE/gl$, donde los grados de libertad (gl) se calculan como antes:

$$gl = (n_1 - 1) + (n_2 - 1) = (3-1) + (2-1) = 3$$

$$CME = SCE/gl = 4/3 = 1.333$$

La fórmula

$$S_{\overline{x}_1 - \overline{x}_2} = \sqrt{\frac{2CME}{n}}$$

se reemplaza por:

$$S_{\overline{x}_1 - \overline{x}_2} = \sqrt{CME\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Por tanto

$$S_{\overline{x}_1 - \overline{x}_2} = \sqrt{1.333\left(\frac{1}{3} + \frac{1}{2}\right)} = 1.054$$

$t_{\%c,gl}$ para 3 gl y un nivel de confianza de 0.95 es igual a 3.182.

Entonces el intervalo de confianza al 95% es

$$\text{límite inferior} = 1 - (3.182)(1.054) = -2.35$$

$$\text{límite superior} = 1 + (3.182)(1.054) = 4.35$$

Entonces escribimos el intervalo de confianza como sigue:

$$-2.35 \leq \mu_1 - \mu_2 \leq 4.35$$

PREGUNTAS

1. Selecciona todas las suposiciones que se necesitan para construir un intervalo de confianza para la diferencia entre medias.

a) al menos se muestrea el 2% de la población.

b) las muestras son independientes.

c) las varianzas son homogéneas.

d) las poblaciones se distribuyen en forma normal.

2. Comparas las horas que se pasan viendo la televisión los hombres y las mujeres. Seleccionas a 12 hombres y a 14 mujeres y calculas un intervalo de confianza para la diferencia entre las medias. ¿Cuántos grados de libertad tiene el valor de "*t*"?

a) Respuesta _____.

3. Comparas las horas que pasan estudiando por día los estudiantes de primer semestre y los del último semestre. Seleccionas una muestra de 11 personas de cada grupo. Para los del primer semestre la media fue de 3, con una varianza igual a 1.2 y para los del último semestre la media fue de 2 con una varianza de 1. Calculas un intervalo de confianza al 90% para la diferencia entre medias (primer semestre-último semestre). ¿Cuál es el límite inferior del intervalo?

a) Respuesta _____.

4. Se realizó un examen a los alumnos del primero y segundo cursos de estadística para comparar su desempeño. Cinco estudiantes del primer curso obtuvieron los siguientes puntajes: 4, 3, 5, 7, 4, y los cinco estudiantes del segundo curso obtuvieron los siguientes resultados: 7, 9, 8, 6, 9. Calcula un intervalo de confianza al 95% para la diferencia entre las medias (segundo curso-primer curso). ¿Cuál es el límite superior del intervalo de confianza?

a) Respuesta _____.

Intervalo de confianza para el coeficiente de correlación de Pearson

El cálculo de un intervalo de confianza para el coeficiente de correlación poblacional, ρ, es complicado debido a que la distribución muestral de r no se distribuye en forma normal. La solución a este problema se subsana mediante la transformación de Fisher, z', descrita en la sección de la distribución muestral del coeficiente de correlación r de Pearson. Los pasos a seguir para calcular un intervalo de confianza para ρ son:

1. Transformar r en z'.
2. Calcular un intervalo de confianza para z'.
3. Transformar el intervalo de confianza a r.

Tomemos como ejemplo la problemática del transporte en la FES-C. En este estudio se correlacionó el grado de aceptación de un aumento en las tarifas y la calidad del servicio que proporcionan los permisionarios. Seguramente esperas una relación positiva entre estas dos variables: cuantos más estudiantes acepten el aumento a las tarifas, más estudiantes calificarán como buena la calidad del servicio. Al realizar el estudio, tal como se esperaba, se encontró una relación positiva entre estas dos variables: cuantos más estudiantes aceptaban el aumento, más opinaban que la

calidad proporcionada era buena (aunque la mayoría se resiste a pagar más por el servicio). La correlación utilizando una muestra de 34 estudiantes fue de 0.654. El problema ahora es calcular un intervalo de confianza al 95% para ρ, utilizando el valor calculado de $r = 0.654$.

La conversión de r a z' puede hacerse utilizando las tablas. La tabla contiene únicamente valores positivos de r, pero esto no es problema. El valor de z' asociado con un coeficiente de 0.654 es 0.78. Entonces el valor de z' asociado con un coeficiente r de 0.654 es 0.78.

La distribución muestral de z' es aproximadamente normal y tiene un error estándar de

$$\frac{1}{\sqrt{n-3}}$$

Para nuestro ejemplo, $n = 34$ y por tanto el error estándar es 0.180. El valor de z para un intervalo de confianza al 95% ($z_{0.95}$) es 1.96 que se puede encontrar en tablas de la distribución normal. El intervalo de confianza lo calculamos como sigue:

$$\text{límite inferior} = 0.78 - (1.96)(0.18) = 0.43$$

$$\text{límite superior} = 0.78 + (1.96)(0.18) = 1.13$$

El paso final es convertir este intervalo en un intervalo de r usando las tablas. El coeficiente r asociado con un valor de z' de 1.13 es 0.81 y el coeficiente r asociado con un valor de z' de 0.43 es 0.40. Entonces, el coeficiente de correlación poblacional, ρ, se encontrará entre 0.40 y 0.81 con un 95% de confianza.

Por tanto, podemos escribir el intervalo de confianza al 95% como:

$$0.40 \leq \rho \leq 0.81$$

Para calcular un intervalo de confianza al 99%, encontramos que z es igual a 2.58 y entonces:

$$\text{límite superior} = 0.78 + (2.58)(0.18) = 1.24$$

$$\text{límite inferior} = 0.78 - (2.58)(0.18) = 0.32$$

Convirtiendo estos valores a r, el intervalo de confianza es:

$$0.31 \leq \rho \leq 0.84$$

El intervalo de confianza al 99% es más ancho que el intervalo al 95%.

PREGUNTAS

1. Selecciona los intervalos de confianza para el coeficiente de correlación poblacional, que sean posibles.

a) (–0.4, 0.6).

b) (0.3, 0.5).

c) (–0.85, –0.47).

d) (0.72, 1.2).

2. Se selecciona una muestra de 28 elementos de una población, y se calcula el coeficiente de correlación, y resulta $r = 0.45$, ¿cuál es el intervalo de confianza al 95% para el coeficiente de correlación poblacional?

a) (0.58, 0.842).

b) (0.093, 0.877).

c) (0.058, 0.687).

d) (0.093, 0.705).

3. El coeficiente de correlación muestral es –0.8. Si el tamaño de la muestra fue de 40, entonces el intervalo de confianza al 99% para el coeficiente de correlación poblacional está entre –0.909 y _____.

Intervalo de confianza para la proporción

Limpia Todo, S. A. es una empresa que se dedica a la elaboración de productos de limpieza y acaba de desarrollar un nuevo aromatizante para el hogar. Un estudio para dar a conocer el producto consideró una muestra de 500 amas de casa; después de cierto periodo de prueba, se determinó que 260 aprobaban el producto. En otras palabras, a una proporción muestral de 0.52 le gustó el producto. Aunque esta estimación puntual de la proporción nos da información, es también importante calcular un intervalo de confianza. El intervalo de confianza se calcula con base en la media y la desviación estándar de la distribución muestral de la proporción. Las fórmulas para determinar estos dos parámetros son:

$$\mu_p = \pi$$

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$

Ya que no se conoce el parámetro de la población π, usaremos la proporción de la muestra p como un estimador. El error estándar de p es, por tanto:

$$S_p = \sqrt{\frac{p(1-p)}{n}}$$

Empezamos tomando nuestro estadístico, p, y construyendo un intervalo que se extienda a partir de la media en ambas direcciones una distancia igual a $(z_{0.95})(s_p)$. Es decir, $z_{0.95}$, en una distribución normal, es el número de desviaciones estándar a partir de la media que se tiene que recorrer en ambas direcciones para incluir el 0.95 del área total (véase la sección sobre el intervalo de confianza para la media). El valor de $z_{0.95}$ se localiza en tablas de la distribución normal y es igual a 1.96. Se hace un ajuste leve para corregir el hecho de que la distribución es discreta.

s_p se calcula de la siguiente manera:

$$s_p = \sqrt{\frac{0.52(1-0.52)}{500}} = 0.0223$$

Para corregir el hecho de que aproximamos una distribución discreta con una continua (distribución normal), restamos $0.5/n$ al límite inferior y sumamos $0.5/n$ al límite superior. Entonces, el intervalo de confianza es:

$$p \pm Z_{0.95}\sqrt{\frac{p(1-p)}{n}} \pm \frac{0.5}{n}$$

límite inferior: $0.52 - (1.96)(0.0223) - 0.001 = 0.475$

límite superior: $0.52 + (1.96)(0.0223) + 0.001 = 0.565$

$$0.475 \leq \pi \leq 0.565$$

En términos porcentuales, podemos decir que entre 47.5 y 56.5% de las amas de casa consumirían el producto con un margen de error de 4.5%. Debes tener presente que el margen de error es el margen para el porcentaje de amas de casa a las cuales les gustó el aromatizante y no el margen de error para la diferencia entre el porcentaje que prefiere al producto y el porcentaje de amas de casa que no lo prefieren. El margen de error para las diferencias es 9%, el doble del margen de error para el porcentaje individual. Ten en cuenta esto cuando escuches o leas los informes, que por lo general lo interpretan en forma equivocada.

PREGUNTAS

1. ¿Por qué restamos $0.5/n$ al límite inferior y sumamos $0.5/n$ al límite superior cuando calculamos un intervalo de confianza para la proporción poblacional?

a) porque necesitamos esta corrección debido a que estamos aproximando una distribución discreta (distribución muestral de p) con una distribución continua (distribución normal).

b) necesitamos corregir el hecho de que el estimador de la proporción poblacional tiene un ligero sesgo.

c) necesitamos corregir el hecho de que el estimador del error estándar de la proporción tiene un ligero sesgo.

2. El periódico *Reforma* llevó a cabo una encuesta para determinar quién es el candidato de mayor preferencia a gobernador del Estado de México. Los encuestadores calcularon un intervalo de confianza al 95% y encontraron que el porcentaje de votantes en Cuautitlán Izcalli, que muestran preferencia por el candidato A, está entre 51 y 59%. ¿Cuál es el margen de error (como porcentaje)?

a) Respuesta _____.

3. Un investigador está interesado en saber cuánta gente de la ciudad está de acuerdo con un nuevo impuesto. Selecciona una muestra de 100 personas de la localidad y encuentra que el 40% está de acuerdo, ¿cuál es el límite superior del intervalo de confianza al 95% para la proporción poblacional?

a) Respuesta _____.

Actividades

I. Contesta y resuelve los siguientes ejercicios para reafirmar los conceptos.

1. ¿Cuándo se podría considerar el promedio del examen final de una asignatura de un grado escolar como estadístico? ¿Cuándo se podría considerar como un parámetro?

2. Define el sesgo en términos del valor esperado.

3. ¿Es posible que un estimador sea insesgado pero muy impreciso? ¿Cómo sería un estimador preciso pero sesgado?

4. ¿Por qué un intervalo de confianza al 99% es más ancho que un intervalo de confianza al 95 por ciento?

5. ¿Cuándo se construye un intervalo de confianza al 95%, al que le tienes 95% de confianza?

6. ¿Cuál es la diferencia en el cálculo de los intervalos de confianza cuando se conoce la desviación estándar de la población y cuando no se conoce y se tiene que estimar?

7. ¿Cuál es el efecto del tamaño de la muestra en el ancho del intervalo de confianza?

8. ¿Cómo es la distribución *t* en comparación con la distribución normal? ¿Cómo afecta esta diferencia al tamaño de los intervalos de confianza? ¿Es el tamaño de la muestra determinante en esta diferencia?

9. Se investigó la efectividad de una droga para controlar la presión de la sangre. ¿Cómo puede un investigador demostrar que la reducción promedio en la presión sistólica es de 20 unidades o más?

10. Se conoce que una población se distribuye normalmente con una desviación estándar de 2.8. *a*) Calcula un intervalo de confianza al 95% para la media, si una muestra de 9 elementos proporcionó los siguientes resultados: 8, 9, 10, 13, 14, 16, 17, 20 y 21. *b*) Ahora calcula un intervalo de confianza al 99% usando los mismos datos.

11. Una persona afirma que puede predecir el resultado del lanzamiento de una moneda. Acierta 16 de 25 lanzamientos. Calcula un intervalo de confianza al 95% para la proporción de veces que esta persona puede predecir en forma correcta el resultado. ¿Qué puedes concluir acerca de la habilidad que tiene la persona en predecir el futuro?

12. ¿Qué significa que la varianza (calculada con el divisor *N*) sea un estimador sesgado?

13. Un intervalo de confianza para la media de la población calculada a partir de un tamaño de muestral de 16 fue de [12 – 28]. Se tomará una nueva muestra de 36 observaciones. No se puede saber de antemano cuál será el intervalo de confianza porque éste depende de la muestra y ésta es aleatoria. Sin embargo, debes tener una idea de cómo será el intervalo. Proporciona las ideas que tengas acerca de esta nueva estimación.

14. Se selecciona una muestra de 22 observaciones de una población y su media fue 60. *a*) Se conoce que la desviación estándar de la población es 10. Calcula el intervalo de confianza al 99% para la media de la población. *b*) Ahora, supón que no se conoce la desviación estándar de la población, pero que la desviación estándar de la muestra es 10. Calcula de nuevo el intervalo de confianza al 99% para la media de la población.

15. Leíste la encuesta de un periódico donde 70% de 250 personas entrevistadas mostraron preferencia por el candidato A. Te muestras sorprendido porque piensas que lo más probable es que el 50% de la población prefiera a este candidato. Con base en la muestra seleccionada por el diario, ¿es posible que la proporción en la población de las personas con preferencia al candidato A sea 50%? Calcula un intervalo de confianza al 95% para fundamentar tu respuesta.

16. Se estudiaron las estaturas de adolescentes hombres y mujeres. La estatura promedio para una muestra de 12 muchachos fue de 174 cm con una varianza de 62. Para una muestra de 12 muchachas, el promedio fue de 166 cm con una varianza de 65. *a*) Calcula un intervalo de confianza al 95% para la diferencia de las medias poblacionales. *b*) Calcula un intervalo de confianza al 99% para la diferencia de las medias poblacionales. *c*) ¿Puedes pensar que la diferencia de las medias puede ser alrededor de 5? ¿Por qué sí o por qué no?

17. Se desea conocer cuánto tiempo en promedio un estudiante de Estadística estudia esta asignatura durante la noche. Se les hizo esta pregunta a 10 estudiantes y los resultados en horas fueron los siguientes: 2, 1.5, 3, 2, 3.5, 1, 0.5, 3, 2 y 4.

a) Encuentra un intervalo de confianza al 95% para la media de la población.

b) Encuentra un intervalo de confianza al 99% para la media de la población.

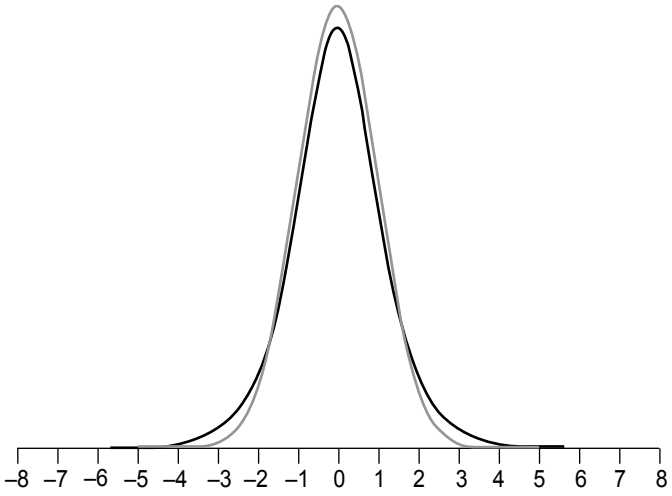
18. Verdadero/Falso: Según aumenta el tamaño de la muestra, la probabilidad de que el intervalo de confianza contenga a la media de la población es más alta.

19. Verdadero/Falso: Tienes una muestra de 9 hombres y una muestra de 8 mujeres, los grados de libertad para el valor de *t* del intervalo de confianza de la diferencia entre medias es 16.

20. Verdadero/Falso: Las letras griegas se usan para los estimadores.

21. Verdadero/Falso: Para construir un intervalo de confianza para la diferencia entre medias es necesario suponer que las poblaciones tienen la misma varianza y que se distribuyen normalmente.

22. Verdadero/Falso: La distribución más alta representa a la distribución *t* y la otra a la distribución normal.



II. Resuelve los siguientes ejercicios de aplicación.

1. Después de tomar una muestra aleatoria de 80 profesores, se descubre que 30 de ellos no votaron en las elecciones federales pasadas. Estima en forma puntual la proporción de todos los profesores que no votaron en dichas elecciones.

2. Una muestra de 200 consumidores reveló que 166 prefieren la pasta dental marca Colgate. Realiza una estimación puntual de la proporción de consumidores que prefieren dicha marca de pasta dental.

3. Para estimar el gasto diario de los estudiantes por concepto de transporte se tomó una muestra y se encontraron los siguientes resultados (en pesos): 10, 15, 25, 20, 25, 15, 28, 9, 10, 15. Determina una estimación puntual para la media y la varianza del gasto de transporte de los estudiantes.

4. Una muestra de golosinas que se expenden en una cooperativa escolar tienen los siguientes precios en pesos: 0.80, 1.50, 8.00, 4.00, 2.50, 1.00, 0.50, 3.50, 6.00, 12.00, 3.00.

a) determina una estimación puntal de la proporción de golosinas que cuestan por arriba de los 6 pesos.

b) determina una estimación puntual del precio promedio de las golosinas.

5. En una muestra de 200 profesores del estado de Oaxaca, 184 expresaron una insatisfacción extrema ante el plan presentado por las autoridades para modificar las condiciones de trabajo. Determina una estimación puntual de la proporción de profesores que están en total desacuerdo con las autoridades.

6. Una pizzería con entregas a domicilio ha prosperado mucho, pues entrega las pizzas en muy poco tiempo. La pizzería garantiza que sus productos se recibirán en 30 minutos o menos después de hacer el pedido; si la entrega se atrasa, la pizza es gratis. El tiempo que se tarda en realizar cada pedido surtido puntualmente se anota en el registro de la pizzería, y el tiempo de entrega de los pedidos que se surten con retraso se anota como 30 minutos en el registro. A continuación se incluyen 12 anotaciones aleatorias del registro: 15.3, 10.8, 29.5, 12.2, 30.0, 14.8, 10.1, 30.0, 30.0, 22.1, 19.6, 18.3

a) determina en forma puntual el tiempo promedio de entrega de la muestra anterior.

b) ¿de qué población se extrajo esta muestra?

c) ¿puede esta muestra emplearse para estimar el tiempo promedio que tarda la pizzería en entregar un pedido? Explica tu respuesta.

7. El departamento de personal de una fábrica sabe que la desviación estándar de llegada tarde de los obreros al trabajo es de 5.1 minutos. Se toma una muestra de 40 obreros y se encuentra que en promedio llegan 12.5 minutos tarde al trabajo.

a) encuentra el error estándar de la media.

b) establece una estimación por intervalo alrededor de la media, usando el error estándar de la media.

8. La Universidad está realizando un estudio socioeconómico sobre el ingreso promedio familiar de los estudiantes que solicitaron beca de Pronabes. Se enviaron trabajadores sociales a los hogares de una muestra de 200 solicitantes. Se encontró un ingreso mensual promedio familiar de 4 200 pesos. La desviación estándar, según un estudio anterior es de 800 pesos. ¿Cuál es el intervalo alrededor de la media muestral que incluirá la media de la población 95.44% de las veces?

9. En una muestra de 36 cajeros de Teléfonos de México se encontró que recibían pagos de 14 300 pesos en promedio por día. Si la variancia de la población es de 20 700 pesos por día:

a) encuentra el error estándar de la media.

b) establece una estimación por intervalo que incluya la media de la población 68.26% de las veces.

10. El dueño de un restaurante recién inaugurado ha tenido problemas al estimar la cantidad de comida que debe prepararse cada día; por eso, decidió determinar el número promedio de clientes a quienes atiende cada noche. Seleccionó una muestra de 25 noches, que dio por resultado una media de 70 clientes. La desviación estándar de la población ha sido establecida como 23.8 clientes. Determina un intervalo de confianza al 95% para estimar la media verdadera de clientes que acuden al restaurante.

11. El gerente del departamento de crédito de Fabricantes Muebleros, S. A. (FAMSA), desea estimar los saldos deudores de sus clientes que no han mostrado falta de pago, con la finalidad de otorgarles una nueva línea de crédito. Tomó una muestra de 100 clientes con estas características y encontró una deuda promedio de 5 400 pesos con una desviación estándar de 1 900 pesos.

a) determina un intervalo de confianza al 90% para estimar la deuda real promedio de los clientes que siempre han pagado.

b) supón que se tiene una población de 810 clientes con estas características. Determina un intervalo de confianza al 99% para estimar la deuda real promedio.

12. Verónica Velazco es una universitaria con gran sentido del ahorro y quiere comprar un automóvil compacto usado. Aleatoriamente seleccionó 125

anuncios en el "Aviso Oportuno" y descubrió que el precio promedio de un automóvil en esa muestra era de 48 845 pesos, con una desviación estándar de 14 275 pesos.

a) establece una estimación por intervalo al 98% para el precio promedio real de un automóvil compacto usado.

b) establece una estimación por intervalo del precio promedio de un automóvil, para que la señorita Velazco tenga una confianza de 95% de que la media de la población se hallará dentro de ese intervalo.

13. El departamento de servicios escolares considera que el número promedio de estudiantes por grupo para las licenciaturas de Administración y Contaduría es muy importante para optimizar las instalaciones de la Facultad. Para estimar esta variable tomó una muestra de 40 grupos y encontró un promedio de 39.8 alumnos por grupo con una desviación estándar de 9.1 alumnos por grupo. Construye un intervalo de confianza al 99% para estimar el promedio real de estudiantes por grupo.

14. Supón que deseamos utilizar un nivel de confianza de 80%. Da el límite superior del intervalo de confianza en función de una media muestral, y el error estándar.

15. En una muestra de 50 estudiantes del CCH Azcapotzalco se obtuvo que utilizaban Internet en promedio 12 horas por semana, con una desviación estándar de 4.5 horas. Establece una estimación por intervalo al 95% para la media verdadera del tiempo de uso de Internet.

16. El gerente de la división de focos de General Electric debe estimar el número promedio de horas que durará un foco fabricado con nuevos materiales. Una muestra de 50 focos de este tipo mostró un promedio de vida de 840 horas con una desviación estándar de 35 horas. Construye un intervalo de confianza de 90% para la verdadera media de la población.

17. Juan Garibay, un graduado sumamente escrupuloso, acaba de terminar el primer borrador de su tesis de 200 páginas. Mandó capturar su trabajo y quiere conocer el número promedio de errores ortográficos por página, pero no quiere leer toda la tesis, seleccionó al azar 40 páginas para leerlas y descubrió que el número promedio de errores por página es 4.3, mientras que la desviación estándar de la muestra es 1.2 errores por página.

a) calcula el error estándar estimado de la media.

b) construye un intervalo de confianza de 90% para el valor promedio verdadero de errores por página en la tesis de Juan.

18. En una muestra de 35 profesores de una población de 360, se descubre que la media en años de antigüedad es de 20.9 años con una desviación estándar de 72 meses.

a) calcula el error estándar estimado de la media.

b) construye un intervalo de confianza de 96% para la media.

19. Un corredor de la Bolsa Mexicana de Valores desea saber el tiempo que transcurre entre la colocación y la ejecución de una orden del mercado. Muestreó 45 órdenes y descubrió que el tiempo medio de ejecución era de 24.3 minutos con una desviación estándar de 3.2 minutos. Construye un intervalo de confianza de 95% para el tiempo medio de ejecución.

20. Al gerente de producción de Jumex, S. A, le preocupa que en los tres últimos años las heladas hayan dañado los plantíos que posee la empresa. Con el objetivo de averiguar el posible daño causado a los árboles, ha muestreado el número de naranjas producidas por árbol en 42 de ellos y ha observado que la producción promedio fue de 525 naranjas por árbol, con una desviación estándar de 30 naranjas por árbol.

a) estima el error estándar de la media.

b) construye un intervalo de confianza de 98% para la producción media por árbol de las naranjas.

21. Si el rendimiento promedio de naranjas por árbol fue de 600 naranjas el año pasado, ¿qué puede decir el gerente sobre la posible existencia de un daño en este momento?

22. Un jefe policiaco de la delegación Tlalpan implantó medidas enérgicas contra los distribuidores de droga en esa zona del Distrito Federal. Desde que las implantó, han sido atrapados 525 distribuidores de los que se tienen detectados. El valor medio en pesos de las drogas confiscadas a esos individuos es de 15 000 000 de pesos. La desviación estándar del valor monetario de las drogas es de 350 000 pesos. Constrúyele al jefe de policía un intervalo de confianza de 90% para el valor monetario medio que tienen las drogas de los distribuidores.

23. Una proveedora de repuestos para computadora que compra tarjetas de memoria DDR al mayoreo y sin probar está pensando en cambiar de proveedor y en acudir a uno que le ofrezca las tarjetas de memoria probadas y garantizadas, aunque a un precio más alto. Con el objeto de decidir si su plan es rentable, debe determinar la proporción de tarjetas defectuosas que le surte el actual proveedor. Probó una muestra de 200 tarjetas y descubrió que 5% están defectuosas. Construye un intervalo de confianza de 98% para la proporción de tarjetas de memoria defectuosas.

24. Una muestra de 80 personas de la tercera edad fueron entrevistadas con la finalidad de estimar la proporción de ellas que estaban a favor del nuevo programa de ayuda de Marcelo Ebrard Casaubon, y se encontró que 51 votaron por dicho candidato. Encuentra los límites superior e inferior de la proporción de personas que están a favor del nuevo programa de ayuda.

25. El gerente de producto de un nuevo aderezo para ensaladas está preocupado por las ventas tan bajas del producto y por el futuro de este último. Ante la sospecha de que la estrategia de mercado no hubiera identificado bien los atributos del producto, muestreó a 1 500 consumidores y se enteró de que 956 pensaban que el producto era un complemento para postres. Construye un intervalo de confianza de 96% para la verdadera proporción de la población que desconocía los atributos del producto.

26. Un jugador profesional de futbol lanzó 150 tiros libres y metió 126 goles.

a) construye un intervalo de confianza de 98% para la proporción de tiros libres que convierte en goles.

27. La Secretaría de Educación Pública (SEP) encuestó a una muestra de 500 alumnos que terminaron su ciclo de primaria y descubrió que 42% de ellos eran incapaces de sumar correctamente los quebrados. Construye un intervalo de confianza de 99% para la verdadera proporción de alumnos que terminaron su ciclo de primaria que no saben sumar quebrados.

28. Una sucursal de Elektra revisó aleatoriamente 200 de las 3 800 cuentas de crédito de las que actualmente tiene y determinó que 60% se hallaban en excelente estado (cuentas que no presentan pagos vencidos).

a) encuentra un intervalo de confianza de 95% para la proporción de cuentas en excelente estado.

b) encuentra un intervalo de confianza de 95% para la proporción de cuentas que presentan problemas de falta de pagos.

29. Durante el último año, las ventas han ido disminuyendo constantemente en una cadena de 1 500 restaurantes de comida rápida. Se determinó que 30% de una muestra de 95 restaurantes mostraba claros indicios de una mala administración. Construye un intervalo de confianza de 98% para esta proporción.

30. El responsable de la biblioteca está preocupado por la cantidad de libros que presentan daños por el uso de los estudiantes. Tomó una muestra de 45 libros que los estudiantes entregaron en el día y descubrió que 18 presentaban algún tipo de daño. Da un intervalo de la proporción de libros con daños que ofrezcan una seguridad de 96% de contener la verdadera proporción.

31. Un conocido analista financiero de la Bolsa Mexicana de Valores desea estimar la proporción de accionistas que planean vender por lo menos la cuarta parte de sus acciones el próximo mes. Para tal efecto, tomó una muestra aleatoria de 800 individuos que poseen acciones y encontró que una octava parte proyecta vender al menos una cuarta parte de ellas el próximo mes. El analista está a punto de publicar su informe y le gustaría poder ofrecer un intervalo de confianza a sus lectores. Construye un intervalo de confianza de 92% para la verdadera proporción de accionistas que planean vender al menos una cuarta parte de sus acciones el próximo mes.

32. Una muestra de 12 tiendas de alfombras vende en promedio 62 alfombras por semana con una varianza de 8 alfombras por semana. Construye un intervalo de confianza de 95% para estimar la media verdadera de las ventas por semana de alfombras.

33. Los siguientes datos representan los minutos diarios que escuchan música grabada una muestra de 8 personas: 88.3, 99.2, 34.9, 81.7, 85.4, 0.0, 53.3, 60.0

a) construye un intervalo de confianza de 98% para la media verdadera del tiempo que escuchan música las personas.

34. Siete amas de casa fueron muestreadas aleatoriamente y se averiguó que gastaban un promedio de 29.20 pesos por día en insumos para realizar sus tareas domésticas, con una desviación estándar de 3.20 pesos por día. Construye un intervalo de confianza de 95% para la media de la población.

35. La desviación estándar de todos los productos que se venden en una tienda de abarrotes es de 25 pesos. Encuentra el tamaño necesario de la muestra para estimar la verdadera media, si se desea tener un error de 5 pesos, con un nivel de confianza de 95 por ciento.

36. Para un estudio de mercado, calcula el tamaño necesario de la muestra para estimar la verdadera proporción de los consumidores satisfechos con un producto nuevo, con un error de 0.05 y un nivel de confianza de 90%. Supón que no tienes una idea muy clara respecto de cuál es la proporción de los consumidores satisfechos.

37. Se desea estimar el salario de los docentes de la facultad. Si se sabe que la desviación estándar es de 2 100 pesos, ¿qué tamaño de la muestra se necesita para estimar la media de la población, si se desea tener un error de 500 pesos, con una confianza de 99 por ciento?

38. La proporción de estudiantes fumadores es del 80%. Encuentra el tamaño de la muestra necesario para estimar la proporción real, con un error del 6 por ciento.

39. La gerencia de una empresa textil ha recibido últimamente muchos ataques por los efectos supuestamente nocivos que su proceso de manufactura ejerce sobre la salud. Un sociólogo propuso la teoría de que los empleados que mueren de causa natural muestran una distribución normal a lo largo de su vida. Si los límites superior e inferior de la duración de su vida en la empresa difieren en no más de 12 años, para un nivel de confianza de 98%, ¿de qué tamaño ha de ser una muestra examinada si se quiere calcular la vida promedio de esos empleados, con un error de 30 semanas?

40. Una tienda de comestibles vende bolsas para basura sin marca y ha recibido muchas quejas respecto a la resistencia de ellas. Parece ser que las bolsas sin

marca son más débiles que las de marca que vende el competidor. El dueño quiere determinar el peso máximo promedio que puede meterse en las primeras bolsas sin marca, sin que se rompan. Si la desviación estándar del peso que rompe las bolsas es de 1 kg, determina el número de bolsas que hay que probar para que el dueño tenga una seguridad de 95% de que el peso de rotura promedio de la muestra se halla dentro de 0.5 kg del promedio verdadero.

41. Un curso de lectura rápida garantiza cierto incremento en la velocidad de lectura en un plazo de dos días. El maestro sabe que habrá pocos que no logren ese aumento, por lo cual antes de señalar el incremento garantizado quiere tener una confianza de 95% de que el porcentaje haya sido estimado dentro de 3% del valor verdadero. ¿Cuál es el tamaño más conservador de la muestra necesario en este problema?

42. Una tienda que se especializa en lámparas artesanales quiere obtener una estimación por intervalo del número promedio de clientes que entra diariamente en la tienda. El dueño tiene una seguridad razonable de que la desviación estándar del número diario de clientes es 15. Ayúdalo a resolver el problema determinando el tamaño de la muestra que él debe usar si quiere lograr un intervalo de confianza de 96% para la verdadera media y un error de 4 clientes solamente.

43. En una muestra de 42 cooperativas escolares se comprobó que el precio promedio de un cierto tipo de dulce era 1.12 pesos, con una desviación estándar de 4 centavos. ¿En qué intervalo caerá 99.7% de las veces la verdadera media del precio del dulce?

44. El departamento de asuntos estudiantiles de una facultad desea saber qué proporción de estudiantes tienen promedios por debajo de 8.0. ¿Cuántos historiales académicos deben examinarse a fin de determinar la proporción dentro de 0.05 con un nivel de confianza de 95 por ciento?

45. Un minisúper compró una carga de cajas de jitomate. Una muestra aleatoria de 50 de ellas reveló un peso neto promedio de 23.2 kg con una desviación estándar de 300 g. ¿Cuáles son los límites superior e inferior del intervalo de confianza del peso neto medio?

46. Si se tiene una media muestral de 91, una desviación estándar de la población de 5.1 y un tamaño de muestra de 41, encuentra el nivel de confianza asociado a los siguientes intervalos:

a) (89.4, 92.6).

b) (89, 93).

c) (90.328, 91.672).

47. Por encuestas anteriores se sabe que la desviación estándar del número de horas de televisión vistas por semana por una familia es de 1.1 horas. Se desea determinar el número promedio de horas de televisión vistas a la semana por cada familia mexicana, con un nivel de confianza de 98% de que el número promedio de la muestra cae dentro de 30 minutos del promedio nacional. En forma conservadora, ¿qué tamaño de muestra debe usarse?

48. Juan Beltrán acaba de comprar un *software* especial que calcula las probabilidades de que una acción de la Bolsa Mexicana de Valores aumente de precio, con una precisión de 85%. ¿En cuántas acciones deberá el señor Beltrán probar el programa a fin de tener una certidumbre de 98% de que el porcentaje de acciones que aumentan de valor en la siguiente semana se encuentra dentro de $\pm 3\%$ de la muestra de la población?

49. Al evaluar la eficacia de un programa de rehabilitación en un Cereso, en una encuesta a 52 presos de un total de 900 se descubrió que 35% de ellos ya habían cometido antes algún delito. Construye un intervalo de confianza de 90% para la proporción de los reincidentes entre los presos de esta cárcel.

50. El gerente de un hotel quiere estimar el número promedio de huéspedes diarios. Los siguientes datos muestran el número de usuarios que se registraron en una muestra de 16 días seleccionados al azar: 59, 61, 57, 50, 54, 53, 60, 87, 60, 60, 54, 64, 57, 61, 57, 58.

a) determina una estimación puntual de la media poblacional.

b) construye un intervalo de confianza para estimar la media real del número de huéspedes por día.

51. El SAT muestreó 200 declaraciones de impuestos y descubrió que el ingreso promedio de la muestra por concepto de devolución de impuestos era de 4 253.9 pesos con una desviación estándar de la muestra de 1 071 pesos.

a) usando esta información, estima la devolución media de la población y su desviación estándar en forma puntual.

b) construye un intervalo de confianza para la media de la población de las devoluciones de impuestos al 90 y 99%.

52. Un escuadrón de rescate está efectuando un estudio para analizar su eficiencia. En una muestra de 49 llamadas, el tiempo promedio de respuesta fue de 15.2 minutos. Se sabe que la desviación estándar de dicho tiempo es de 2.5 minutos, dato que se consiguió en un estudio precedente. Construye un intervalo de confianza para el tiempo medio de respuesta con un nivel de confianza de 90 y 99 por ciento.

52. Un ingeniero de una planta purificadora de agua mide todos los días el cloro contenido en 200 muestras. A través de varios años, ha comprobado que la desviación estándar de la población es de 1.4 mg de cloro por litro. Las últimas muestras dieron un promedio de 4.6 mg de cloro por litro. Establece un intervalo de confianza al 68.7%, para estimar la media verdadera.

54. Un ingeniero industrial mide los tiempos de diversas tareas en un proceso de montaje que hacen los obreros. Se tomó una muestra de 7 obreros, cada uno de las cuales realizaba la misma tarea de montaje y obtuvo los siguientes tiempos de montaje en minutos: 1.9, 2.5, 2.9, 1.3, 2.6, 2.0, 3.0. Construye un intervalo de confianza de 98% para el tiempo medio de montaje.

55. Una empresa de ropa está pensando en reintroducir corbatas anchas con diseños llamativos. Entrevistó a 75 ejecutivos jóvenes (su mercado primario) y averiguó que 70 de ellos opinaban que ese tipo de corbatas eran elegantes y querían comprar una. Usando un nivel de confianza de 98%, construye un intervalo de confianza para la proporción de ejecutivos jóvenes a quienes gustan las corbatas elegantes y llamativas.

56. El dueño de un restaurante piensa adquirir nuevo mobiliario. Para decidir con mayores datos sobre la cantidad que puede invertir en mesas y sillas, desea determinar el ingreso promedio por cliente. Muestreó aleatoriamente a 9 clientes, cuya cuenta promedio fue de 183 pesos con una desviación estándar de 36 pesos. Construye un intervalo de confianza de 95% para el monto de la cuenta promedio por cliente.

4 Conceptos de la prueba de hipótesis

Cuando se interpreta un resultado experimental, una pregunta natural surge respecto a si el resultado pudo haber ocurrido por casualidad. La prueba de hipótesis es un procedimiento estadístico para evaluar si el azar es una explicación plausible de un resultado experimental. Las ideas falsas sobre la verificación de las hipótesis son comunes tanto entre los profesionales como entre los estudiantes. Para ayudar a prevenir estas interpretaciones erróneas, en este capítulo se revisa con detalle la lógica de la prueba de hipótesis.

El estadístico R. Fisher explicó el concepto de la prueba de hipótesis con la historia de una dama probando té. Aquí presentamos un ejemplo en relación con la insistencia de Juan Garibay, profesor de estadística, acerca de la forma de preparar las sopas. Garibay insiste en que las sopas deben ser preparadas con caldo de pollo natural y nunca con caldo de pollo concentrado, que comercialmente se presenta en cubitos o en polvo. Consideremos un experimento hipotético para determinar si Garibay puede distinguir entre una sopa preparada con caldo natural y una con concentrado de pollo. Supón que le damos a probar una serie de 16 tazas de sopa. En cada prueba, lanzamos una moneda para ver si le damos una con caldo natural o una con concentrado. Luego le damos la sopa y le pedimos que determine si es con caldo natural o con concentrado. Digamos que Garibay contestó en forma correcta a 13 de las 16 tazas. ¿Prueba este resultado que Garibay tiene al menos algún grado de habilidad para distinguir entre una sopa preparada con caldo de pollo natural y una preparada con caldo de pollo concentrado?

Alguien podría decir que este resultado no prueba que tiene algún grado de habilidad; puede ser que el resultado sólo se deba a la suerte y por eso adivinó correctamente 13 veces de las 16. Pero, ¿qué tan plausible es la explicación de que este resultado sólo se debe a que Garibay es suertudo? Para evaluar si nos debemos conformar con esta explicación, determinemos la probabilidad de que alguien que sólo esté adivinando conteste correctamente 13/16 veces o más. Esta probabilidad se puede calcular utilizando la distribución binomial. Usando la calculadora de la distribución binomial encontramos que esta probabilidad es 0.0106. Ésta es una probabilidad muy baja y, por tanto, alguien tendría que ser muy suertudo para acertar 13 de las 16 veces, si no tuviera ninguna habilidad, es decir, sólo adivinando. Entonces, el problema es decidir si Garibay es muy suertudo, o tiene algún grado de habilidad para distinguir entre una sopa preparada con caldo de pollo natural y una preparada con caldo de pollo concentrado. No se ha comprobado que la hipótesis de que el resultado se debe al azar, es decir, que Garibay sólo adivinaba, sea falsa, pero se tiene duda considerable sobre ella. Por tanto, hay fuertes indicios que nos permiten decir que Garibay puede distinguir entre una sopa preparada con caldo natural y otra con caldo concentrado.

Consideremos otro ejemplo: una compañía de aviación de bajo costo está planeando abrir vuelos de varias ciudades de la República a la ciudad de México. Empezará con un solo vuelo Monterrey-México-Monterrey y desea captar a clientes con perfil empresarial y de negocios. Una encuesta realizada dirigida a este sector le indica que muchos de sus clientes potenciales realizan negocios en las zonas industriales del área conurbada del Estado de México, como Naucalpan, Tlalnepantla, Tepotzotlán, etc. Por problemas de tráfico aéreo, no es posible que le den permiso de operar en el aeropuerto de la ciudad de México, por lo que tiene dos opciones: Querétaro y Cuernavaca. Un aspecto relevante es el que se refiere al tiempo de transporte del aeropuerto elegido a las zonas industriales. Se seleccionaron al azar 30 taxis que dan servicio en cada uno de los aeropuertos, contratándolos para hacer el recorrido hasta un punto de referencia en distintos días durante una semana y en los horarios de llegada y salida del vuelo piloto.

Se determinó que el tiempo promedio desde Cuernavaca fue de 2.75 horas, comparado con el tiempo promedio de 2.5 horas desde Querétaro. ¿A qué se debe esta diferencia? Una posibilidad es que se necesitó más tiempo para trasladarse desde Cuernavaca, debido no a la distancia, sino a las condiciones del tráfico. Por otro lado, quizá por casualidad, los taxistas que vienen de Querétaro son más rápidos. Hay innumerables factores que pueden afectar el tiempo de recorrido, por ejemplo el clima (si llueve se reducirá la velocidad), la presencia de accidentes, tramos en reparación, manifestaciones en la ciudad, etc. Entonces, ¿es posible que

estas diferencias debidas al azar entre los dos grupos sean las responsables de la diferencia en los tiempos de transporte?

Evaluar qué tan plausible es la hipótesis de que las diferencias detectadas se deban al azar nos lleva a calcular la probabilidad de obtener una diferencia igual o mayor a la diferencia detectada ($2.75 - 2.5 = 0.25$ horas), si suponemos que la diferencia sólo se debe al azar.

Utilizando métodos que se presentan en otra sección, se determina que esta probabilidad es igual a 0.451. Ya que este valor es muy alto, tendremos confianza en decir que la diferencia entre los tiempos de recorrido se debe al azar y no al aeropuerto de origen.

El valor de la probabilidad

Es muy importante entender en forma precisa qué significa el valor de la probabilidad. En el ejemplo de Garibay, la probabilidad fue de 0.0106, y es la probabilidad de que él acertará correctamente 13 de las 16 pruebas, si sólo estuviera adivinando.

Es fácil confundir esta probabilidad de 0.0106, como la probabilidad de que él no pueda distinguir la diferencia entre las sopas. Esto no es lo que representa el valor de esta probabilidad.

La probabilidad de 0.016 es la probabilidad de obtener un cierto resultado (13 o más de 16 pruebas), suponiendo cierto estado de la naturaleza (el estado de la naturaleza es: Garibay estaba solamente adivinando). No es la probabilidad que el estado de la naturaleza sea verdadero. Aunque esto podría parecer una diferencia sin importancia, no lo es. Considere el ejemplo siguiente: los domingos en Coyoacán, un pajarero te apuesta a que su canario puede seleccionar los números que son divisibles entre 7. En un experimento para evaluar esta afirmación se deja al ave realizar 16 ensayos. En cada prueba, un número se despliega sobre una pantalla y el ave picotea en una de dos teclas para indicar su elección. Los números son escogidos de tal manera que la probabilidad de que aparezca un número divisible entre 7 es 0.5. El ave acierta 9 de las 16 veces. Usando la distribución binomial, calculamos la probabilidad de acertar 9/16 o más, considerando que el ave sólo está adivinando y esta probabilidad resulta ser 0.40. Es decir, el ave, si sólo estuviera adivinando, acertaría 40% de las veces, por lo que con base en esto no hay pruebas convincentes para suponer que el ave pueda distinguir los números.

Como científico, serías muy escéptico en creer que el ave tuviera esa habilidad. ¿Llegarías a la conclusión de que hay 0.40 de probabilidad de que el ave puede detectar la diferencia? ¡Por supuesto que no! Pensarías que la probabilidad es mucho menor a 0.0001.

Insistimos en que el valor de la probabilidad es la probabilidad de que ocurra un resultado (9/16 o más) y no la probabilidad de que suceda un determinado estado de la naturaleza (el ave puede distinguir los números divisibles entre 7). En estadística, convencionalmente llamamos a los posibles estados de la naturaleza "hipótesis", ya que son estados hipotéticos de la naturaleza. Usando esta terminología, el valor de la probabilidad es la probabilidad de obtener un resultado dada una hipótesis. No es la probabilidad de la hipótesis dado un resultado.

Esto no quiere decir que ignoremos la probabilidad de las hipótesis. Si la probabilidad de obtener un resultado dada una hipótesis es suficientemente baja, tenemos pruebas de que la hipótesis es falsa. Sin embargo, nosotros en ningún momento estamos calculando la probabilidad de que la hipótesis sea falsa. En el ejemplo de Garibay, la hipótesis es que él no puede distinguir entre las sopas. El valor de la probabilidad fue bajo (0.0106), lo que proporcionó indicios de que él sí puede distinguir entre las sopas. Sin embargo, no calculamos la probabilidad de que

pueda distinguir las sopas. Una rama de la estadística conocida como estadística Bayesiana proporciona métodos para calcular las probabilidades de las hipótesis. Estos cálculos requieren que se especifique la probabilidad de las hipótesis antes de considerar los datos, lo que supone dificultades en algunos contextos.

La hipótesis nula

La hipótesis de que un efecto detectado se debe únicamente al azar se llama hipótesis nula. En el ejemplo de la selección de un aeropuerto, la hipótesis nula es que en la población el tiempo promedio de recorrido desde Querétaro es igual al tiempo de recorrido desde Cuernavaca. La hipótesis nula se expresa de la siguiente manera:

$$\mu_{\text{Cuernavaca}} = \mu_{\text{Querétaro}}$$

$$\mu_{\text{Cuernavaca}} - \mu_{\text{Querétaro}} = 0$$

Ahora bien, la hipótesis nula en un estudio de correlación sería considerar que la correlación en la población es cero, es decir:

$$\rho = 0$$

donde ρ es el coeficiente de correlación en la población (no lo confundas con r, el coeficiente de correlación en la muestra). Aunque en la hipótesis nula el valor del parámetro por lo general es 0, hay ocasiones en las cuales tiene un valor distinto de 0. Por ejemplo, si se estuviera evaluando la habilidad de un individuo para decir correctamente el resultado del lanzamiento de una moneda, la hipótesis nula sería $\pi = 0.5$.

Ten en cuenta que la hipótesis nula por lo general es lo opuesto a la hipótesis de investigación. En el estudio de los aeropuertos, la hipótesis de los investigadores es que se invierte menos tiempo desde Querétaro a pesar de que es mayor la distancia. La hipótesis nula de que el tiempo de recorrido es igual se plantea con la esperanza de rechazarla. Si la hipótesis nula fuera verdadera, una diferencia igual o mayor que la diferencia encontrada en la muestra, que fue igual a 0.25 horas, es muy probable que ocurra. El valor de p, cuyo cálculo se verá en secciones posteriores, fue de 0.451. Por tanto, los investigadores aceptarían la hipótesis nula que expresa que no existe diferencia entre los tiempos de recorrido.

Si se hubiera rechazado la hipótesis nula, entonces se debería aceptar la alternativa a ella, llamada hipótesis alternativa. La hipótesis alternativa es simplemente lo contrario de la hipótesis nula. Si la hipótesis nula:

$$\mu_{\text{Cuernavaca}} = \mu_{\text{Querétaro}}$$

se hubiera rechazado, entonces hay dos alternativas:

$$\mu_{\text{Cuernavaca}} < \mu_{\text{Querétaro}}$$

$$\mu_{\text{Cuernavaca}} > \mu_{\text{Querétaro}}$$

Naturalmente, la dirección de la diferencia entre las medias muestrales determina qué alternativa utilizar. Algunos textos han argumentado incorrectamente que rechazar la hipótesis nula que indica que las medias de las poblaciones son iguales no permite concluir cuál media de la población es mayor.

Un valor pequeño de probabilidad pone en duda la hipótesis nula. ¿Qué tan bajo debe ser el valor de probabilidad para llegar a la conclusión de que la hipótesis nula es falsa? Aunque evidentemente no hay respuesta correcta o equivocada para

esta pregunta, es convencional concluir que la hipótesis nula es falsa, si el valor de la probabilidad es menor a 0.05. Los investigadores más conservadores concluyen que la hipótesis nula es falsa, solamente si el valor de la probabilidad es menor de 0.01. Cuando un investigador llega a la conclusión de que la hipótesis nula es falsa, el investigador dice que rechaza la hipótesis nula. El valor de probabilidad bajo el cual se rechaza la hipótesis nula se conoce como nivel de significancia.

PREGUNTAS

1. Tomás asegura que puede adivinar a ciegas las respuestas correctas de una prueba que contiene 20 preguntas de la modalidad falso o verdadero. Realiza la prueba y acierta en un 80%. Usando la calculadora binomial se encontró que la probabilidad de obtener 16 o más respuestas correctas, cuando $p = 0.5$ es 0.0059. La probabilidad de 0.0059:

a) es la probabilidad de que Tomás obtendrá un 80% de respuestas correctas si vuelve a realizar el examen.

b) es la probabilidad de que obtenga este puntaje o un puntaje mayor, si se considera que solamente está adivinando.

c) es la probabilidad de que él esté adivinando a ciegas en la prueba.

2. Un investigador cree que los alumnos del segundo curso de estadística tienen calificaciones mayores que los del primer curso en una prueba particular. ¿Cuál de las siguientes opciones es la hipótesis nula?

a) media del primer curso < media del segundo curso.

b) media del primer curso > media del segundo curso.

c) media del primer curso = media del segundo curso.

3. Los investigadores tienen la hipótesis de que hay correlación entre la estatura y el peso de los estudiantes de nuevo ingreso a la Universidad. La hipótesis nula es que la correlación de la población es igual a:

a) Respuesta: _____.

Prueba de significancia

Cuando la hipótesis nula se rechaza, se dice que el efecto es estadísticamente significante. Por ejemplo, en un estudio dirigido para comparar la preferencia de las amas de casa por dos marcas de detergentes en polvo, el valor de la probabilidad fue de 0.0057. Por tanto, el efecto del detergente en el grado de aceptación es estadísticamente significativo y la hipótesis nula de que para las amas de casa es indistinto usar cualquiera de los dos se rechaza. Es muy importante tener en cuenta que la significancia estadística quiere decir solamente que se rechaza la hipótesis nula, que indica que exactamente no existe ningún efecto; no quiere decir que el efecto sea importante. Entonces, en general ¿qué se entiende por significante? Cuando un efecto es significante, puedes tener confianza en que el efecto no es exactamente cero. Detectar que un efecto es significativo no indica qué tan grande o importante es.

No confundas la significancia estadística con la significancia práctica. Un efecto pequeño puede ser muy significante si el tamaño de la muestra es suficientemente grande.

¿Por qué la palabra "significante" en la frase "estadísticamente significante", representa algo tan diferente a los otros usos de la palabra? Curiosamente, porque

el significado del "significante" en el lenguaje cotidiano ha cambiado. En el siglo XIX, cuando se desarrollaron los procedimientos para la prueba de hipótesis, algo "significante" sí significaba algo. Por tanto, encontrar que un efecto es estadísticamente significativo quiere decir que el efecto es real y no se debe al azar. Con el paso de los años, el significado de "significante" cambió, conduciendo a interpretaciones potencialmente equivocadas.

Hay dos enfoques (por lo menos) para realizar las pruebas de significancia. En uno (preferido por R. Fisher) se realiza la prueba de significancia y el valor de probabilidad refleja la fuerza de los resultados obtenidos en contra de la hipótesis nula. Si la probabilidad es menor que 0.01, los datos revelan fuertes indicios de que la hipótesis nula es falsa. Si el valor de la probabilidad es menor que 0.05, pero mayor que 0.01, entonces la hipótesis nula, por lo general, se rechaza, pero no con la misma confianza que para el caso de que el valor de la probabilidad fuera menor a 0.01. Valores de probabilidad entre 0.05 y 0.10 proporcionan pruebas débiles en contra de la hipótesis nula, y no se consideran suficientemente pequeñas para justificar el rechazo de ésta. Probabilidades más altas proporcionan menores pruebas de que la hipótesis nula sea falsa.

El enfoque alternativo (preferido por los estadísticos Neyman y Pearson) es especificar un nivel a, antes de analizar los datos. Si al analizar los datos resulta un valor de probabilidad menor al nivel a, entonces la hipótesis nula se rechaza; si no es así, entonces la hipótesis nula no se rechaza. De acuerdo con esta perspectiva, si un resultado es significante, no importa qué tan significante es. Además, si no es significativo, no importa qué tan cerca está de serlo. Por tanto, si se usa un nivel de 0.05, entonces da igual obtener valores de probabilidad, por ejemplo, 0.049 o 0.001. De forma semejante, da lo mismo obtener valores de probabilidad, por ejemplo de 0.06 o 0.34.

El primer enfoque (preferido por Fisher) es más apropiado para la investigación científica y será el que usemos en este texto. El segundo es más apropiado para las aplicaciones en las que la decisión es del tipo pasa o no pasa. Por ejemplo, si un análisis estadístico fuera realizado para determinar si una máquina en una fábrica está funcionando mal, el análisis estadístico se usaría para determinar si la máquina debe o no ser reparada. El gerente de la planta estaría menos interesado en evaluar la fuerza de las pruebas que en decidir qué acción tomar. No hay necesidad de tomar una decisión inmediata en el campo de la investigación científica, donde un investigador puede llegar a la conclusión de que hay pocos indicios en contra de la hipótesis nula, pero que es necesario realizar más trabajo de investigación, antes de concluir en forma definitiva.

PREGUNTAS

1. En la investigación es convencional rechazar la hipótesis nula si el valor de la probabilidad es menor a:

a) Respuesta _____.

2. Selecciona lo que aplique. El valor de probabilidad bajo el cual la hipótesis nula se rechaza, también se llama:

a) probabilidad clave.
b) nivel de significancia.
c) nivel α.
d) valor objetivo.

3. Cuando comparamos las calificaciones de dos grupos, una diferencia de un punto nunca es altamente significativa, aun si tienes una muestra muy grande.

a) verdadero.

b) falso.

4. Hay dos maneras de aproximar la prueba de significancia. En una aproximación, el valor de probabilidad refleja la fuerza de la prueba en contra de la hipótesis nula. Cuanto menor sea el valor de la probabilidad, hay más pruebas de que la hipótesis nula es falsa. ¿Qué estadístico apoya esta aproximación?

a) Fisher.

b) Neyman.

c) Pearson.

Errores Tipo I y Tipo II

En el caso de estudio de las "Reacciones de los médicos ante el peso de los pacientes", el valor de probabilidad asociado con la prueba de significancia es 0.0057, por lo que la hipótesis nula se rechaza y se concluye que los médicos pretenden pasar menos tiempo con los pacientes obesos. A pesar de este pequeño valor de probabilidad, es posible que la hipótesis nula que indica ninguna diferencia en el tiempo de consulta dedicado a personas obesas y a personas con peso normal sea verdadera y que la gran diferencia encontrada entre las medias de las muestras se deba al azar. Si éste es el caso, entonces la conclusión de que los médicos pasan menos tiempo con los pacientes obesos es errónea. Este tipo de error se conoce como error Tipo I. En forma general, un error Tipo I sucede cuando una prueba de significancia da como resultado el rechazo de una hipótesis nula verdadera.

Por una convención común, si el valor de probabilidad es menor de 0.05 se rechaza la hipótesis nula. Otra convención, aunque ligeramente menos común, es rechazar la hipótesis nula si el valor de probabilidad es menor de 0.01. El umbral para el rechazo de la hipótesis nula se llama nivel de significancia α. Como se discutió en la introducción de la prueba de hipótesis, lo más adecuado es interpretar el valor de la probabilidad como un indicador del peso o de la fuerza de la prueba en contra de la hipótesis nula y no como parte de una regla de decisión elaborada para el rechazo o el no rechazo. Por tanto, ten en cuenta que el rechazo de la hipótesis nula no es una decisión de todo o nada.

El error Tipo I depende del nivel α: cuanto menor sea el nivel, menor será el valor del error Tipo I. Podría decirse que el nivel α es la probabilidad de que suceda el error Tipo I. En forma correcta, el nivel α es la probabilidad de que suceda el error Tipo I, siempre y cuando la hipótesis nula sea verdadera. Si la hipótesis nula es falsa, entonces es imposible que suceda el error Tipo I.

El segundo tipo de error que puede suceder en una prueba de significancia es no rechazar (o sea, aceptar) una hipótesis nula falsa. Esta clase de error se conoce como error Tipo II.

A diferencia del error Tipo I, el Tipo II no es realmente un error. Cuando una prueba estadística no es significante, quiere decir que los datos no proporcionan un fuerte indicio para considerar falsa la hipótesis nula. La falta de significancia no respalda la conclusión de que la hipótesis nula sea verdadera. Por tanto, un investigador no debe cometer el error de concluir que la hipótesis nula es verdadera cuando la prueba estadística no es significante. El investigador debe considerar la prueba como no concluyente. Esto contrasta con el error Tipo I, en el cual el investigador erróneamente concluye que la hipótesis nula es falsa cuando en realidad es verdadera.

Un error Tipo II sólo puede ocurrir cuando la hipótesis nula es falsa. Si la hipótesis nula es falsa, entonces la probabilidad de cometer un error Tipo II es β. La probabilidad de rechazar correctamente una hipótesis nula que sea falsa es igual a $1 - \beta$ y se denomina potencia. La potencia se revisa en detalle en otra sección.

PREGUNTAS

1. Se ha visto repetidamente en una cierta prueba de memoria que reconocer es sustancialmente mejor que recordar. Sin embargo, el valor de probabilidad para los datos de una muestra fue de 0.12, por lo que no puedes rechazar la hipótesis nula que asegura que reconocer y recordar produce los mismos resultados. ¿Qué tipo de error estás cometiendo?

a) error Tipo I.

b) error Tipo II.

2. En la población no hay diferencia entre hombres y mujeres en una cierta prueba. Sin embargo, encontraste diferencia en tu muestra. El valor de probabilidad calculado para los datos fue de 0.03, por lo que rechazas la hipótesis nula. ¿Qué tipo de error estás cometiendo?

a) error Tipo I.

b) error Tipo II.

3. A medida que el nivel α disminuye, ¿cuál es el error que también disminuye?

a) error Tipo I.

b) error Tipo II.

4. β es la probabilidad del:

a) error Tipo I.

b) error Tipo II.

5. Si la hipótesis nula es falsa, ¿qué tipo de error podrías cometer?

a) error Tipo I.

b) error Tipo II.

Pruebas de una o dos colas

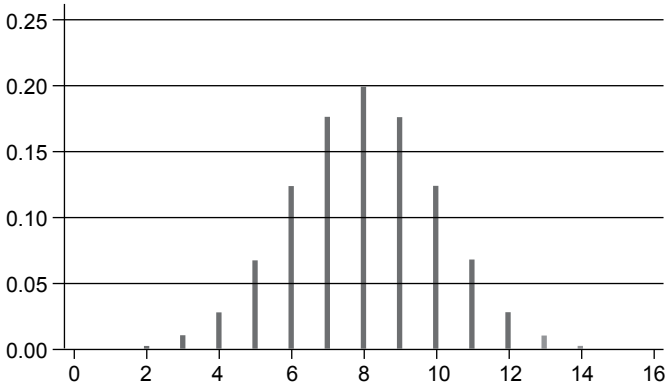


Figura 4.1 La distribución binomial. La cola superior se muestra en las barras de tono más claro.

En el ejemplo de Garibay, se le dieron 16 tazas de sopa para que distinguiera si se habían elaborado con caldo de pollo natural o concentrado. Él contestó acertadamente 13 veces. Utilizando la distribución binomial, sabemos que la probabilidad de acertar correctamente 13 veces en 16 ensayos, si no tiene ninguna habilidad y sólo está adivinando, es 0.0106. En la figura 4.1 se muestra la gráfica de la distribución binomial. Las barras de tono más claro muestran los valores iguales o mayores a 13. Como puedes ver, el valor calculado de la probabilidad corresponde a la cola derecha de la distribución. La probabilidad calculada en una cola de la distribución se conoce como "probabilidad de una cola".

Se puede hacer una pregunta ligeramente diferente: "¿Cuál es la probabilidad de obtener un resultado tan extremo o más que el que observamos, que es igual a 13/16?" Ya que el valor esperado es 8/16, un resultado de 3/13 es tan extremo como 13/16, de tal manera que el cálculo de la probabilidad debe considerar las dos colas de la distribución. Ya que la distribución binomial es simétrica cuando $\pi = 0.5$, entonces la probabilidad pedida es exactamente el doble de la probabilidad calculada anteriormente de 0.0106. Por tanto, $p = 0.0212$. La probabilidad calculada en ambas colas de la distribución se llama probabilidad de dos colas.

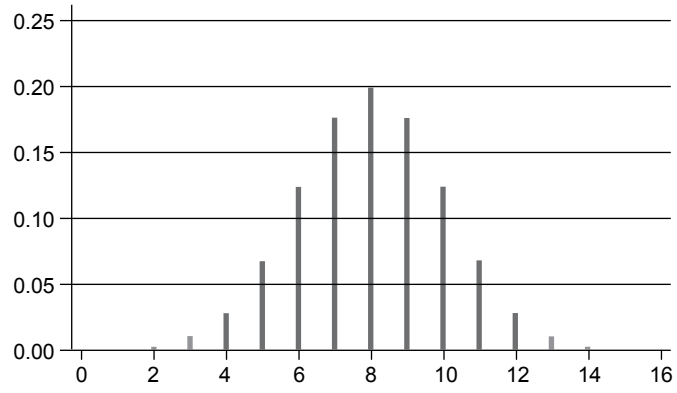


Figura 4.2 La distribución binomial, ambas colas en tono más claro.

¿Cuál probabilidad debes utilizar para evaluar la habilidad del señor Garibay: la de una cola o la de dos colas? Esto depende de la manera en que se plantea la pregunta: si la pregunta es si Garibay puede distinguir entre una sopa preparada con caldo de pollo natural y otra con caldo de pollo concentrado, entonces concluiríamos que él sí tiene esa habilidad si su actuación es mucho mejor o mucho peor que la debida al azar. Observa que la pregunta es "si puede distinguir", no se pregunta si puede "acertar". Es decir, puede distinguir cuando su actuación es mejor que el azar; en este caso, acertaría en distinguir las sopas. Pero también las distingue si su actuación es mucho peor que la debida al azar, aunque en este caso no sabe cuál es cuál. Por tanto, ya que se va a rechazar la hipótesis nula que sostiene que Garibay tiene un desempeño muy bueno o muy pobre en los 16 ensayos, entonces se utilizará una probabilidad de dos colas. Por otro lado, si la pregunta es si Garibay acierta más veces que las que lo haría cualquier persona sin ninguna habilidad (sólo adivinando), es decir, que su actuación es mejor que la debida al azar, usaríamos una probabilidad de una cola. ¿Cuál sería la conclusión al usar la probabilidad de una sola cola si Garibay contestó correctamente a sólo 3 de las 16 pruebas? Ya que la probabilidad de una cola, en nuestro caso, es la probabilidad de la cola derecha, calcularíamos la probabilidad de tener tres o más aciertos. Esta probabilidad es muy alta; por tanto, la hipótesis nula no debe ser rechazada.

La hipótesis nula para una prueba de dos colas es $\pi = 0.5$. En contraste, la hipótesis nula para una prueba de una cola es de $\pi \leq 0.5$.

De acuerdo con esto, rechazamos la hipótesis de dos colas si la proporción de aciertos en la muestra se aparta bastante de 0.5 en cualquier dirección. Para la prueba de una sola cola la hipótesis nula se rechaza únicamente si la proporción en la muestra es mucho mayor que 0.5.

La hipótesis alternativa en una prueba de dos colas es $\pi \neq 0.5$. En la prueba de una cola es $\pi \geq 0.5$.

Debes decidir qué tipo de prueba usar antes de ver tus datos. Las pruebas estadísticas que calculan las probabilidades de una cola se llaman pruebas de una sola cola; y las que calculan probabilidades de dos colas se llaman pruebas de dos colas. En la investigación científica, las pruebas de dos colas son más comunes que las de una sola cola debido a que un resultado extremo significa que está operando algo que no es el azar, y es importante tomarlo en cuenta. Las pruebas de una sola cola son apropiadas cuando no es importante distinguir entre un efecto y el efecto nulo en una dirección no esperada. Por ejemplo, considera un experimento diseñado para probar la eficacia de un tratamiento para el resfriado común: el investigador

sólo está interesado en saber si el tratamiento es mejor que un placebo o control y no es importante distinguir el caso en que el tratamiento fuera peor que el placebo, ni tampoco el caso en que fueran iguales, debido en que en ambos casos la droga sería inútil.

Algunos autores argumentan que las pruebas de una sola cola se justifican siempre y cuando el investigador haya predicho el efecto en esta dirección. El problema o la crítica a este argumento es que si el resultado del efecto es muy fuerte en la dirección no predicha, el investigador no podrá justificar que el efecto no es cero. Ya que esto no es real, las pruebas de una sola cola por lo general deben verse con escepticismo si se justifican solamente sobre esta base.

PREGUNTAS

1. Selecciona lo que aplique para la prueba de dos colas.

a) son apropiadas cuando no es importante distinguir entre el efecto y el no efecto en una dirección no esperada.

b) son más comunes que las pruebas de una cola.

c) se calcula la probabilidad en las dos colas.

d) son más controversiales que las pruebas de una cola.

2. Estás probando la diferencia entre los alumnos de primero y noveno semestres en cierta prueba. Piensas que los del noveno semestre tendrán mejor desempeño, pero estás interesado en saber si tal vez los del primer semestre tendrán un mejor desempeño. ¿Cuál es la hipótesis nula?

a) la media de noveno semestre $\leq$ media de los de primer semestre.

b) la media de noveno semestre $\geq$ media de los de primer semestre.

c) la media de noveno semestre $=$ media de los de primer semestre.

3. Piensas que una moneda no es legal y que con ella caerán más veces águilas que soles. ¿Cuál es la probabilidad de que en 22 lanzamientos obtengas 16 o más águilas? Escribe tu respuesta con al menos tres decimales.

a) Respuesta _____.

4. Piensas que una moneda no es legal y estás interesado en verificarlo. ¿Cuál es la probabilidad de obtener 8 o menos águilas en 30 lanzamientos? Escribe tu respuesta con al menos 3 decimales.

a) Respuesta _____.

Interpretación de los resultados significantes

Cuando el valor de probabilidad es menor que el nivel α, el efecto es estadísticamente significante y la hipótesis nula se rechaza. Sin embargo, no todos los efectos estadísticamente significantes deben ser tratados de la misma manera. Por ejemplo, tendrás menos confianza en decidir el rechazo de la hipótesis nula, si $p = 0.049$ en lugar de $p = 0.003$. Entonces, rechazar la hipótesis nula no es una proposición de todo o nada.

Si la hipótesis nula se rechaza, entonces la alternativa, llamada hipótesis alternativa, se acepta. Considera la prueba de una sola cola en el ejemplo de Garibay: a él se le proporcionaron 16 tazas de sopa y se le pidió que para cada taza indicara si la sopa había sido preparada con caldo natural o concentrado. La interrogante

es determinar si puede distinguir las sopas correctamente, es decir, si tiene más aciertos que los atribuidos al azar. La hipótesis nula para la prueba de una sola cola es $\pi \leq 0.5$, donde π es la probabilidad de acertar correctamente en cualquier ensayo. Si la hipótesis nula se rechaza, entonces se acepta la hipótesis alternativa, de $\pi > 0.5$. Si $\pi > 0.5$, entonces Garibay es mejor que el azar, es decir, sí distingue correctamente las sopas.

Ahora consideremos la prueba de dos colas utilizada en el ejemplo del estudio de preferencia de los detergentes. La hipótesis nula es:

$$\mu_A = \mu_B$$

Si la hipótesis nula se rechaza, entonces hay dos alternativas:

$$\mu_A > \mu_B$$

$$\mu_A < \mu_B$$

Naturalmente, la dirección de las medias de las muestras determina cuál alternativa usar. Si la media de los puntajes en la muestra para el detergente "A" es significativamente mayor que para el detergente "B", entonces debemos concluir que la media de la población de los puntajes para el detergente "A" es mayor que la media de la población de los puntajes para el detergente "B".

Hay muchas situaciones en las cuales es muy improbable que dos condiciones tengan exactamente las mismas medias en la población. Por ejemplo, es imposible que la aspirina y el paracetamol proporcionen el mismo grado de alivio. Por tanto, aun antes de realizar un experimento para comparar su efectividad, los investigadores saben que la hipótesis nula de no diferencia es falsa. Sin embargo, ellos no saben cuál droga proporciona mayor alivio. Si la prueba de la diferencia es significante, entonces se conoce la dirección de ésta. Este punto también se trata en la sección de intervalos de confianza y prueba de significancia.

Algunos textos incorrectamente dicen que el rechazo de la hipótesis nula, de que dos medias de poblaciones son iguales, no justifica la conclusión de que una media de la población es más grande, en vez de esto dicen que uno puede concluir que las medias de la población son diferentes. La validez de concluir la dirección del efecto es claro, si notas que una prueba de dos colas al nivel de 0.05 es equivalente a dos pruebas separadas de una cola, cada una al nivel de 0.025. Las dos hipótesis nulas son, entonces:

$$\mu_A \geq \mu_B$$

$$\mu_A \leq \mu_B$$

Si la primera de estas hipótesis se acepta, entonces la conclusión es que la media de la población de los puntajes para el detergente "A" es mayor que la media de la población de los puntajes para el detergente "B", es decir, se muestra mayor preferencia por el detergente "A". Si la última se acepta, entonces la conclusión es que se muestra mayor preferencia por el detergente "B", ya la media de los puntajes en la población para este detergente es mayor que para el "A".

Interpretación de los resultados no significativos

Cuando en una prueba de significancia obtenemos un valor alto de probabilidad, quiere decir que los datos no proporcionan evidencia que permita rechazar la hipótesis nula. Sin embargo, un valor alto de probabilidad no quiere decir que la hipótesis nula sea verdadera. El problema consiste en la imposibilidad de distinguir entre un efecto nulo y uno muy pequeño. Por ejemplo, en el caso de Garibay, vamos a suponer que es, en efecto, apenas un poco mejor que el azar al distinguir entre las

sopas preparadas con caldo natural o concentrado. Supongamos que tiene una probabilidad de 0.51 de acertar en una prueba dada ($\pi = 0.51$). Digamos que el experimentador Martínez (quien no sabe que π es igual a 0.51), le da a probar diferentes tazas de sopa y encuentra que acierta 49 de 100 veces. ¿Cuál sería el resultado de la prueba de significancia? El experimentador realizaría la prueba de significancia con base en la suposición de que Garibay tiene una probabilidad de acertar en cada ensayo de 0.50 ($\pi = 0.50$). Dada esta suposición, la probabilidad de acertar 49 o más veces en 100 ensayos es 0.62. Es decir, el valor de probabilidad es 0.62, que es un valor muy grande en comparación al nivel de significancia de 0.05. Este resultado no proporciona ninguna base que nos permita rechazar la hipótesis nula. Sin embargo, nosotros sabemos (pero el experimentador Martínez no lo sabe) que π es igual a 0.51 y no a 0.50 y por tanto la hipótesis nula es falsa. De tal manera que si el experimentador Martínez hubiera concluido aceptar la hipótesis nula basado en el análisis estadístico, habría cometido un error. Hacer esto es un error serio.

Entonces, ¿cómo deben interpretarse los resultados no significativos? El experimentador debe informar que no hay pruebas suficientes para decir que Garibay puede distinguir correctamente entre las dos sopas, pero tampoco puede decir que él no las puede distinguir. Es por lo general imposible probar lo negativo. ¿Qué pasa si dices que fuiste Pancho Villa en una vida anterior? Debido a que no tienes pruebas, tendrías muchas dificultades para convencer a cualquiera de que esto es verdad. Sin embargo, nadie podría ser capaz de probar definitivamente que no fuiste Pancho Villa.

A menudo, un resultado no significante incrementa la confianza de que la hipótesis nula es falsa. Considera el ejemplo hipotético siguiente: un investigador desarrolla un tratamiento para la ansiedad que cree que es mejor que el tratamiento tradicional. Conduce un estudio para evaluar la eficacia relativa de los dos tratamientos: 20 sujetos seleccionados aleatoriamente se dividen en dos grupos de 10 personas cada uno. Un grupo recibe el nuevo tratamiento y el otro el tradicional. El nivel promedio de ansiedad es inferior para el grupo que recibe el nuevo tratamiento, en comparación con el promedio calculado para el que recibe el tratamiento tradicional. Sin embargo, la diferencia no es significante. El análisis estadístico muestra que el valor de la probabilidad de que resulte una diferencia igual o mayor que la obtenida en el experimento sería de 0.11, si no hubiera una diferencia real entre los tratamientos. En otras palabras, debido a que el valor de probabilidad es 0.11, un investigador ingenuo interpretaría este resultado como prueba de que el nuevo tratamiento no es más eficaz que el tratamiento tradicional. Sin embargo, un investigador más refinado, aunque decepcionado porque el efecto resultó no significativo, estaría animado a decir que el nuevo tratamiento dirige la ansiedad a niveles menores que el tratamiento tradicional. Los datos soportan la tesis de que el nuevo tratamiento es mejor que el tradicional, aunque el efecto no es estadísticamente significativo. El investigador debe tener mayor confianza en que el nuevo tratamiento es mejor, en comparación con la confianza que tenía antes de que se realizara el experimento. Sin embargo, el soporte es débil y los datos no son concluyentes.

¿Qué debe hacer el investigador? Una acción razonable sería repetir el experimento o realizar experimentación adicional. Digamos que el investigador repitió el experimento y encontró otra vez que el nuevo tratamiento era mejor que el tratamiento tradicional. Sin embargo, de nuevo el efecto no fue significativo y esta vez el valor de probabilidad fue de 0.07. El investigador novato pensaría que debido a que los dos experimentos no encontraron significancia, el nuevo tratamiento es improbable que sea mejor que el tratamiento tradicional. El investigador experimentado notaría que en los dos experimentos, el nuevo tratamiento es mejor que

el tratamiento tradicional. Además, los dos experimentos proporcionan un soporte débil para decir que el nuevo tratamiento es mejor, pero cuando se examinan en forma conjunta pueden proporcionar un soporte fuerte. Usando un método para combinar las probabilidades, se puede determinar que al combinar los valores de probabilidad, 0.11 y 0.07, se obtiene como resultado un valor de probabilidad de 0.045. Por tanto, los dos resultados no significativos examinados en forma conjunta dan un resultado significativo.

Aunque nunca hay una base estadística para llegar a la conclusión de que un efecto es exactamente igual a cero, un análisis estadístico puede demostrar que un efecto dado es probablemente muy pequeño. Esto se hace construyendo un intervalo de confianza. Si los tamaños de todos los efectos en el intervalo son pequeños, entonces se puede llegar a la conclusión de que el efecto es pequeño. Por ejemplo, considera un experimento para probar la eficacia de un tratamiento para el insomnio. Supón que el tiempo medio para quedarse dormido fue de 2 minutos menos para las personas que recibieron el tratamiento en comparación con las personas del grupo control, y que esta diferencia no es significativa. Si el intervalo de confianza al 95% se extendiera de –04 a 8 minutos, entonces se justificaría que el investigador llegara a la conclusión de que el beneficio es de 8 minutos o menos. Sin embargo, no se justificaría si concluyera que la hipótesis nula es verdadera, o incluso decir que la hipótesis nula está soportada por los datos.

PREGUNTAS

1. Analizaste los resultados de un experimento y calculaste que $p = 0.13$. ¿Qué conclusión puedes sacar? Selecciona todo lo que aplica.

a) rechazar la hipótesis nula.

b) aceptar la hipótesis nula.

c) no rechazar la hipótesis nula.

d) aceptar la hipótesis alternativa.

2. Se aplica un examen a dos grupos de alumnos: uno de Estadística descriptiva y otro de Inferencia estadística. Se encuentra que los de Inferencia estadística tuvieron un mejor desempeño, pero el valor de p calculado fue de 0.08, el cual no es significativo a un nivel de 0.05. ¿Qué debes pensar acerca de la diferencia entre los dos grupos?

a) tienes mayor confianza en que hay diferencia.

b) tienes menor confianza en que hay diferencia.

c) ahora sabes que la diferencia es realmente 0.

Pasos de la prueba de hipótesis

1. El primer paso consiste en plantear la hipótesis nula. Para una prueba de dos colas, la hipótesis nula indica por lo general que el parámetro es cero aunque hay excepciones. Una hipótesis nula típica es $\mu_1 - \mu_2 = 0$, lo que es equivalente a $\mu_1 = \mu_2$. Para una prueba de una cola, la hipótesis nula es que el parámetro es mayor o igual a cero, o bien que el parámetro es menor o igual a cero. Si la predicción es que μ_1 es mayor que μ_2, entonces la hipótesis nula (lo contrario a la predicción) es $\mu_2 - \mu_1 \geq 0$. Esto es equivalente a $\mu_1 \leq \mu_2$.

2. El segundo paso es especificar el nivel α, llamado nivel de significancia. Normalmente estos valores son 0.05 y 0.01.

3. El tercer paso es calcular el valor de la probabilidad (también conocido como el valor *p*). Ésta es la probabilidad de obtener un estadístico muestral tan diferente o más que el parámetro especificado en la hipótesis nula, dado que ésta es verdadera.
4. El cuarto paso es comparar el valor de probabilidad con el nivel α. Si el valor de probabilidad es menor, entonces se rechaza la hipótesis nula. Ten en cuenta que el rechazo de la hipótesis nula no es una decisión de todo o nada. Cuanto más pequeño sea el valor de la probabilidad, tienes más confianza en decir que la hipótesis nula es falsa. Si el valor de la probabilidad es mayor que el nivel α de 0.05, muchos científicos consideran que sus resultados no son concluyentes. No rechazar la hipótesis nula no constituye un soporte para aceptarla, sólo significa que los datos no proporcionan fuertes señales que permitan rechazarla.

PREGUNTAS

1. Primero planteas la hipótesis nula, analizas los datos y calculas *p*; observas este valor de *p* y, dependiendo de este valor, seleccionas un nivel α apropiado. Finalmente decides si aceptas o rechazas la hipótesis nula.
   a) verdadero.
   b) falso.
2. El objetivo de las investigaciones es comprobar que la hipótesis nula es verdadera.
   a) verdadero.
   b) falso.

Pruebas de significancia e intervalos de confianza

Existe una relación cercana entre los intervalos de confianza y las pruebas de significancia. Específicamente, si un estadístico es significativamente diferente de cero a un nivel de 0.05, entonces el intervalo de confianza al 95% no contiene al 0. Todos los valores incluidos en el intervalo de confianza son valores posibles del parámetro y se considera que los valores no incluidos no son valores posibles de éste. En el ejemplo del tiempo de recorrido de los aeropuertos a un punto de referencia, el intervalo de confianza al 95% para la diferencia entre las medias se extiende desde −0.41 a 0.91. Por tanto, cualquier valor menor a −0.41 horas o mayor a 0.91 horas, se rechaza como un valor posible del parámetro "diferencia entre las medias de las poblaciones". Ya que 0 está incluido en el intervalo, se acepta como un valor posible y por tanto la prueba de la hipótesis nula de no diferencia entre las medias no es significante. El valor de la probabilidad para este ejemplo se había calculado en 0.451, lo que indica significancia. En forma similar, hay una relación entre el intervalo de confianza y la prueba de significancia al nivel de 0.01.

Siempre que un efecto sea significativo, todos los valores del intervalo serán positivos o serán negativos (no incluirán al cero). Por tanto, un resultado significante permite que el investigador determine la dirección del efecto. Hay muchas situaciones en las cuales es muy difícil que se tengan exactamente las mismas medias poblacionales. Por ejemplo, es prácticamente imposible que la aspirina y el paracetamol proporcionen exactamente el mismo grado de alivio del dolor. Por tanto, aun antes de que se realice el experimento, ya se conoce que la hipótesis nula de que no exista exactamente ninguna diferencia es falsa. Sin embargo, el investigador no sabe cuál de las dos drogas proporciona mayor alivio. Si la prueba de la diferencia

es significativa, entonces se puede establecer la dirección de la diferencia porque los valores en el intervalo de confianza son o todos positivos o todos negativos.

Si, por ejemplo, el intervalo de confianza al 95% no contiene al cero (en forma más precisa, el valor especificado en la hipótesis nula), entonces el efecto es significativo al nivel de 0.05.

Por otro lado, observa que un efecto no significativo en términos del intervalo de confianza aclara por qué la hipótesis nula no debe ser aceptada, aun cuando decidimos no rechazarla. Cada valor incluido en el intervalo de confianza es un valor posible del parámetro. Debido a que cero está incluido en el intervalo, la hipótesis nula no debe ser rechazada. Sin embargo, hay un número infinito de valores en el intervalo (suponiendo medidas continuas) y ninguno de ellos tampoco puede ser rechazado.

PREGUNTAS

1. La hipótesis nula para un experimento particular es que la media es 20. Si el intervalo de confianza al 99% es (18, 24), ¿puedes rechazar la hipótesis nula al nivel de $\alpha = 0.01$?

a) sí.

b) no.

2. Selecciona lo que aplica: ¿cuál de los intervalos de confianza al 95% para la diferencia entre medias representa una diferencia significativa a un nivel de 0.05?

a) (−4.6, −1.8)

b) (−0.2, 8.1)

c) (−5.1, 6.7)

d) (3, 10.9)

3. Si un intervalo de confianza al 95% contiene al cero, también lo contendrá al 99 %.

a) verdadero.

b) falso.

4. Selecciona todo lo que aplique. Estás interesado en probar que una moneda que usa un mago está sesgada. Lanzas muchas veces la moneda y registras el número de águilas obtenidas. Después de analizar los resultados del experimento, el valor de *p* resultó ser 0.21, por lo que concluyes que no hay pruebas de que la moneda no sea legal. Con base en esta información, ¿cuáles son los intervalos de confianza posibles al 95% para la proporción poblacional de águilas?

a) (0.43, 0.55).

b) (0.32, 0.46).

c) (0.48, 0.64).

d) (0.76, 0.98).

e) (0.81, 1.33).

Conceptos falsos

Es muy común tener conceptos erróneos en las pruebas de significancia. En esta sección revisaremos tres conceptos falsos importantes.

1. Concepto erróneo: El valor de probabilidad es la probabilidad de que la hipótesis nula sea falsa.

Interpretación apropiada: El valor de probabilidad es la probabilidad de obtener un resultado o uno mayor, dado que la hipótesis nula es verdadera. Es la probabilidad de obtener un valor bajo la hipótesis nula. No es la probabilidad de que la hipótesis nula sea falsa.

2. Concepto erróneo: Un valor de probabilidad pequeño indica un efecto grande.

   Interpretación apropiada: Un valor pequeño de probabilidad indica que el resultado proporcionado por la muestra o uno mayor, es muy improbable de obtener si la hipótesis nula es cierta. Un valor pequeño de probabilidad puede ocurrir con efectos muy pequeños, en particular si el tamaño de la muestra es grande.

3. Concepto erróneo: Un resultado no significativo quiere decir que la hipótesis nula es cierta.

   Interpretación apropiada: Un resultado no significativo indica que los datos no demuestran en forma concluyente que la hipótesis nula es falsa.

PREGUNTAS

1. Un valor bajo de la probabilidad *p* indica un efecto grande.
   a) verdadero.
   b) falso.
2. El valor de la probabilidad *p* indica la probabilidad de la hipótesis nula.
   a) verdadero.
   b) falso.
3. Un resultado no significante quiere decir que a partir de esos datos no se puede concluir que la hipótesis nula es falsa.
   a) verdadero.
   b) falso.
4. El valor de la probabilidad *p* es la probabilidad de que la hipótesis nula es falsa.
   a) verdadero.
   b) falso.

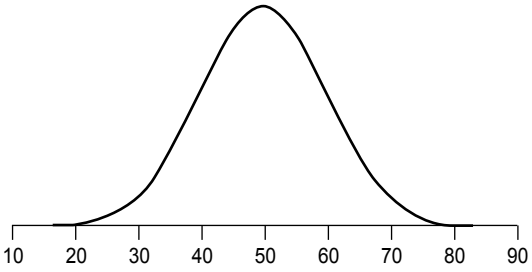
Actividades

I. Contesta y resuelve los siguientes ejercicios para reafirmar los conceptos.

1. Se realizó un experimento con el fin de probar la afirmación de que James Bond puede reconocer la diferencia entre un Martini mezclado y uno agitado. ¿Cuál es la hipótesis nula? ¿Por qué los experimentadores al probar una hipótesis suponen que es falsa?
2. Establece la hipótesis nula para:
   a) un experimento que se realiza para probar si la longitud de un vegetal disminuye con el frío.
   b) un estudio de correlación para determinar la relación entre el tamaño del cerebro y la inteligencia.
   c) una investigación sobre un psíquico impostor que asegura predecir el resultado del lanzamiento de una moneda.

d) un estudio realizado a fin de comparar una droga con un placebo en su efectividad para el alivio del dolor. (Utiliza pruebas de una sola cola).

3. Supón que la hipótesis nula es $\mu = 50$ y que la gráfica que se muestra en seguida es la distribución muestral de la media $\overline{x}$. ¿El valor de $\overline{x} = 60$ obtenido en una muestra es significativo en una prueba bilateral con un nivel de 0.05? ¿Aproximadamente qué valor de $\overline{x}$ sería necesario para detectar significancia?



4. Un investigador desarrolla una nueva teoría que sostiene que los vegetarianos tienen mayor cantidad de una vitamina particular en la sangre. Un experimento se lleva a cabo y los vegetarianos realmente tienen más cantidad de la vitamina, pero la diferencia no es significativa. El valor de probabilidad es 0.13. ¿La confianza del experimentador en su teoría debe aumentar, disminuir, o mantenerse igual?

5. Un investigador supone que la disminución del colesterol que se asocia a la pérdida de peso, en realidad se debe al ejercicio. Para probar lo anterior, el investigador controla con cuidado el tipo de ejercicio, mientras compara los niveles de colesterol de un grupo de personas que pierden peso debido a una dieta con el nivel de colesterol de un grupo control que no hace la dieta. La diferencia en el colesterol entre estos grupos no es significativa. ¿Puede el investigador concluir que la pérdida de peso no tiene ningún efecto en la disminución del colesterol?

6. Se realiza una prueba de significancia y se encuentra un valor de *p* igual a 0.20. ¿Por qué el experimentador no puede afirmar que la probabilidad de que la hipótesis nula es verdadera es igual a 0.20?

7. Para que una droga pueda ser aprobada por las autoridades de salud, debe demostrar seguridad y efectividad. Si la droga es significativamente más efectiva que un placebo, entonces es considerada efectiva. ¿Qué puede decir acerca de la efectividad de una droga que ha sido aprobada? (Suponiendo que no se cometió un error Tipo I).

8. ¿Cuándo es válido utilizar una prueba de una cola? ¿Cuál es la ventaja de utilizar una prueba de una cola? Da un ejemplo de una hipótesis nula que sería probada con una prueba de una cola.

9. Distingue entre el valor de probabilidad y el nivel de significancia.

10. Supón que se realizó un estudio para investigar la efectividad de un curso de preparación sobre "Cómo realizar exámenes de admisión". Se compararon los puntajes en un examen de admisión, obtenidos por un grupo de personas que tomaron el curso (grupo experimental) y los obtenidos por un grupo control. En cada grupo participaron 100 personas. El puntaje medio del grupo experimental fue 503 y el del grupo control fue 499. La diferencia entre los

puntajes medios fue significativa, $p = 0.037$. ¿Cuáles son sus conclusiones acerca de la efectividad del curso?

11. ¿Es más conservador usar un nivel α de 0.01 o un nivel a de 0.05? ¿El valor de β será mayor para una α de 0.05 o para un α de 0.01?

12. ¿Por qué H_o: $x_1 - x_2$, no es una hipótesis nula apropiada?

13. Un experimentador espera que un efecto ocurra en una determinada dirección. ¿Justifica esta suposición el uso de una prueba unilateral? ¿Por qué sí o por qué no?

14. ¿En qué se diferencia el error Tipo I y el Tipo II en las pruebas de una y de dos colas?

15. La probabilidad en una prueba bilateral es 0.03. ¿Cuál sería la probabilidad en una cola si el efecto estuviera en la dirección especificada? ¿Cuál sería si el efecto estuviera en la otra dirección?

16. Escoges un nivel α de 0.01 y luego analizas tus datos. *a*) ¿Cuál es la probabilidad de cometer el error Tipo I, si la hipótesis nula es verdadera? *b*) ¿Cuál es la probabilidad de cometer el error Tipo I, si la hipótesis nula es falsa?

17. Juegas en un casino. Tienes que sacar una de 4 cartas, y el casino asegura que la probabilidad de ganar es de ¼. Piensas que esta probabilidad es menor y decides probar tu hipótesis. Observas a muchas personas que están jugando, y notas que solamente 2 de cada 20 ganan. *a*) Suponiendo que la probabilidad de ganar es realmente 0.25, ¿cuál es la probabilidad de que dos personas o menos ganen? *b*) ¿Se puede rechazar la hipótesis nula con un nivel de 0.05?

18. Crees que la moneda que un mago utiliza está cargada, pero no estás seguro si caerán más águilas o soles. Observas al mago tirar la moneda y registras en qué porcentaje caen águilas. *a*) ¿Realizarías una prueba unilateral o bilateral? *b*) Suponiendo que la moneda es legal, ¿cuál es la probabilidad de que en 30 lanzamientos, caiga un lado 23 veces o más? *c*) ¿Se puede rechazar la hipótesis nula con un nivel de 0.05? ¿Qué sucedería con un nivel de 0.01?

19. ¿Por qué no tiene sentido probar la hipótesis de que la media muestral es por ejemplo 42?

20. Falso o verdadero: es más fácil rechazar la hipótesis nula si el investigador usa un nivel α más pequeño.

21. Falso o verdadero: se tiene una mayor probabilidad de cometer un error Tipo I cuando se utiliza una muestra pequeña.

22. Falso o verdadero: se acepta la hipótesis alternativa cuando se rechaza la hipótesis nula.

23. Falso o verdadero: no se acepta la hipótesis nula cuando se rechaza la alternativa.

24. Falso o verdadero: un investigador tiene riesgo de cometer el error Tipo I cada vez que rechaza una hipótesis nula.

5 Pruebas de hipótesis

La mayoría de los experimentos se diseñan para comparar las medias. Los experimentos pueden planearse para comparar una sola media con un valor especificado. O se puede diseñar el experimento para probar las diferencias de distintas condiciones experimentales. En este capítulo se presenta el método para realizar las pruebas de significancia para comparar una media con un valor dado y los métodos para comparar las medias de dos condiciones experimentales.

Prueba para una sola media

Esta sección muestra cómo probar la hipótesis nula cuando ésta establece que la media de la población es igual a un valor hipotético. Por ejemplo, supongamos que un profesor de estadística utiliza un nuevo método de enseñanza utilizando el empleo de applets, los cuales son pequeños programas en Java que permiten al estudiante tener un ambiente interactivo, y mediante el empleo de éstos poder construir el conocimiento de diferentes conceptos. Para este fin, el maestro seleccionó a nueve estudiantes al azar y un tema del programa les fue impartido con este nuevo método. El promedio histórico de calificaciones obtenidas en este tema es de 6 con una desviación estándar de 1.6. La pregunta es: si aumenta el promedio de calificaciones obtenidas por los estudiantes con este nuevo método es 6. En otras palabras, la hipótesis nula es que la media de la población, μ, es igual a 6. Observa que la hipótesis de investigación es que el promedio de calificaciones sea mayor a 6 y recuerda que en secciones anteriores hemos mencionado que quisiéramos que se rechazara la hipótesis nula. Las calificaciones obtenidas por los nueve estudiantes se muestran en la tabla 5.1 y puedes observar que la media de la muestra, $\overline{x}$, es de 7. Por tanto, la diferencia entre la media de la muestra y la media hipotética de la población es 1.

Calificaciones con el nuevo método de enseñanza
4.0
5.0
5.0
7.0
7.0
7.0
8.0
10.0
9.0

Tabla 5.1 Frecuencias de las medias para $n = 2$.

La prueba de significancia consiste en calcular la probabilidad de que la media de la muestra se aleje de la media de la población μ, en 1 (la diferencia entre la media de la muestra y el valor hipotético de la media poblacional) o más. El primer paso es determinar la distribución muestral de la media. Como se ha visto, la media y la desviación estándar de la distribución muestral de la media son:

$$\mu_{\bar{x}} = \mu$$

y

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

respectivamente. Es claro que $\mu_{\bar{x}} = 6$. Con el fin de calcular la desviación estándar de la distribución muestral de la media, debemos conocer la desviación estándar de la población, σ. En la práctica, es muy difícil que se conozca el valor de σ y entonces lo tenemos que estimar con los datos de la muestra. Sin embargo, para fines didácticos, veamos cómo se calcula el valor de la probabilidad si se conoce σ, antes de proceder al cálculo de la probabilidad cuando lo tenemos que estimar.

En este ejemplo conocemos que los valores históricos del promedio y de la desviación estándar de las calificaciones obtenidas por los estudiantes en el tema del programa seleccionado fueron 6 y 2.0 respectivamente. Para un valor de σ de 2.0 y $n = 9$, la desviación estándar de la distribución muestral de la media es $2.0/3 = 0.666$. Recuerda que la desviación estándar de la distribución muestral se llama error estándar.

Deseamos conocer la probabilidad de que la media de una muestra sea igual o mayor a 7, cuando la distribución muestral de la media tiene una media de 6 y una desviación estándar de 0.666. Para calcular esta probabilidad, recordemos que la distribución muestral de la media tiene una distribución aproximadamente normal. Entonces, utilizando la tabla de la distribución normal, tenemos la figura 5.1.

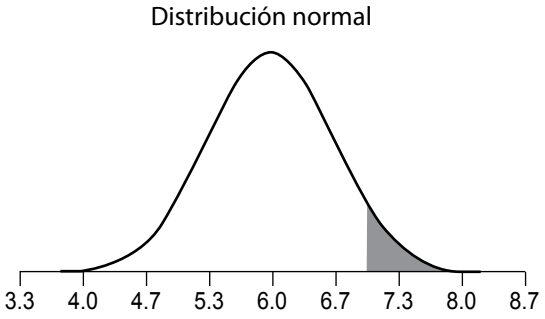


Figura 5.1 Probabilidad de que la media de la muestra sea igual o mayor a 7.

Entonces, se calcula el valor de z

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

donde z es el valor estandarizado de la distribución normal, $\bar{x}$ es la media de la muestra, μ es el valor hipotético de la media de la población, y $\sigma_{\bar{x}}$ es el error estándar de la media. Para nuestro ejemplo,

$$z = (7 - 6)/0.666 = 1.5$$

Usando la tabla de la distribución normal para un valor de $z = 1.5$, obtenemos un valor de 0.9332, por lo que el valor del área en la cola es igual a $1 - 0.9332 = 0.0668$.

En la figura 5.1 se muestra sombreada el área de interés.

Por tanto, la probabilidad de obtener una media muestral igual o mayor a 7 es 0.0668. Es decir, que una media igual o mayor a 7 no es improbable de obtener bajo la hipótesis de que el promedio de calificaciones es igual a 6, por lo que decimos que la diferencia detectada $(7 - 6 = 1)$ no es significante y la hipótesis nula no se rechaza.

Observa que el valor de la probabilidad 0.0668 es mayor que el nivel de significancia de 0.05 y, por tanto, los datos no proporcionan indicios de que la diferencia sea significante.

La prueba anterior fue de una sola cola porque se calculó la probabilidad de que la media de la muestra fuera mayor que la media hipotética en un punto o más, por lo que se calculó la probabilidad del área mayor o igual a 7.

En una prueba de hipótesis de dos colas, debes calcular la probabilidad de que la media de la muestra difiera en un punto o más, en cualquier dirección de la media hipotética, que en nuestro caso es igual a 6. Esto quiere decir que debes calcular la probabilidad de que la media sea igual o menor a 5, e igual o mayor a 7. Las áreas a calcular se muestran en la figura 5.2.

Al observar en la figura 5.2 el área de las dos colas será dos veces el área anteriormente calculada, es decir 0.1336.

Como puedes haber notado, en un análisis real de datos es muy raro que conozcas el valor de σ. Por lo general, σ no es conocida y se estima con la desviación estándar de la muestra s, y $\sigma_{\bar{x}}$ se estima con $s_{\bar{x}}$.

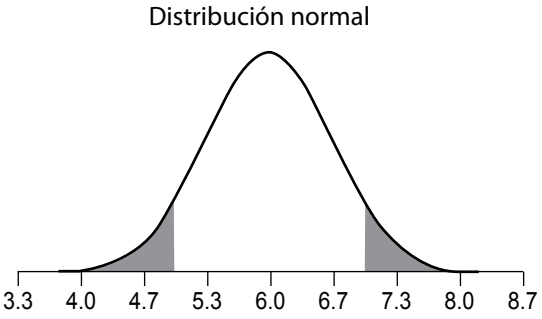


Figura 5.2 Probabilidad de que $\overline{x} \geq 7$ o $\overline{x} \leq 5$.

Consideremos otro ejemplo: en varias colonias de un municipio del Estado de México de alta densidad de población, ante los altos índices de robos y criminalidad los vecinos pusieron en marcha un programa de vigilancia al cual denominaron "*el vecino siempre vigila*". El número de delitos, antes y después del programa, registrados para diez colonias se muestra en la tabla 5.2. Es de interés particular la columna "diferencia", que muestra la diferencia en el número de delitos antes y después del programa para cada una de las colonias. Los valores son positivos cuando hubo más delitos antes del programa y negativo en caso contrario. Si el programa tuvo un efecto positivo, entonces la media de las diferencias en la población será positiva (recuerda que la diferencia que estamos estudiando es, d = antes − después). La hipótesis nula es que la media de las diferencias de los delitos antes y después en la población es 0 (por supuesto, nuestro mejor deseo es que se rechace).

La hipótesis de investigación es suponer que los delitos disminuyen después del programa; es decir, que la media de las diferencias en la población sea mayor de 0 (no pierdas de vista que d = antes − después, y si el programa tuvo efecto, supones que el número de delitos es menor después y por tanto la diferencia es positiva).

Para probar la hipótesis nula, se calcula t, usando un caso especial de la siguiente fórmula general:

$$t = \frac{\text{estadístico} - \text{valor hipotético}}{\text{estimación del error estándar del estadístico}}$$

El caso especial de esta fórmula aplicable a la prueba de una sola media es:

$$t = \frac{\overline{x} - \mu}{s_{\overline{x}}}$$

Donde t es el valor de "t", que se calcula para la prueba de significancia, $\overline{x}$ es la media de la muestra, μ es el valor hipotético de la media de la población, y $s_{\overline{x}}$ es la estimación del error estándar de la media. Observa que esta fórmula es similar a la de z.

En el ejemplo anterior asumimos que la distribución de los valores es normal. En este caso supondremos que la distribución de la diferencia de los valores también es normal.

Colonia	Antes	Después	Diferencia
Estrella	57	62	5
Lomas Altas	27	49	22
Atlanta	32	30	−2
Río Hondo	31	34	3
Cofradías	34	38	4
Parques	38	36	−2
Barrio Bajo	71	77	6
San Marcos	33	51	18
Olimpia	34	45	11
Loma Bonita	53	42	−11
Río Seco	36	43	7
El Salitre	42	57	15
Sta. Bárbara	26	36	10
La Cañada	52	58	6
Pastores	36	35	−1
El Ejido	55	60	5
San Sebastián	36	33	−3
La Concha	42	49	7
Potrero	36	33	−3
La Viga	54	59	5
Puente Viejo	34	35	1
La Colmena	29	37	8
Colina Verde	33	45	12
Monte Seco	33	29	−4

Tabla 5.2 Número de delitos cometidos antes y después del programa de vigilancia.

La media $\overline{x}$, de las $n = 24$ diferencias es 4.9583, el valor hipotético de μ es 0, y la desviación estándar s es 7.5382. La estimación del error estándar de la media es:

$$s_{\overline{x}} = \frac{s}{\sqrt{n}} = \frac{7.5382}{\sqrt{24}} = 1.54$$

Entonces, $t = 4.96/1.54 = 3.22$. El valor de la probabilidad para t depende de los grados de libertad. Los grados de libertad son iguales a $n - 1 = 23$. Utilizando la tabla de la distribución t, se determina que la probabilidad de obtener un valor de t igual o menor a –3.22 o igual o mayor a 3.22 es menor que 0.004. Entonces, si el nuevo método no tuviera efecto (hipótesis nula), la probabilidad de encontrar diferencias entre la media de la muestra y la media hipotética poblacional (en cualquier dirección) tan grandes o mayores como la encontrada es muy pequeña. De esta manera, la hipótesis nula que indica que la media poblacional de las diferencias es cero se puede rechazar. La conclusión es que la media de la población de los delitos después es menor que la media antes del programa de vigilancia.

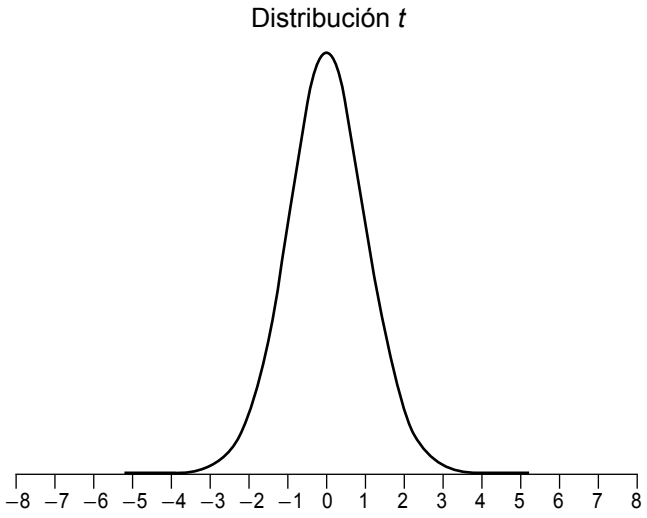


Figura 5.3

Revisión de los supuestos:

1. Cada uno de los valores se selecciona independientemente de cualquier otro valor.
2. Los valores se muestrean de una población normal.

PREGUNTAS

1. Debes hacer la prueba de "*z*" en lugar de "*t*", cuando:
   a) los datos se seleccionan de una distribución normal.
   b) el tamaño de la muestra es grande.
   c) la desviación estándar de la muestra es conocida.
   d) la desviación estándar de la población es conocida.
2. Supón que conoces que la desviación estándar es igual a 10 y que la distribución es normal. Seleccionas 16 elementos de la población y encuentras que la media es igual a 26. Calcula el valor de *p*, para una prueba de dos colas, si la hipótesis nula establece que la media es igual a 20.
3. ¿Cuál es la desviación estándar de esta muestra de datos? –2, 1,3, 2, –1, 0, 4, 6.
4. ¿Cuál es el estimador del error estándar con base en los siguientes datos? –2, 1, 3, 2, –1, 0, 4, 6.
5. ¿Cuál es el valor de la prueba "*t*", si la hipótesis nula establece que la media es igual a 0? Considera los siguientes datos: –2, 1, 3, 2, –1, 0, 4, 6.
6. ¿Cuál es el valor de la probabilidad *p* para una prueba de dos colas si la hipótesis nula establece que la media es igual a 0? Considera los siguientes datos: –2, 1, 3, 2, –1, 0, 4, 6.

Diferencias entre dos medias (grupos independientes)

Los investigadores por lo general están más interesados en conocer la diferencia entre las medias que un valor específico de ellas. En esta sección revisamos cómo

se realiza la prueba para la diferencia entre las medias de dos grupos diferentes. En una sección posterior describiremos cómo realizar la prueba para la diferencia entre medias para dos condiciones, cuando se tiene un solo grupo, pero en cada elemento del grupo se prueban las dos condiciones.

Tomemos como ejemplo la encuesta para determinar la cantidad de cerveza (número de cervezas) que consumen los viernes los estudiantes de una universidad. En la tabla 5.3 se resumen, para el grupo de varones y mujeres, el tamaño de las muestras, las medias y las varianzas.

Género	*n*	Media	Varianza
Hombres	17	5.353	2.743
Mujeres	17	3.882	2.985

Tabla 5.3 Medias y varianzas de una encuesta referente al consumo de cerveza.

Como se puede observar, los hombres consumen más cerveza que las mujeres. La diferencia entre la media muestral de los hombres de 5.35 y la de las mujeres de 3.88 es 1.47. Sin embargo, la diferencia por género en estas muestras particulares no es lo que nos importa. Lo que es de nuestro interés es la diferencia entre las medias poblacionales.

Con el fin de probar si hay diferencia entre las medias poblacionales, tenemos que hacer tres suposiciones:

1. Las dos poblaciones tienen la misma varianza. Esta suposición es conocida como la suposición de homogeneidad de varianzas.
2. Las poblaciones se distribuyen en forma normal.
3. Cada dato se selecciona en forma independiente de cualquier otro. Esta suposición requiere que cada elemento proporcione sólo un valor. Si un elemento proporciona dos valores, entonces éstos no son independientes. El análisis de los datos para el caso de dos valores por elemento se revisa en la sección sobre prueba *t* para datos correlacionados en este mismo capítulo.

Por ahora, es suficiente decir que violaciones moderadas de las suposiciones 1 y 2 no producen mucha diferencia. Es importante no violar la tercera suposición.

Hemos visto en la sección "Prueba para una sola media", que la fórmula general para la prueba de significancia es:

$$t = \frac{\text{estadístico} - \text{valor hipotético}}{\text{estimación del error estándar del estadístico}}$$

En este caso, nuestro estadístico es la diferencia entre las medias muestrales, y nuestro valor hipotético es 0. El valor hipotético es el valor que indica la hipótesis nula, es decir, indica que la diferencia entre las medias poblacionales es 0.

Realizaremos la prueba de significancia para la diferencia entre las calificaciones de los hombres y las de las mujeres. Para realizar la prueba, hacemos las tres suposiciones indicadas anteriormente.

El primer paso es calcular el estadístico, el cual simplemente es la diferencia entre las medias muestrales:

$$\bar{x}_1 - \bar{x}_2 = 5.35 - 3.88 = 1.47$$

Ya que el valor hipotético es 0, no necesitamos restarle nada al estadístico para encontrar el valor del numerador de *t*.

El siguiente paso es calcular un estimador para el error estándar del estadístico. En este caso, el estadístico es la diferencia entre las medias muestrales, por lo que el estimador del error estándar del estadístico es $s_{\bar{x}_1 - \bar{x}_2}$. Recuerda que en la sección de la distribución muestral para la diferencia entre medias vimos que la fórmula para el error estándar de la diferencia entre medias en la población viene dada por:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}} = \sqrt{\frac{2\sigma^2}{n}}$$

Para estimar esta cantidad, estimamos a su vez a σ^2 y usamos esa estimación en lugar de σ^2. Ya que asumimos que las varianzas poblacionales son iguales, estimamos la varianza con el promedio de las varianzas muestrales:

$$CME = \frac{s_1^2 + s_2^2}{2}$$

donde *CME*, cuadrado medio del error, es nuestra estimación de σ^2. En este ejemplo,

$$CME = (2.743 + 2.985)/2 = 2.864$$

Ya que *n* (el número de datos para cada condición o grupo) es 17, tenemos:

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{2CME}{n}} = \sqrt{\frac{(2)(2.864)}{17}} = 0.5805$$

El siguiente paso es calcular *t*, sustituyendo los valores calculados anteriormente, en la fórmula:

$$t = 1.47/0.5805 = 2.533$$

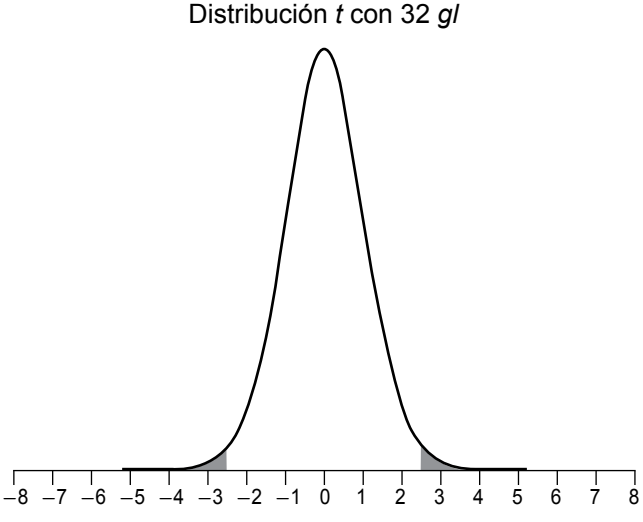


Figura 5.4 Probabilidad para la prueba de dos colas.

Finalmente, se calcula la probabilidad de obtener un valor de *t* igual o mayor a 2.53 o igual o menor a –2.53. Para calcular el valor de la probabilidad, necesitamos conocer los grados de libertad. Los grados de libertad son el número de estimaciones independientes de la varianza, con las cuales se calculó el *CME*. Éstas son iguales a $(n_1 - 1) + (n_2 - 1)$ donde n_1 es el tamaño de la muestra del primer grupo y n_2 es el tamaño de la muestra del segundo. Para nuestro ejemplo, $n_1 = n_2 = 17$. Cuando $n_1 = n_2$, por convención se usa "*n*" para referirse al tamaño de muestra de cada grupo. Por tanto, los grados de libertad son $16 + 16 = 32$.

Una vez que tenemos los grados de libertad, podemos usar la tabla de la distribución *t* para calcular el valor de la probabilidad. El valor de la probabilidad para una prueba de dos colas es menor a 0.02 y este valor corresponde a las áreas sombreadas que se muestran en la figura 5.5. La prueba de dos colas

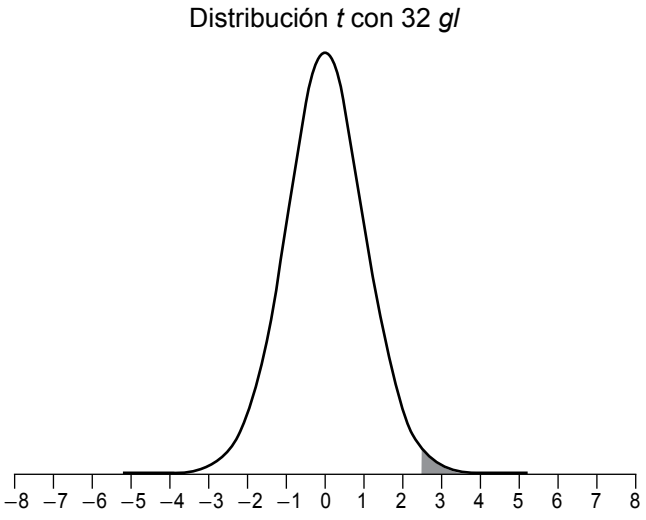


Figura 5.5 Probabilidad para la prueba de una cola.

se usa cuando la hipótesis nula puede ser rechazada sin considerar la dirección del efecto. Es, como se muestra en la figura 5.5, la probabilidad de obtener un valor de $t < -2.533$ o > 2.533.

Los resultados para la prueba de una cola se muestran en la figura 5.6. Como puedes ver, el valor de la probabilidad es la mitad del valor de la probabilidad obtenida en la prueba de dos colas.

La mayoría de los programas estadísticos en computadora requieren que los datos se proporcionen en una forma específica a fin de procesar pruebas de *t*. Considera los datos de la tabla 5.4.

Grupo 1	Grupo 2
3	5
4	6
5	7

Tabla 5.4 Datos de ejemplo.

Grupo	Y
1	3
1	4
1	5
2	5
2	6
2	7

Tabla 5.5 Datos formateados.

Hay dos grupos, cada uno con tres observaciones. El formato de los datos para su uso en computadora, por lo general, consiste en identificar los datos del primer grupo con un 1 y a los del segundo con un 2, como se muestra en la tabla 5.5.

Los cálculos cambian cuando los tamaños de muestra son diferentes. Una suposición importante es considerar que en la estimación de la varianza, *CME*, tiene mayor influencia la muestra de mayor tamaño. Se calcula la suma de las desviaciones cuadradas como sigue:

$$SCE = \sum(x - \bar{x}_1)^2 + \sum(x - \bar{x}_2)^2$$

donde $\bar{x}_1$ es la media del grupo 1 y $\bar{x}_2$ es la media del grupo 2. Consideremos el ejemplo, con unos cuantos datos para realizar en forma rápida los cálculos (véase tabla 5.6):

$$\bar{x}_1 = 4 \text{ y } \bar{x}_2 = 3.$$

$$SCE = (3-4)^2 + (4-4)^2 + (5-4)^2 + (2-3)^2 + (4-3)^2 = 4$$

Entonces, el *CME* se calcula de la forma siguiente:

$$CME = \frac{SCE}{gl}$$

donde los grados de libertad, *gl,* se calculan como antes:

$$gl = (n_1 - 1) + (n_2 - 1) = (3-1) + (2-1) = 3$$

$$CME = SCE/gl = 4/3 = 1.333$$

La fórmula

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{2CME}{n}}$$

Grupo 1	Grupo 2
3	2
4	4
5	

Tabla 5.6 Datos de ejemplo.

se reemplaza por:

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{CME}{n_1} + \frac{CME}{n_2}} = \sqrt{\frac{1.333}{3} + \frac{1.333}{2}}$$

$$s_{\bar{x}_1 - \bar{x}_2} 1.054$$

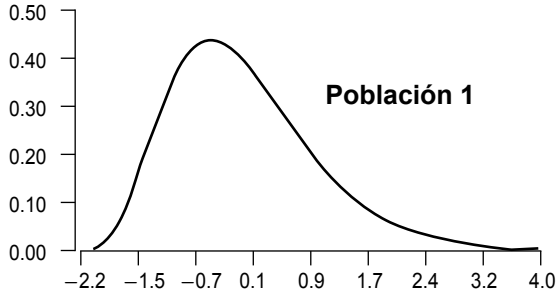
Entonces:

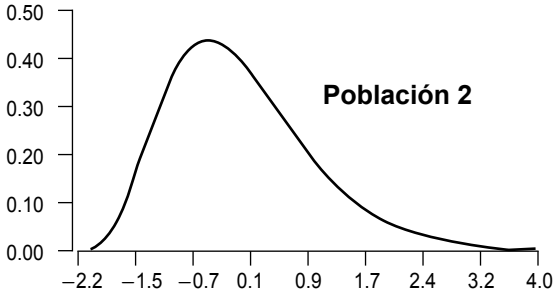
$$t = (4 - 3)/1.054 = 0.949$$

y el valor de p para la prueba de dos colas es mayor a 0.2.

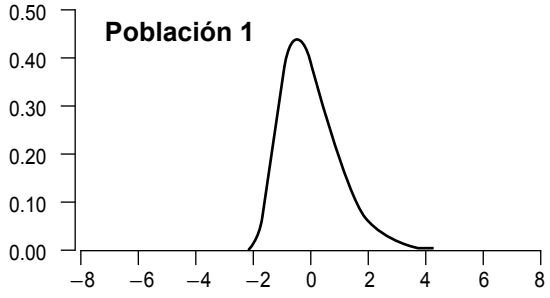
PREGUNTAS

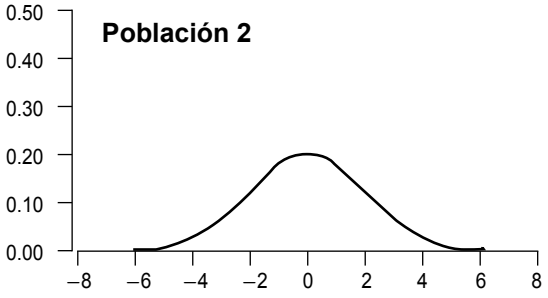
1. Las gráficas muestran una violación a la suposición de:
a) normalidad.
b) homogeneidad de la varianza.



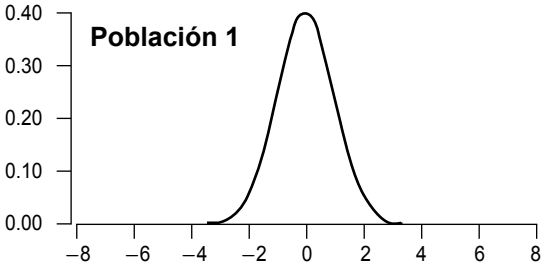


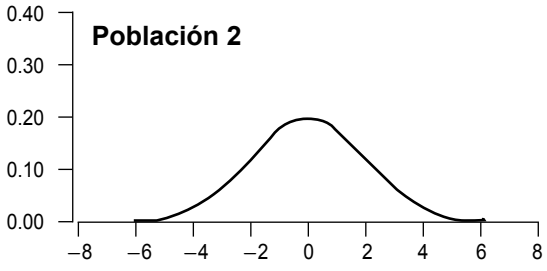
2. Las gráficas muestran una violación a la suposición de:
a) normalidad.
b) homogeneidad de la varianza.





3. Las gráficas muestran una violación a la suposición de:
a) normalidad.
b) homogeneidad de la varianza.





4. En la fórmula de "*t*" el estadístico es:

a) la hipótesis nula.

b) la media de todos los datos.

c) la diferencia entre las medias de las muestras.

d) el nivel de significancia.

5. En la fórmula para *t* el "valor hipotético" es:

a) el valor esperado de *t.*

b) la diferencia entre las medias de las poblaciones.

c) el nivel de significancia.

6. Si la hipótesis nula es que las dos medias de las poblacionales son iguales, entonces el valor hipotético es:

a) cero.

b) la media de la población.

7. El denominador en la prueba de "*t*" es:

a) el estimador del error estándar.

b) el estimador del error estándar de la diferencia entre las medias.

c) SCE/2.

8. Si hay 4 datos por grupo y el valor de "*t*" es 2.34, ¿cuál es el valor de la probabilidad para una prueba de dos colas? (Responde con tres decimales).

a) Respuesta _____.

9. ¿Cuál es el valor de "*t*" para la diferencia entre las medias para estos datos?

Rango	66	44	39	44	39	33	31	51
Frecuencia	43	45	40	43	29	23	41	40

a) Respuesta _____.

Diferencia entre dos medias (pares correlacionados)

Consideremos cómo realizar el análisis estadístico de los datos del ejemplo del programa de vigilancia. Los datos se refieren al número de delitos en diferentes colonias de un municipio muy poblado, antes y después del programa. En la tabla 5.7

se muestran los datos. Es de interés particular la columna "diferencia", que muestra la diferencia en el número de delitos para cada una de las condiciones (antes y después de implantar el programa de vigilancia). La primera pregunta es por qué la diferencia entre las medias no se prueba utilizando el procedimiento descrito en la sección "Diferencia entre dos medias (grupos independientes)". La respuesta se apoya en el hecho de que en este experimento no tenemos grupos independientes. El número de delitos antes y después se contabilizaron a las mismas colonias. Es decir, cada colonia nos proporciona dos respuestas (número de delitos) para cada condición o tratamiento (sin programa y con programa).

En la figura 5.7 se muestra un diagrama de dispersión para el número de delitos antes y después del programa de vigilancia para cada colonia. Es claro que las colonias que tenían mayor criminalidad tienden a tener mayor criminalidad aun después de implantado el programa de vigilancia. La correlación entre el número de delitos antes y después es alta: $r = 0.80$. Es claro que las dos variables no son independientes.

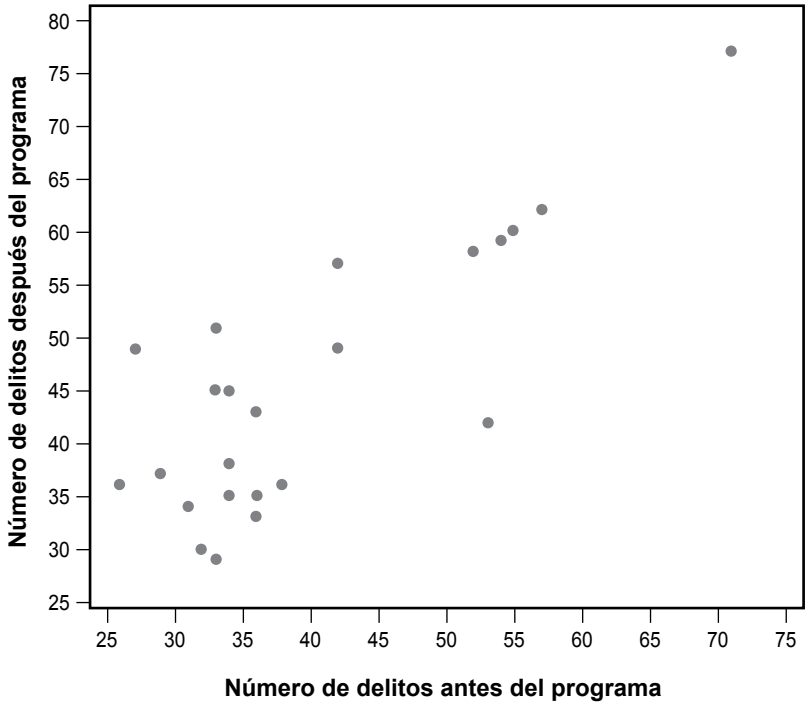


Figura 5.6 Diagrama de dispersión para el número de delitos antes y después del programa de vigilancia.

El procedimiento para la prueba *t*, para la diferencia entre medias para pares, se desarrolló en la sección "Prueba para una sola media". El procedimiento es: calcular la diferencia entre el número de delitos cometidos antes y después, y probar si la media de estas diferencias es significativamente diferente de 0. En la tabla 5.7 se muestran las diferencias. Como se vio en la sección "Prueba para una sola media", la media de las diferencias es 4.96, la cual es significativamente diferente de 0; $t = 3.22$, $gl = 23$, $p < 0.04$. Esta prueba *t* se conoce con varios nombres: Prueba de *t* para grupos correlacionados y Prueba *t* para datos por pares.

Si hubieras utilizado equivocadamente para analizar estos datos el procedimiento de la prueba *t* para grupos independientes, habrías encontrado: $t = 1.42$,

Colonia	Antes	Después	Diferencia
Estrella	57	62	5
Lomas Altas	27	49	22
Atlanta	32	30	−2
Río Hondo	31	34	3
Cofradías	34	38	4
Parques	38	36	−2
Barrio Bajo	71	77	6
San Marcos	33	51	18
Olimpia	34	45	11
Loma Bonita	53	42	−11
Río Seco	36	43	7
El Salitre	42	57	15
Sta. Bárbara	26	36	10
La Cañada	52	58	6
Pastores	36	35	−1
El Ejido	55	60	5
San Sebastián	36	33	−3
La Concha	42	49	7
Potrero	36	33	−3
La Viga	54	59	5
Puente Viejo	34	35	1
La Colmena	29	37	8
Colina Verde	33	45	12
Monte Seco	33	29	−4

Tabla 5.7 Número de delitos cometidos antes y después del Programa de Vigilancia.

$gl = 46$, y $p = 0.15$. Por tanto, la diferencia entre las medias no es estadísticamente significante. Las pruebas de *t* para pares siempre tienen mayor potencia que la prueba *t* para grupos independientes. Esto se debe a que en la prueba *t* para pares, cada diferencia es una comparación de las dos condiciones (o tratamientos), evaluada en la misma unidad experimental (colonia, individuo, etc.). Esto hace que cada unidad experimental sea "su propio control" y que las diferencias entre ellas se bloqueen. El resultado es que el error estándar de la diferencia entre medias es menor en la prueba *t* para pares (grupos correlacionados) y debido a que este término es el denominador de la fórmula de *t*, entonces *t* resultará mayor para este tipo de prueba.

Detalles acerca del error estándar de la diferencia entre las medias

Para ver por qué el error estándar de la diferencia entre las medias es menor en una prueba de *t* para grupos correlacionados, consideremos la varianza de la diferencia de dos variables. De acuerdo con la Ley de la suma de las varianzas, la varianza de la suma o diferencia de dos variables *X* y *Y* está dada por:

$$s^2_{X\pm Y} = s^2_x + s^2_y \pm 2rs_x s_y$$

Entonces, la varianza de la diferencia de dos variables es la varianza del primer grupo (*X*), más la varianza del segundo grupo (*Y*) menos el producto de la correlación por

la desviación estándar de X por la desviación estándar de Y. Para nuestro ejemplo, $r = 0.80$ y en la tabla 5.8 se muestran las varianzas y las desviaciones estándar.

	Antes	Después	Antes-Después
Varianza	128.02	151.78	56.82
Desviación estándar	11.31	12.32	7.53

Tabla 5.8 Varianzas y desviaciones estándar.

La varianza de las diferencias es 56.82, que fue calculado como sigue:

$$128.02 + 151.78 - (2)(0.80)(11.31)(12.32)$$

Observa que cuanto mayor sea el coeficiente de correlación, menor será el error estándar de la media.

PREGUNTAS

1. Considera un diseño experimental de pares correlacionados. El experimentador analizó los datos con una prueba *t* para pares correlacionados e incorrectamente realizó una prueba *t* para muestras independientes. Encontró significancia en la prueba *t* para pares correlacionados y en la prueba *t* de muestras independientes no encontró significancia. El experimentador llegó a la conclusión de que esto se debe a que la diferencia entre las unidades experimentales fue mayor que la diferencia entre las medias de los tratamientos. ¿Es correcta esta conclusión?

a) sí.

b) no.

2. Las pruebas *t* para pares correlacionados tienen mayor potencia que las pruebas *t* para muestras independientes, porque las pruebas para pares correlacionados tienen menor error estándar, lo que da como resultado un mayor valor de *t*.

a) verdadero.

b) falso.

3. Si se incrementa la correlación entre las mediciones:

a) aumenta el valor absoluto de *t*.

b) disminuye el valor absoluto de *t*.

c) no tiene efecto sobre el valor de *t*.

4. La hipótesis nula en pruebas *t* de pares correlacionados es que la media de las diferencias en la población es 0.

a) verdadero.

b) falso.

5. Usando los siguientes datos, ¿cuál es el valor de *t* para una prueba *t* de pares correlacionados?

C1	6.54	12.71	5.68	10.27	18.53	8.46	3.5	9.96
C2	14.31	9.98	8.3	7.55	12.69	7.52	12.25	8.42

Actividades

I. Contesta y resuelve los siguientes ejercicios para reafirmar los conceptos.

1. Los puntajes de una prueba de física realizada por 8 estudiantes seleccionados al azar fueron: 60, 62, 67, 69, 70, 72, 75 y 78.

a) realiza una prueba para decidir si la media de la muestra es significativamente diferente de 65 a un nivel de 0.05. Calcula los valores de t y p.

b) el investigador registra por equivocación un dato como 67 cuando debería haber sido 76. Con los datos corregidos realiza una prueba para decidir si la media de la muestra es significativamente diferente de 65 a un nivel de 0.05.

2. Un experimento (hipotético) se realiza para estudiar el efecto de alcohol sobre la habilidad motora de los individuos. A diez personas se les califica su habilidad motora dos veces, una vez después de consumir dos bebidas alcohólicas y después de consumir dos vasos con agua. El experimento se realizó durante dos días diferentes para dar oportunidad a que el alcohol se degradara. A la mitad de las personas se les dio a consumir primero las bebidas alcohólicas y a la otra mitad el agua. Las calificaciones se muestran en la siguiente tabla. El primer número para cada sujeto es su calificación en la condición "agua". Los puntajes más altos reflejan mayor habilidad motora. Realiza una prueba para determinar si el alcohol tiene un efecto significativo. Calcula los valores de t y p.

Agua	Alcohol
16	13
15	13
11	10
20	18
19	17
14	11
13	10
15	15
14	11
16	16

3. Los puntajes hipotéticos sobre una prueba de vocabulario de un grupo de personas de 20 años y de un grupo de personas de 60 se muestran a continuación.

20 años	60 años
27	26
26	29
21	29
24	29
15	27
18	16
17	20
12	27
13	

a) realiza una prueba de hipótesis para la diferencia de las medias, usando un nivel del 0.05.

b) escribe las suposiciones necesarias para obtener la respuesta.

4. La distribución de muestreo de un estadístico se distribuye normalmente con un error estándar de 12 y 20 grados de libertad. *a*) ¿Cuál es la probabilidad de obtener una media igual o mayor a 107 si el parámetro poblacional es 100? ¿Esta probabilidad es significativa a un nivel de 0.05 (bilateral)? *b*) ¿Cuál es la probabilidad de obtener una media igual o menor a 95 (unilateral)? ¿Es esta probabilidad significativa a un nivel de 0.05?

5. ¿Cómo se puede decidir si usar una prueba de *t* para dos medias independientes o una prueba de *t* para muestras por pares?

6. Los participantes de un juego lanzaron dardos a una diana. En una oportunidad usaron su mano preferida; en la otra oportunidad, usaron la otra mano. El orden en que usaron las manos fue aleatorio. Los datos se muestran a continuación.

Mano preferida	Mano no preferida
12	7
7	9
11	8
13	10
10	9

a) ¿qué clase de prueba *t* debe usarse?

b) realiza una prueba *t* bilateral y calcula el valor de *p*.

c) realiza una prueba *t* unilateral y calcula el valor de *p*.

7. Supón que los datos del problema anterior fueron obtenidos usando dos grupos diferentes de personas: un grupo usó la mano preferida y otro grupo usó la mano no preferida. Analiza los datos y compara los resultados con los del problema anterior.

8. Se realiza un estudio para investigar si los estudiantes son mejores cuando estudian todo de una sola vez o cuando lo hacen en diferentes sesiones. Un grupo de 12 participantes realizó una prueba después de estudiar durante una hora continua. Otro grupo de 12 participantes realizó la prueba después de estudiar por tres sesiones de veinte minutos cada una. El primer grupo obtuvo una media de 75 y una varianza de 120. El segundo grupo obtuvo una media de 86 y una varianza de 100.

a) ¿cuál es el valor de *t* calculada? ¿Son significativamente diferentes las medias de los dos grupos a un nivel de 0.05?

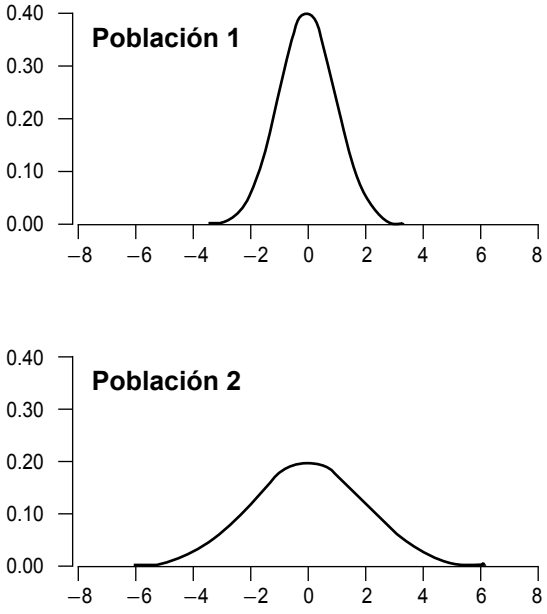
b) ¿qué valor tendría *t* calculada si sólo hubiera 6 participantes en cada grupo? ¿Serían diferentes significativamente las medias a un nivel de 0.05?

9. Una nueva prueba se diseñó con una media de 80 y una desviación estándar de 10. Se seleccionó una muestra aleatoria de 20 estudiantes y se les aplicó la prueba, encontrándose una media de 85. ¿Esta media es diferente significativamente de la media esperada de 80?

10. Se realiza una prueba *t* para una muestra y se obtiene una *t* calculada de 3.0. La media de la muestra fue 1.3 y la desviación estándar 2.6. ¿De qué tamaño fue la muestra?

11. Verdadero/Falso: Las pruebas *t* para muestras por pares siempre tienen mayor potencia que pruebas *t* para muestras independientes.

12. Verdadero/Falso: Las gráficas siguientes representan una violación de la suposición de homogeneidad de las varianzas.



13. Verdadero/Falso: Cuando se realiza una prueba *t* para una sola muestra y se conoce la desviación estándar poblacional, se debe buscar el valor crítico de *t* en las tablas, utilizando los grados de libertad.

II. Resuelve los siguientes ejercicios de aplicación

1. Una máquina automática de Nescafé está ajustada para verter, en vasos, un promedio de 200 ml de café. Para probar si la máquina no ha sufrido desajustes, se toma una muestra aleatoria de 50 vasos llenos, y se encuentra una media de 196 ml, con una desviación típica de 12 ml. Usa un nivel de significación de 5% para determinar si la máquina está o no desajustada.

2. El director de una editorial mexicana debe decidir si publica un texto escrito por un catedrático universitario. Aquél aprobará la publicación si hay pruebas de que al menos 35% de los estudiantes universitarios habrán de adoptar el libro como texto. Al seleccionar una muestra aleatoria de 40 estudiantes universitarios, resulta que 32% adoptarán el libro como texto. A la luz de este resultado, ¿aprobará el director la publicación? Usa un nivel de significación de 0.05.

3. Bicicletas Mexicanas, S. A., necesita obreros calificados para trabajar en una línea de montaje que requiere en promedio 42 segundos para terminar la operación. Un obrero que aspira al trabajo se pone a prueba durante 25 ciclos de esta operación y los resultados son: tiempo total, 1300 segundos con una desviación estándar de 6 segundos. De acuerdo con esta prueba, ¿darías el empleo a este obrero, asumiendo un error alfa de 0.01?

4. El grupo 1301 de la licenciatura en Administración se dividió en dos secciones en la clase de Estadística: una conformada por todas las mujeres y otra por hombres, cada una con 28 estudiantes. Hicieron el mismo examen de distribuciones de probabilidad. Las mujeres obtuvieron una calificación media de 7.2 puntos con una desviación típica de 0.9 puntos, mientras que la media de los hombres fue 8.2 con una desviación típica de 0.5.

a) comprueba la hipótesis de que los hombres son mejores en su rendimiento promedio del conocimiento del tema. Usa un nivel de significancia del 5%.

b) comprueba la hipótesis de que no existe diferencia significativa entre hombres y mujeres en cuanto a su comprensión del tema. Usa un nivel de significancia del 1%.

5. La utilidad por casa vendida varía según el tipo de casa. La utilidad promedio por ventas registradas en el mes pasado fue de 21 000, 30 000, 12 000, 62 000, 45 000 y 51 000 pesos. Con base en esta información, determina si hay suficientes indicadores de que el vendedor ha alcanzado su meta propuesta, que es de 48 000 pesos. Usa un error alfa de 0.01.

6. La empresa Bicicletas Mexicanas, S. A., afirma que al menos 95% de las piezas que fabrica cumplen con las especificaciones requeridas. Si examinamos una muestra aleatoria de 200 piezas y encontramos que 24 de ellas resultaron defectuosas, ¿podemos decir que el dato muestral proporciona suficientes pruebas para rechazar la afirmación del fabricante? Utiliza un error alfa de 0.05.

7. La dirección de la Biblioteca de la FES-C está interesada en saber si ha aumentado el número promedio de libros que cada estudiante se lleva en préstamo por visita. Anteriormente, el promedio era de tres libros, con una desviación típica de un libro. Para comprobarlo se toma una muestra aleatoria de 32 estudiantes, y su media resulta ser de cuatro libros. Determina si realmente ha aumentado el promedio de libros llevados en préstamo por los universitarios. Usa un error alfa de 1%.

8. La máquina de empacado de una empresa de cereales vierte el producto en cajas de tamaño económico. Se efectúan verificaciones constantes de los pesos netos de las cajas para mantener el ajuste de la maquinaria que controla el peso neto. Dos muestras tomadas en dos días presentan la siguiente información: la primera muestra de 19 cajas mostró una media de 470 g con una desviación estándar de 17 g, la segunda muestra de 15 cajas tuvo una media de 498 g con una varianza de 515 g. Determina si es necesario un ajuste a la maquinaria de empacado con un nivel de significación de 0.05.

9. La cadena de tiendas COMEX tiene dos planes de cuentas de cargo disponibles para sus clientes con cuenta de crédito. La gerencia general de la cadena desea recopilar información acerca de cada plan y estudiar las diferencias entre los dos planes. Le interesa conocer el porcentaje de saldos mensuales superiores a 100 pesos. Se seleccionó una muestra aleatoria de 75 cuentas del plan A y 80 del plan B, con los siguientes resultados: de las 75 cuentas del plan A, 28% fueron superiores a 100 pesos; del plan B, el porcentaje fue 25%. A la luz de estos resultados, determina si es significativa la diferencia entre los porcentajes de ambos planes.

10. Un contratista construyó un gran número de casas de aproximadamente el mismo tamaño y el mismo precio, durante el año pasado. Él asegura que el valor promedio de estas casas no excede los 600 000 pesos. Un corredor de bienes raíces seleccionó aleatoriamente 15 de las casas construidas por el contratista el año pasado y encontró que el precio promedio de esta muestra es 615 000 pesos con una desviación típica de 11 250 pesos. Determina si está en lo cierto el contratista, con un nivel de significación de 5%.

11. El jefe del personal de una empresa editora desea determinar el tiempo que necesitan sus empleados para llegar a su trabajo. Una muestra aleatoria de 12

empleados dio un tiempo promedio de 48 minutos con una desviación típica de 11 minutos. Con un error alfa de 0.01, determina si hay pruebas suficientes para afirmar que el tiempo promedio de viaje de los empleados es a lo sumo de 55 minutos.

12. Un gerente de un complejo de cines Cinemex afirma que el 72% de los cinéfilos que acuden al complejo prefieren las películas de acción. De una muestra de 200 cinéfilos que acudieron al cine el miércoles pasado, 160 manifestaron su predilección por las películas de acción. Establece si es cierta la afirmación del gerente, con un nivel de significación de 2%.

13. Un profesor de inglés de la FES-C cree que si se introducen ilustraciones alusivas a los distintos temas de conversación en la enseñanza del idioma inglés, el estudiante adquiere mayor dominio de dicho idioma. Para poner a prueba tal hipótesis, a un grupo de 36 estudiantes se les impartió clases durante un periodo de 10 semanas con ilustraciones y a un grupo similar de 40 estudiantes se le impartieron los mismos temas, pero sin utilizar ilustraciones. Al finalizar el curso, se obtuvieron los siguientes resultados: el grupo experimental "con ilustraciones" obtuvo un total de 2 340 puntos con una desviación típica de 9 puntos, y el grupo control "sin ilustraciones" obtuvo un total de 2 400 puntos con una desviación estándar de 12 puntos. Prueba si esos recursos audiovisuales mejoraron el aprendizaje del idioma inglés. Usa $\alpha = 0.05$.

14. Una compañía constructora está preocupada por el tiempo que se pierde debido a accidentes de trabajo. Por ello, dispuso montar un programa de seguridad para reducir el tiempo perdido debido a dichos accidentes. El programa duró 30 meses y, al finalizar, el tiempo perdido por accidentes tuvo un promedio de 96 h mensuales, con una desviación típica de 15 h. En los 36 meses anteriores al programa de seguridad, el tiempo perdido por accidentes promedió 110 h mensuales con una desviación estándar de 18 h. Determina si fue efectivo el programa de seguridad para disminuir el tiempo perdido por accidentes de trabajo. Usa un nivel de significación de 5%.

15. Se hizo una encuesta sobre hábitos de consumo de cigarrillos en cierta universidad entre estudiantes de un sexo y otro. El gasto promedio y la desviación estándar de los hombres fue de 65 pesos con una desviación estándar de 10 pesos. El de las mujeres fue de 60 pesos con una varianza de 64 pesos. Ambas muestras fueron de tamaño 64.

a) usa un nivel de significación de 0.05 para probar la hipótesis de que no hay diferencia significativa en la cantidad promedio gastada en cigarrillos por los hombres y las mujeres de dicha universidad.

b) prueba la hipótesis de que los hombres gastan más en promedio en el consumo de cigarrillos que las mujeres. Usa un error alfa de 0.01.

16. Se aplicó una encuesta a estudiantes de los últimos semestres de dos carreras de la FES-C con el objeto de conocer la proporción de casados. En la licenciatura en Administración se tomó una muestra aleatoria de 100 estudiantes de los cuales 22% manifestaron estar casados; en la licenciatura en Contaduría, el porcentaje de casados fue 25%, calculado en una muestra de 100 estudiantes.

a) determina si la proporción de casados es mayor en Contaduría, con un nivel alfa del 5%.

17. El supervisor de la oficina del SAT en Cuautitlán Izcalli tiene sospechas de que ha aumentado el número de quejas de los usuarios, por lo que llamó a junta a todo su personal de atención al público y les aseguró que había un promedio de 25 quejas diarias por parte del público. Se hizo un registro del número de quejas ocurridas en nueve días: 28, 24, 24, 20, 23, 19, 23, 18,17, 26.

a) ¿la evidencia muestral demuestra que el supervisor tiene razón, utilizando un nivel de significancia $\alpha = 0.05$?

18. Se desarrolló un nuevo sistema para realizar un montaje en una línea de producción con la esperanza de reducir los tiempos en la operación; 36 obreros desarrollan el trabajo según el nuevo sistema y 34 obreros lo desarrollan según el método anterior. La prueba dio los siguientes resultados sobre los tiempos requeridos para terminar la operación: con el nuevo sistema, se obtuvo una media de 54 minutos y una desviación estándar de 6 minutos, y con el sistema anterior se obtuvo una media de 58 minutos y una desviación estándar de 5 minutos.

a) prueba la hipótesis de que no existe diferencia significativa entre ambos métodos. Usa $\alpha = 0.5$.

b) ¿el nuevo método redujo los tiempos para realizar la operación? Usa $\alpha = 0.5$.

19. Se realizó una prueba para determinar la eficiencia de dos procedimientos, *A* y *B*, para detectar declaraciones con facturas que no son deducibles de impuestos; para ello se tomó una muestra de 26 empleados de una oficina del SAT, dividida en dos grupos de 13. Después de un entrenamiento adecuado en el procedimiento asignado, los resultados, en términos del número de declaraciones que presentan facturas que no son deducibles de impuestos, fueron: con el procedimiento *A* encontraron una media de 28, con una desviación estándar de 4 declaraciones con facturas que no son deducibles de impuestos; con el procedimiento *B* se encontró una media de 24 y una desviación estándar de 6 declaraciones con facturas que no son deducibles de impuestos. Prueba la hipótesis de que el procedimiento *A* es mejor para detectar declaraciones con facturas que no son deducibles de impuestos. Usa un nivel de significancia del 5%.

20. Se desea comprobar si existe una diferencia significativa en la calificación promedio entre los varones y las mujeres de un curso de Estadística; para tal efecto se tomaron muestras de unos y otras del grupo 2 401 y se encontraron los siguientes resultados de un examen:

Hombres	4.4	4.8	2.5	3.7	4.8	9.4	8.4	4.9
Mujeres	3.7	2.2	7.1	8.1	1.5	6.8	6.7	4

a) realiza la prueba con $\alpha = 0.05$.

21. Un gerente de personal piensa que 18% de los empleados de la fábrica La Favorita trabajan horas extras todas las semanas. Si la proporción observada en esta semana es de 13% en una muestra de 250 de los 2 500 empleados, ¿podemos aceptar como razonable su opinión o hemos de concluir que es más adecuado algún otro valor? Usa $\alpha = 0.05$.

22. En un taller de maquinaria, algunos de los obreros ya no reciben sueldo fijo, sino que trabajan a destajo. Para averiguar si esto ha modificado la productividad de los obreros, al supervisor se le pide llevar un registro de la producción diaria (número de piezas terminadas) de cada uno. Con los datos que se incluyen en seguida, prueba en un nivel de significancia del 10% si existen diferencias significativas de productividad en las dos clases de remuneración.

Remuneración	**Producción**											
Sueldo	118	115	122	99	106	125	102	100	92	103	113	129
Destajo	115	126	113	110	135	102	124	137	108	128		

23. La Profeco investiga acusaciones contra una embotelladora porque no llena los refrescos adecuadamente; ha muestreado 40 botellas y ha descubierto que el contenido promedio es de 1.98 litros de líquido. En la propaganda se anuncia que las botellas contienen 2.0 litros de líquido. Se sabe que la desviación estándar del proceso de producción es de 40 ml de líquido. ¿Debe la Profeco llegar a la conclusión de que las botellas no están siendo llenadas correctamente? Usa $\alpha = 1\%$.

24. Un estudio de mercado efectuado en una tienda de Soriana mostró que en una muestra aleatoria de 120 amas de casa de la ciudad, 42 prefieren usar el lavatrastos *Axión*; no obstante, en una muestra de 100 amas de casa de una tienda de Comercial Mexicana, 40 prefieren ese producto. Prueba la hipótesis de que es mayor la proporción de amas de casa de la Comercial Mexicana que prefieren jabón *Victoria*. Usa un nivel de significación de 0.02.

25. Alarmas ADT pensaba que su nuevo sistema para casas capturaría 48% del mercado en el Estado de México en 2 años, debido al bajo costo. En la región hay miles de usuarios que tienen alarmas de ADT. Luego de muestrear a 200 de ellos dos años más tarde, la compañía descubrió que 43% de ellos estaban empleando los nuevos sistemas. Con $\alpha = 0.01$, ¿debemos llegar a la conclusión de que la compañía no logró su meta de participación en el mercado?

26. El auditor de la empresa distribuidora de materiales de construcción El Surtidor informa al dueño de la misma que 40% de sus clientes pagan las facturas dentro de los 30 días de entrega; sin embargo, el dueño cree que el porcentaje es significativamente mayor. Una muestra de 200 facturas indicó que 90 fueron pagadas dentro de los 30 días de entrega. Determina quién tiene la razón, con un error α de 0.05.

27. El propietario de la tintorería Kelin desea saber las órdenes que tienen 120 días o más y que no han recogido sus clientes. Según una auditoría anterior, la proporción de este tipo de órdenes fue del 18%. Para probar si ha habido un cambio significativo en la proporción de órdenes que no recogen los clientes, se toma una muestra aleatoria de 80 órdenes, y se encontró que 16 tenían 120 días o más y no han sido recogidas por los clientes. Con un nivel $\alpha = 1\%$, indica si ha habido cambio real en la proporción de órdenes que no recogen los clientes.

28. La revista *Muy interesante* publicó que el 72% de los matrimonios han tenido relaciones prematrimoniales. ¿Qué conclusión podemos sacar sobre la nota publicada, si en una muestra de 80 matrimonios escogidos aleatoriamente, 56 manifestaron, en una encuesta anónima, que habían tenido relaciones prematrimoniales? Utiliza $\alpha = 0.05$.

29. Un distribuidor de pantalones Levi's tiene varios modelos; uno de ellos es el *501*, en tres colores diferentes: azul, blanco y negro. De un pedido de 500 pantalones, 150 fueron de color azul. ¿Concluirías que más de ¼ de todos los clientes tienen preferencia por el color azul? Usa $\alpha = 0.05$.

30. En un anuncio panorámico se publica que al menos 80% del público prefiere la pasta dental Colgate. Un estudiante de estadística toma una muestra aleatoria de 100 personas para verificar la publicidad de la marca, con $\alpha = 0.05$. ¿Qué tan pequeño debe ser el porcentaje en la muestra para poder refutar la publicidad de Colgate?

31. Una empacadora de azúcar debe producir bolsas con un contenido neto de 2 kg. Se hace una inspección para verificar si los pesos netos son correctos; para tal efecto se toma una muestra aleatoria de 40 bolsas, y se encuentra una media de 1.96 kg, con una desviación estándar de 32 g. ¿Se podría concluir que el contenido neto de azúcar está por debajo de lo que deben tener las bolsas, con un nivel de significancia del 5%?

32. Las especificaciones de cierto tornillo para el ensamblado de una computadora es de 5.5 milímetros, con una desviación típica de 0.05 mm. Una muestra

aleatoria de 40 tornillos tomada de un lote que acaba de llegar a la ensambladora mostró una media, de 5.4 mm. ¿Se puede concluir, con un nivel de significación de 5%, que el lote de tornillos cumple con las especificaciones?

33. Se espera que la vida media de cierto tipo de pila para reloj aumente al utilizar nuevos materiales para fabricarla. Se toma una muestra de 50 pilas fabricadas con los nuevos materiales, y al realizar pruebas en el laboratorio se encuentra que su duración media fue de 1 540 h, con una varianza de 10 000 horas^2. La vida media de este tipo de pilas había sido anteriormente 1 500 h. Determina si las pilas fabricadas con los nuevos materiales aumentan su duración. Usa un nivel de significancia del 5%.

34. Se desea determinar si hay diferencia significativa entre la proporción de hombres y mujeres que favorecen la sindicalización del personal de una cadena de supermercados. En una muestra aleatoria de 40 hombres y 40 mujeres, 12 y 10 respectivamente, manifestaron estar de acuerdo con la sindicalización. Utiliza un nivel de significación de 0.05 para probar si es significativa la diferencia entre ambas proporciones.

35. Históricamente, el número de puntos que obtiene un aspirante para ingresar a la UNAM es de 68 en promedio, con una desviación estándar de 13. En el último examen aplicado para ingresar a esta institución se tomó una muestra de 100 y se obtuvo una media de 65 puntos. Determina si ha disminuido significativamente el puntaje obtenido por los aspirantes en el concurso de selección. Usa $\alpha = 0.05$.

36. Los editores de un libro de *Estadística* aseguran que los alumnos entienden con mayor rapidez el conocimiento de la asignatura que los alumnos que utilizan otros libros. Se escogen al azar 100 estudiantes, y se dividen en dos secciones: *A* y *B*, se les imparte la asignatura por dos profesores de igual capacidad. El profesor de la sección *A* utiliza el texto bajo estudio y el profesor de la sección *B* utiliza cualquier otro texto. Al finalizar el curso, se observaron los siguientes resultados: los alumnos que utilizaron el texto bajo estudio obtuvieron una calificación media de 7.8 con una desviación estándar de 1.0 punto, y los alumnos que utilizaron otro texto obtuvieron una media de 7.1 puntos con una desviación estándar de 1.9 puntos. ¿Tienen razón los editores? Usa $\alpha = 0.05$.

37. Una compañía constructora está preocupada por el tiempo que se pierde debido a accidentes de trabajo. Por ello, dispuso montar un programa de seguridad para reducir el tiempo perdido debido a dichos accidentes. El programa duró 30 meses y, al finalizar, el tiempo perdido por accidentes tuvo un promedio de 96 h mensuales, con una desviación típica de 15 h. En los 36 meses anteriores al programa de seguridad, el tiempo perdido por accidentes promedió 110 h mensuales con una desviación estándar de 18 h. Determina si fue efectivo el programa de seguridad para disminuir el tiempo perdido por accidentes de trabajo. Usa un nivel de significancia del 5%.

38. Se aplicó un examen de ortografía a una muestra aleatoria de 25 secretarias que laboran en la FES-C de campo 1, obteniendo un puntaje medio de 78 puntos, y una desviación estándar de 7 puntos. Se aplicó el mismo examen a una muestra aleatoria de 25 secretarias de campo 4 y se obtuvo un promedio de 75 puntos con una desviación estándar de 9 puntos. ¿Consideras que las secretarias de campo 1 obtuvieron mejores calificaciones? Usa $\alpha = 0.01$.

39. En relación con el problema anterior, redacta una pregunta de tal forma que se tenga que hacer una prueba bilateral.

40. En una muestra de 80 focos de Holophane se obtuvo un 5% de productos defectuosos; en otra muestra de 100 focos de Osram, se obtuvo un 7% de focos defectuosos. Comprueba si existe una diferencia significativa en cuanto a la cantidad de productos defectuosos de ambas fábricas, con un nivel α del 1%.

41. A partir de muestras aleatorias de 200 transistores, fabricados por la máquina *A*, y de 150 transistores, fabricados por la máquina *B*, se obtuvieron 19 y 15 transistores defectuosos respectivamente. Ensaya las hipótesis siguientes:

a) las dos máquinas tienen diferente calidad de fabricación. Usa $\alpha = 0.05$.

b) la máquina *B* es mejor que la *A*. Usa $\alpha = 0.025$.

42. Se aplicó un examen de aptitud matemática a un grupo de 40 estudiantes, y 18 de ellos fueron clasificados como aptos para las matemáticas; mientras que en otro grupo de 80 estudiantes a quienes se les aplicó el mismo examen, 32 fueron clasificados como aptos. Determina si es significativa la diferencia entre ambos grupos respecto a la aptitud para las matemáticas. Usa un error alfa del 5%.

43. Se hizo un estudio de mercados a Nescafé. Se desea saber si con el cambio de envase del producto podrá aumentar el volumen de ventas. Una muestra aleatoria de 50 tiendas donde se vende el producto con el nuevo envase reveló que 62% de los clientes compraron el café en esa nueva presentación, mientras que una muestra de 60 tiendas que venden el café con el envase antiguo mostró que sólo 55% lo adquirieron. Con un nivel de significancia de 5% determina si con la nueva presentación del producto aumentó el porcentaje de consumidores.

44. Bimbo desea probar las ventas de su nuevo *panqué sabor a naranja*; se eligieron 14 tiendas; el promedio de ventas por mes fue 2 940 pesos, con una desviación estándar de 147 pesos. La gerencia decidió que el volumen promedio fuese de 3 125 pesos al mes, determina si es necesario establecer un plan nacional de distribución del producto. Usa un error alfa de 0.05.

45. La escuela de manejo Continental en uno de los cursos para conductores se encontró que el tiempo medio que necesitaron 18 mujeres para aprender a conducir fue de 25 h con una desviación típica de 3 h; mientras que el tiempo promedio para aprender a manejar de una muestra de 15 hombres fue 28 h con una desviación estándar de 4 h. Comprueba la hipótesis de que las mujeres aprenden más rápidamente que los hombres a manejar, con un error α 0.01.

46. Omnibús de México considera comprar una de dos marcas de baterías. Antes de tomar la decisión, hace pruebas aceleradas con muestras de las baterías *LTH y América*. El registro de las duraciones de ambas marcas proporcionó la siguiente información:

LTH	América
$n = 10$ baterías promedio de 320 horas Desviación estándar 50 horas	$n = 12$ baterías promedio de 260 horas Desviación estándar 80 horas

De acuerdo con lo anterior, ¿es más conveniente comprar la marca LTH?

47. El gerente de personal de una tienda Walmart de México desea conocer los tiempos que tardan las cajeras de la tienda en cobrar a los clientes, por lo que tomó una muestra de 16 cajeras del turno matutino y 14 del vespertino. Los tiempos requeridos para atender a los clientes fue de:

Turno matutino	Turno vespertino
Promedio 5.4 minutos Desviación estándar 0.6 minutos	Promedio 5.8 minutos Desviación estándar 0.5 minutos

a) prueba la hipótesis de que no existe diferencia significativa entre los tiempos requeridos para cobrar a los clientes entre ambos turnos. Usa un error α de 0.5.

48. Se comprobó un micrómetro con una serie de piezas patrón y se obtuvo lo siguiente:

Patrón	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Lectura	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1

a) con un nivel de significación de 0.01, ¿tiene error el instrumento?

49. A fin de medir el efecto de una campaña de ventas en toda la tienda para los artículos que no se ponen en barata, el director de investigación de una cadena de supermercados tomó una muestra aleatoria de 13 pares de tiendas, según su volumen de ventas semanales. Una tienda de cada par (el grupo experimental) se expuso a la campaña de ventas y la otra no (grupo control). Los siguientes datos indican los resultados para un periodo de una semana.

a) para un nivel de significancia de 0.05, ¿puede el director de investigación concluir que existen pruebas de que la campaña de ventas aumentó las ventas promedio de los artículos que no se ponen en barata?

b) ¿qué suposición es necesaria para realizar esta prueba?

Con campaña de ventas	67	59	80	48	98	38	57	75	95	61	32	49	54
Sin campaña de ventas	65	55	81	40	93	38	52	70	89	58	33	42	54

50. Un profesor de una escuela de Administración desea investigar los precios de los nuevos libros de texto en la librería del campus y de su competidor fuera de él, que es una sucursal de una cadena nacional. El profesor elige al azar los textos requeridos para 12 cursos y compara los precios de las dos librerías. Los resultados se muestran en seguida.

a) Para un nivel de significancia de 0.01, ¿existen indicios de una diferencia del precio promedio de los libros de texto en las dos librerías?

b) ¿Qué suposición es necesaria para realizar esta prueba?

Libro	1	2	3	4	5	6	7	8	9	10	11	12
Librería Campus	55	48	51	39	59	50	40	42	42	45	46	57
Librería competidor	51	46	51	39	56	46	40	40	43	42	44	56

6 Potencia

Supongamos que trabajas en una fundación cuya misión es respaldar a investigadores en matemáticas educativas, y tu papel es evaluar las propuestas y decidir cuáles financiar. Recibes una propuesta de un nuevo método de enseñanza de la estadística en licenciatura. El proyecto de investigación contempla comparar los logros de los estudiantes a los que se les enseña con el nuevo método, con los logros obtenidos por los que cursaron la asignatura con el método tradicional. El proyecto contiene argumentos teóricos interesantes acerca del porqué el nuevo método debe ser mejor y la propuesta metodológica es firme. Adicionalmente a estos elementos positivos, hay una pregunta importante que debe ser contestada: ¿Tiene el experimento alta probabilidad de proporcionar fuerte evidencia acerca de que el nuevo método es mejor que el tradicional, cuando en realidad el nuevo método es mejor? Es posible, por ejemplo, que el tamaño de muestra propuesto sea pequeño y aun cuando exista una diferencia grande en la población no sea fácil detectarla. Esto es, si el tamaño de la muestra es pequeño, entonces una diferencia importante entre las medias de las muestras resulta ser no significante. Si esta diferencia resulta ser no significante, entonces no se puede sacar ninguna conclusión acerca de la diferencia entre las medias de la población. No está justificado aceptar que la hipótesis nula es verdadera (las medias de las poblaciones son iguales) sólo porque la diferencia no fue significativa. Por supuesto, no se justifica concluir que la hipótesis nula es falsa. Por tanto, cuando un efecto es no significante, no se puede tener una conclusión final. Podrías preferir que el dinero de la fundación se use para financiar un proyecto que tenga una probabilidad más alta de ser capaz, con base en sus resultados, de proporcionar una conclusión sólida.

La potencia se define como la probabilidad de rechazar correctamente una hipótesis nula falsa. En relación con nuestro ejemplo, es la probabilidad que dado que hay una diferencia entre las medias de la población del nuevo método y del método tradicional, las medias de las muestras sean significativamente diferentes. La probabilidad de no rechazar una hipótesis nula falsa se representa como β. Por tanto, la potencia se define como:

$$Potencia = 1 - \beta$$

Es muy importante considerar la potencia cuando se diseña un experimento. Debes evitar perder tiempo o dinero en un experimento que tenga muy poca probabilidad de detectar un efecto significante.

PREGUNTAS

1. La potencia es:
   a) la probabilidad de que la hipótesis nula sea verdadera.
   b) la probabilidad de que la hipótesis nula sea falsa.
   c) la probabilidad de que la hipótesis nula si es falsa, sea rechazada.
   d) la probabilidad de que la hipótesis nula si es verdadera, sea rechazada.

2. Si la potencia de un experimento es pequeña, entonces:
   a) el experimento probablemente no puede proporcionar conclusiones sólidas.
   b) cualquier resultado significante obtenido es sospechoso.
   c) los resultados tienen sesgo.

Cálculos

En el ejemplo de las sopas preparadas con caldo de pollo natural o con concentrados, el problema consistía en determinar si el profesor Garibay podía distinguir entre una sopa preparada con caldo o con concentrado. Para este ejemplo, suponemos que él puede distinguir la diferencia el 0.75 de las veces (suponemos que ésta es la realidad, aunque no pierdas de vista que es desconocida). Ahora, consideremos un experimento que se llevó a cabo para determinar si Garibay puede distinguir entre las sopas, específicamente si puede acertar más de 0.50 de las veces. Sabemos que sí puede (es la suposición de la realidad en este ejemplo). Sin embargo, el experimentador no la conoce y le pide al profesor Garibay probar 16 tazas de sopa. El experimentador, realizará una prueba de significancia utilizando la distribución binomial. Específicamente, si la prueba de una cola es significante al nivel de 0.05, el experimentador concluirá que el profesor Garibay puede distinguir la diferencia. El valor de probabilidad se calcula suponiendo que la hipótesis nula es verdadera ($\pi = 0.50$). Por tanto, el experimentador registra cuántas veces acierta el profesor Garibay y calcula la probabilidad de obtener este número de aciertos o más, asumiendo que la hipótesis nula es cierta. La pregunta es: ¿cuál es la probabilidad de que el experimentador rechace la hipótesis nula $\pi = 0.50$, en forma correcta (es decir cuando es falsa)? En otras palabras ¿Cuál es la potencia de la prueba?

En la figura 6.1 se muestra la distribución binomial para $n = 16$ y $\pi = 0.50$. La probabilidad de obtener 11 aciertos o más es 0.105 y la probabilidad de obtener 12 aciertos o más es 0.0308. Por tanto, la probabilidad de obtener 12 aciertos o más

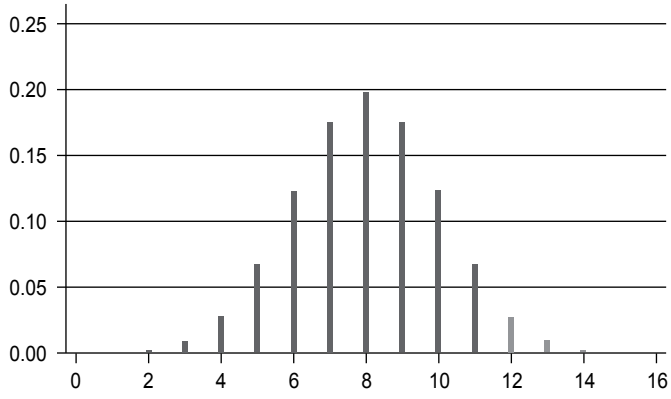


Figura 6.1 La distribución binomial para $n = 16$ y $\pi = 0.5$.

es menor que 0.05. Esto significa que la hipótesis nula sólo puede rechazarse si el profesor Garibay acierta 12 veces o más, y no pude rechazarse en caso contrario.

Sabemos que la realidad es que el profesor Garibay acierta 0.75 de las veces. (Obviamente, el experimentador no conoce esto; de saberlo no habría necesidad de realizar ningún experimento). En la figura 6.2 se muestra la distribución binomial para $n = 16$ y $\pi = 0.75$.

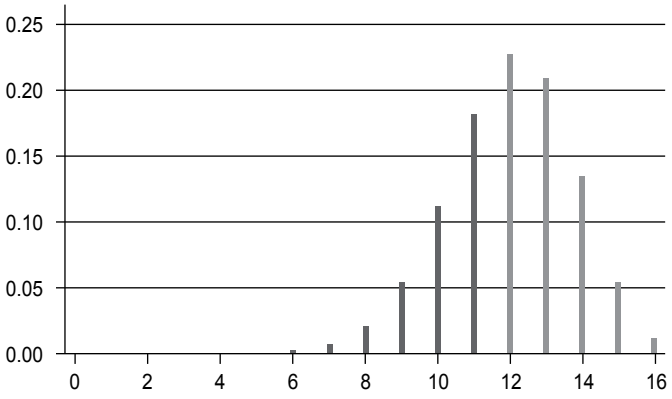


Figura 6.2 La distribución binomial para $n = 16$ y $\pi = 0.75$.

La probabilidad de acertar correctamente 12 o más de las veces es 0.63. Por tanto, la potencia del experimento es 0.63.

En resumen, la probabilidad de acertar 12 o más veces, dado que la hipótesis nula es verdadera, es menor de 0.05. Por tanto, si el profesor Garibay acierta 12 veces o más, la hipótesis nula será rechazada. Dado que el profesor Garibay tiene en realidad habilidad para distinguir entre las bebidas el 0.75 de las veces, la probabilidad de que acierte 12 veces o más es 0.63. Por tanto, la potencia es 0.63.

En la sección "Prueba de una media", el primer ejemplo se realiza bajo la suposición de que el experimentador conocía el valor de la varianza poblacional. Aunque esto no es usual en la práctica, el ejemplo es útil para propósitos didácticos. Por esta razón, en el ejemplo siguiente suponemos que el investigador conoce la varianza poblacional.

Supongamos que en un examen de matemáticas el grupo alcanzó una media de 75 con una desviación estándar de 10. Un investigador está interesado en probar si un nuevo método de enseñanza incrementa el promedio de calificaciones. Supongamos que el investigador no conoce que la media de calificaciones para el nuevo método es 80. El plan de investigación contempla aplicar el nuevo método, seleccionar al azar a 25 estudiantes, aplicar el examen y realizar una prueba de significancia de una cola para probar si la media de la muestra es significativamente mayor a 75. ¿Cuál es la probabilidad de que el investigador rechace la hipótesis nula falsa de que la media de la población es 75? En seguida se muestra cómo calcular esta probabilidad.

El investigador supone que la desviación estándar de la población con el nuevo método es la misma que con el tradicional (10) y que la distribución es normal. Ya que se supone que la desviación estándar de la población es conocida, el investigador puede usar la distribución normal para calcular el valor de p. Recuerda que el error estándar de la media ($\sigma_{\bar{x}}$) es

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Para nuestro ejemplo, es igual a $10/5 = 2$. Como puedes ver en la figura 5.3, si la hipótesis nula es verdadera, es decir, la media de la población es 75, entonces la probabilidad de obtener una media muestral igual o mayor a 78.29 es 0.05. Por tanto, el investigador rechazará la hipótesis, si la media de la muestra es igual o mayor a 78.29.

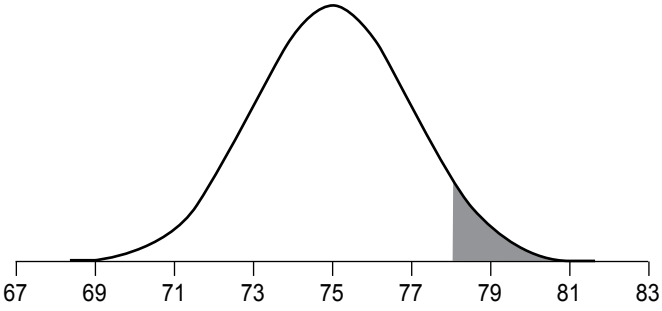


Figura 6.3 La distribución muestral de la media si la hipótesis nula es verdadera.

Nos podemos preguntar: ¿cuál es la probabilidad de que el experimentador obtenga una media muestral igual o mayor a 78.29, dado que la media poblacional es 80? En la figura 6.4 se muestra que esta probabilidad es 0.80.

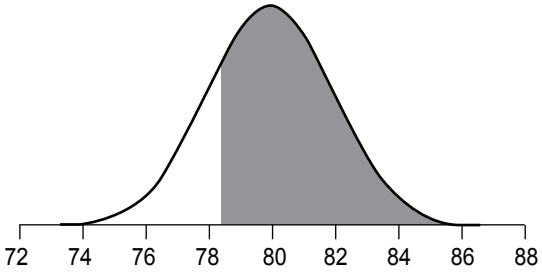


Figura 6.4 La distribución muestral de la media si la media de la población es 75. La prueba es significante si la media muestral es igual o mayor a 78.29.

Por tanto, la probabilidad de que el experimentador rechace la hipótesis nula ($\mu = 75$), cuando ésta es falsa ($\mu = 80$) es 0.80. En otras palabras, la potencia es igual a 0.80.

El cálculo de la potencia es más complejo para la prueba *t*. Utilizando *software* estadístico, se puede determinar la potencia para otros tipos de diseño.

PREGUNTAS

1. Una moneda se lanza 26 veces. Se usa la distribución binomial para realizar una prueba de una cola con la región de rechazo en la cola superior. Lo que se espera es que en más de la mitad de los lanzamientos caigan águilas. Mediante prueba y error, usando algún software que contenga la distribución binomial, encuentra el número de águilas para la cual la probabilidad de obtener este número o más sea menor de 0.05.

a) Respuesta: _____.

2. Una moneda se lanza 26 veces. Se usa la distribución binomial para realizar una prueba de una cola con la región de rechazo en la cola superior. ¿Cuál es la probabilidad de que la hipótesis nula sea rechazada, si la probabilidad de obtener un águila en un lanzamiento es 0.75?

a) Respuesta: _____.

Factores que afectan la potencia

Algunos factores afectan la potencia de una prueba estadística. Algunos de éstos los puede controlar el investigador, otros no. El siguiente ejemplo ilustra varios de estos factores.

Supongamos que en un examen de matemáticas las calificaciones se distribuyen en forma normal con una media de 75 y una desviación estándar σ. Un investigador está interesado en probar que un nuevo método de enseñanza aumenta el promedio del grupo. Aunque no conocemos el valor de la media de la población μ, para las calificaciones con el nuevo método, suponemos que en realidad sí es mayor de 75. El investigador planea seleccionar una muestra de n sujetos y realizar una prueba de significancia de una cola para determinar si la media de la muestra es significativamente mayor de 75. ¿Cuál es la probabilidad de que el investigador en forma correcta rechace la hipótesis nula falsa acerca de que el promedio es 75, con un nivel de 0.05?

Tamaño de muestra

En la figura 6.5 se muestra que a mayor tamaño de la muestra la potencia es mayor. Ya que el tamaño de la muestra normalmente es un factor controlado por el experimentador, incrementarlo es una forma de aumentar la potencia. Sin embargo, trabajar con muestras grandes puede ser dificultoso o muy costoso.

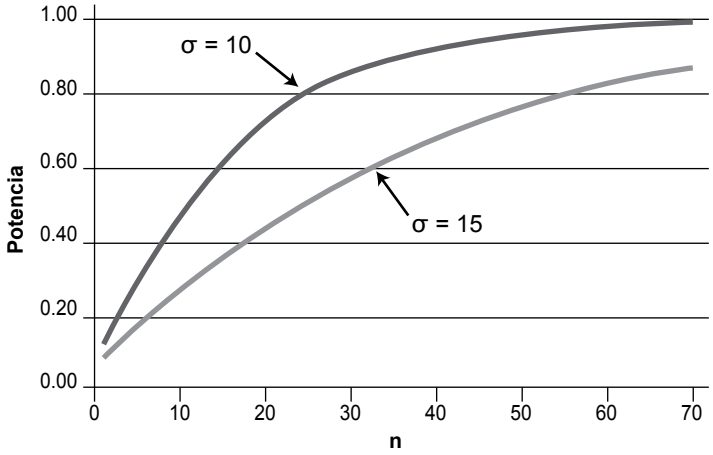


Figura 6.5 La relación entre el tamaño de la muestra y la potencia para H_0: $\mu = 75$, cuando $\mu = 80$, prueba de una cola $\alpha = 0.05$, para valores de σ de 10 y 15.

Desviación estándar

En la figura 6.5 también se muestra que la potencia es grande cuando la desviación estándar es pequeña. Los experimentadores pueden controlar a veces la desviación estándar muestreando poblaciones homogéneas o reduciendo la variabilidad en el proceso de medición.

Diferencias entre la media hipótetica y la verdadera

Naturalmente, cuanto mayor sea el efecto, es más probable que en un experimento se detecte un efecto significativo. En la figura 6.6 se muestra el efecto de aumentar la diferencia entre la media especificada por la hipótesis nula, 75 y la media de la población μ para desviaciones estándar de 10 y 15.

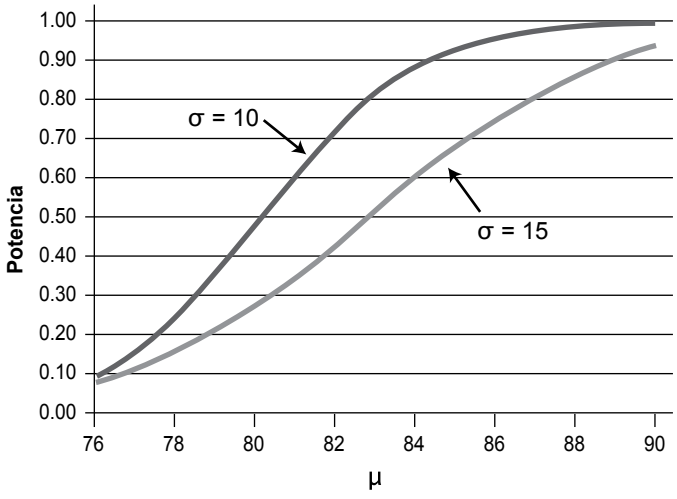


Figura 6.6 La relación entre el tamaño de la muestra y la μ, con H_0: $\mu = 75$, prueba de una cola $\alpha = 0.05$, para valores de σ de 10 y 15.

Nivel de significancia

Cuanto más riguroso sea el nivel de significancia (más bajo), la potencia será más baja. En la figura 6.7 se muestra que la potencia es más baja para el nivel de 0.01, que para el nivel de 0.05.

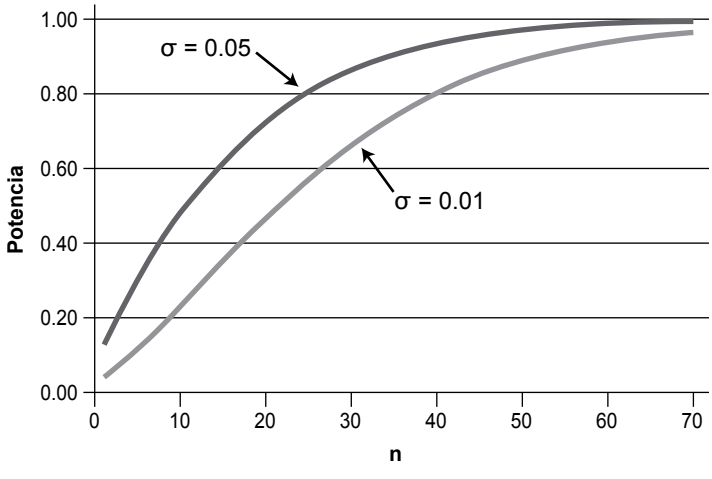


Figura 6.7 La relación entre el tamaño de la muestra y el nivel de significancia para pruebas de una sola cola: H_0: $\mu = 75$, μ verdadera $= 80$ y $\sigma = 10$.

Prueba de una cola contra la prueba de dos colas

La potencia es mayor para la prueba de una cola siempre que esté en la dirección correcta. Una prueba de una cola a un nivel de 0.05 tiene el mismo valor crítico que una de dos colas al nivel de 0.025. Como se mostró, a menor nivel de significancia menor potencia. Una prueba de una cola aumenta el nivel de significancia en comparación con la de dos colas.

PREGUNTAS

1. La potencia es la probabilidad de aceptar la hipótesis nula, dado que la hipótesis nula es verdadera:
   a) verdadero.
   b) falso.

2. Seleccionar lo que incrementa la potencia:
   a) incrementar la desviación estándar.
   b) incrementar el tamaño de la muestra.
   c) incrementar el nivel de significancia.
   d) incrementar el tamaño de la diferencia entre las medias.

3. Seleccionar lo que disminuye la probabilidad de cometer el error Tipo I.
   a) incrementar la desviación estándar.
   b) incrementar el tamaño de la muestra.
   c) disminuir el nivel de significancia.

Actividades

I. Contesta y resuelve los siguientes ejercicios para reafirmar los conceptos.

1. Define la potencia de la prueba con tus propias palabras.

2. Escribe tres formas en las que se puede incrementar la potencia de la prueba de un experimento. Explica tu respuesta.

3. Media poblacional 1 = 36.

Media poblacional 2 = 45

Varianzas poblacionales = 10

En una prueba *t*, qué valor de *p* determinará si una diferencia entre las medias es significativa con un nivel del 0.05? Obtén los resultados para una prueba de una cola y para una prueba de dos colas. Para la prueba de una cola utiliza un nivel del 0.05 y para la prueba de dos colas utiliza un nivel de 0.01.

4. Ordena las letras (*a* – *e*) en términos de su potencia.

	Media poblacional 1	***n***	**Media poblacional 2**	**Varianza**
a	29	20	43	12
b	34	150	40	6
c	105	24	50	27
d	314	4	120	10
e	30	31	41	8

5. Juan, buscando en el sótano de su abuela, tropezó con un objeto brillante que sobresalía de un montón de cajas. Cuando alcanzó el objeto apareció un genio que milagrosamente se materializó y le dijo: "has encontrado mi moneda mágica. Si tiras esta moneda un número infinito de veces, notarás que las águilas caerán el 60% de los volados". Después de lo dicho, el genio desapareció para nunca más ser visto. Juan, emocionado por su nuevo descubrimiento mágico, se acercó a su amigo José y le platicó lo sucedido. José se mostró escéptico acerca de la historia de su amigo; sin embargo, le dijo a Juan que tirara 100 veces la moneda y anotara el número de veces que cayera águila.

a) ¿cuál es la hipótesis nula de José?

b) ¿cómo se le llama a la probabilidad que Juan debe obtener, para convencer a José de que su moneda tiene poderes especiales, y cuyo valor de p es menor a 0.05 (prueba de una cola)?

c) si José dijera a Juan que tire la moneda sólo 20 veces, ¿cuál es la probabilidad con la cual Juan no será capaz de convencer a José?

7 Correlación y regresión

En este capítulo se discute la relación entre dos variables. Por ejemplo, quizá quieras describir la relación entre el peso y la estatura de un grupo de personas para determinar el proceso de producción en una fábrica de ropa. En la sección "Introducción a los datos bivariados" se presentan ejemplos de la relación entre dos variables y la mejor manera de representar gráficamente los datos bivariados. En las siguientes secciones se discute el índice más utilizado para medir la relación entre dos variables, conocido como el coeficiente de correlación de Pearson.

A menudo se recurre a la estadística para predecir el comportamiento de una variable, usando como predictores otras variables. Por ejemplo, se puede tratar de predecir el promedio de calificaciones de un estudiante en la universidad utilizando como predictor el promedio que haya obtenido en preparatoria. O se podría querer predecir el ingreso de una persona usando el número de años de formación académica. En las secciones dedicadas a la regresión lineal, se revisan los métodos estadísticos usados para predecir el comportamiento de una variable, usando como predictor otra variable.

Introducción a los datos bivariados

Las medidas de tendencia central, variabilidad y de forma de una distribución resumen a una variable, porque proporcionan información importante acerca de su distribución. Por lo general, más de una variable se mide en cada elemento. Por ejemplo, en la mayoría de los estudios acerca de la salud de una población, es común obtener medidas de diferentes variables, como edad, peso, sexo, estatura, presión sanguínea, colesterol total, etc., para cada individuo. Los estudios económicos pueden estar interesados en variables como personal de ingreso y años de estudio. En este capítulo consideraremos datos bivariados, es decir, dos variables cuantitativas medidas en cada elemento. Nuestro primer interés es poder resumir estos datos en forma similar a como se hace en el caso de datos univariados (una sola variable).

Consideremos una situación que nos es familiar: la edad. Empecemos por preguntarnos si las personas tienden a casarse con otra persona de la misma edad. Por experiencia, podemos contestar afirmativamente, pero ahora la pregunta sería: ¿qué tan buena es esta correspondencia? Para poder contestar a esto, podemos ver las edades para una muestra de matrimonios. En la tabla 7.1 se muestran las edades para 10 matrimonios. Se puede observar que hombres y mujeres tienden a ser de la misma edad, con una tendencia de los hombres a ser ligeramente mayores que las mujeres. Esto no es una gran sorpresa, pero al menos los datos soportan nuestra experiencia, lo que no siempre es el caso.

Esposos	36	72	37	36	51	50	47	50	37	41
Esposas	35	67	33	35	50	46	47	42	36	41

Tabla 7.1 Edades de una muestra de 10 matrimonios.

Las edades de la tabla 7.1 se seleccionaron de una base de datos de matrimonios. Sabemos que cada variable puede resumirse mediante un histograma (ver figura 7.1), su media y su desviación estándar (ver tabla 7.2).

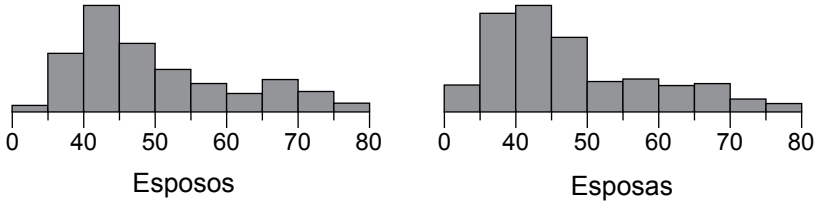


Figura 7.1 Histograma de las edades de los matrimonios.

	Media	Desviación estándar
Esposos	49	11
Esposas	47	11

Tabla 7.2 Media y desviación estándar de las edades de los matrimonios.

Las distribuciones son sesgadas con una cola larga a la derecha. En la tabla 7.1 vemos que no todos los maridos son de mayor edad que sus esposas, y es importante ver que este hecho no tiene interés cuando estudiamos por separado las variables. Es decir, aunque proporcionamos un resumen estadístico para cada variable, la información de cada matrimonio se pierde al separar las variables. No podemos decir, por ejemplo, con base solamente en las medias, qué porcentaje de matrimonios tiene maridos más jóvenes que sus esposas. Tenemos que revisar los pares para determinar este porcentaje. Solamente manteniendo la información en pares se puede responder a preguntas de este tipo. Otro ejemplo de la información no disponible en la estadística descriptiva de las edades de los maridos y de las esposas por separado sería, ¿cuál es la edad promedio de los maridos casados con mujeres de 45 años? En conclusión, no sabemos la relación entre la edad del marido y la edad de la esposa.

Podemos saber mucho de esta relación si graficamos los datos bivariados. En la figura 7.2 se muestra el diagrama de dispersión para los pares de edades. El eje *X* representa la edad del esposo, mientras que el eje *Y*, la edad de la esposa.

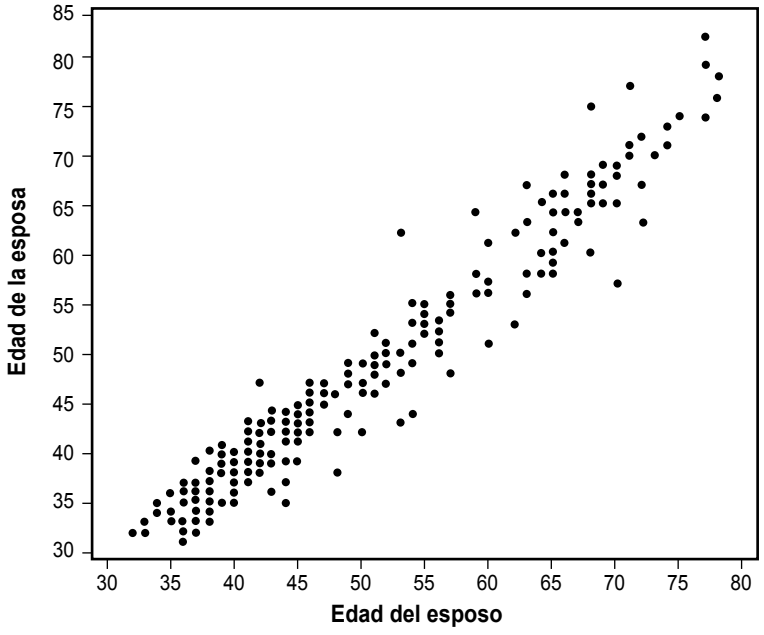


Figura 7.2 Diagrama de dispersión que muestra la edad de la esposa en función de la edad del esposo.

En la figura 7.2 se revelan dos características importantes de los datos. Primero, es claro que hay una fuerte relación entre la edad del hombre y de la mujer. Cuanto mayor sea el esposo, mayor será la esposa. Cuando una variable (Y) aumenta, cuando se incrementa la variable (X), decimos que *X* y *Y* tienen una asociación positiva. Cuando *Y* disminuye, al aumentar *X*, decimos que tienen una asociación negativa. Segundo, los puntos se agrupan a lo largo de una línea recta. Cuando esto ocurre, decimos que tenemos una relación lineal.

En la figura 7.3 se muestra el diagrama de dispersión de las ventas en miles de litros de una muestra de 149 gasolineras. La correlación se hizo entre los años de servicio y la cantidad de litros de gasolina Magna vendidos. Hay una asociación positiva entre estas dos variables. No es ninguna sorpresa que cuantos más años esté en servicio una gasolinera, mayores sean sus ventas. Aunque los puntos se agrupan a lo largo de una línea recta, no están tan cerca de ella como en el caso de las edades de los matrimonios.

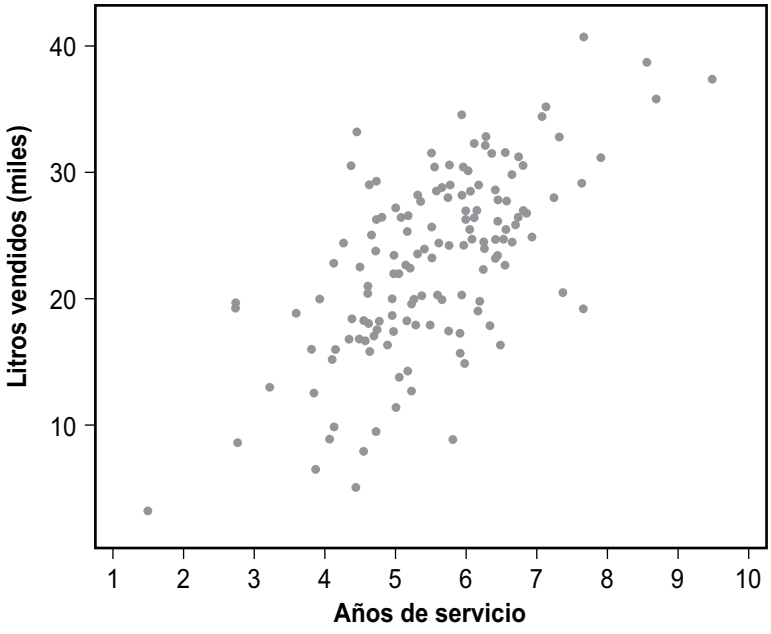


Figura 7.3 Gráfica de dispersión de los años de servicio de las gasolineras y los litros vendidos.

No todos los diagramas de dispersión muestran relaciones lineales. En la figura 7.4 se muestran los rendimientos por hectárea de un producto agrícola obtenidos en una misma superficie en los últimos 8 años. Es claro que la relación entre "Rendimiento" y "Año" no describe una línea recta: si dibujas una línea recta que una el primer punto y el último, todos los demás puntos quedan por arriba de esta recta. Los datos se ajustan mejor a una parábola.

Los diagramas de dispersión que muestran relaciones lineales entre las variables pueden diferir en la pendiente de la línea, en el intercepto con el eje *Y* y en la forma en que se agrupan los datos alrededor de la línea. Una medida estadística de la fuerza de la relación entre las variables, que toma en cuenta estas diferencias, es el tema de la sección siguiente.

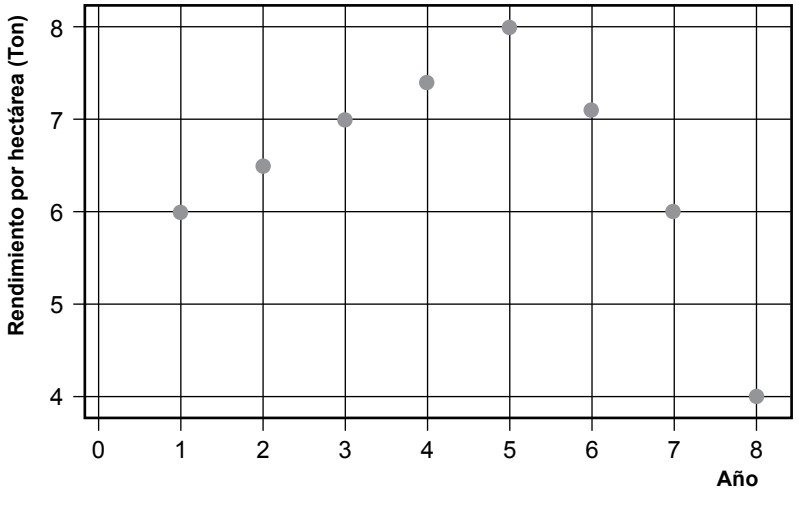


Figura 7.4 Rendimientos por hectárea de un producto agrícola obtenidos en los últimos 8 años.

PREGUNTAS

1. Una buena forma de representar la relación entre dos variables es con:

a) una tabla de medias.

b) una gráfica de caja.

c) una gráfica de tallo y hojas.

d) un diagrama de dispersión.

2. Cuando los puntos tienden a agruparse alrededor y a lo largo de una línea, la relación se conoce como:

a) relación lineal.

b) relación lineal recta.

3. Galileo encontró que la relación entre la altura que se desciende en un plano inclinado y la distancia recorrida es lineal.

a) falso.

b) verdadero.

Coeficiente de correlación de Pearson

El coeficiente de correlación de Pearson es una medida de qué tan fuerte es la relación lineal entre dos variables. Si la relación entre las variables no es lineal, entonces el coeficiente de correlación no representa en forma adecuada la relación entre las variables.

El símbolo para el coeficiente de correlación es "ρ", cuando se refiere a la población y "r" cuando se calcula con datos muestrales. Debido a que, por lo general, trabajaremos con muestras, usaremos r para referirnos al coeficiente de correlación; en caso de usar el poblacional, lo aclararemos específicamente.

El rango del coeficiente de correlación de Pearson es de −1 a 1. $r = -1$ indica una relación lineal perfecta negativa entre las variables; $r = 0$ indica que no existe relación entre las variables, y $r = 1$ indica una relación lineal perfecta positiva entre las variables. En la figura 7.5 se muestra un diagrama de dispersión para $r = 1$.

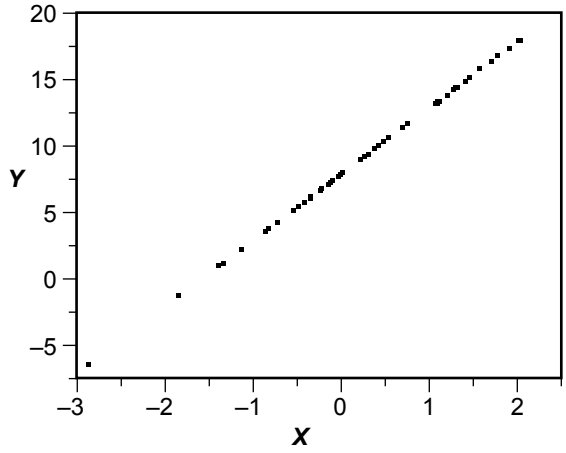


Figura 7.5 Una relación lineal perfecta, $r = 1$.

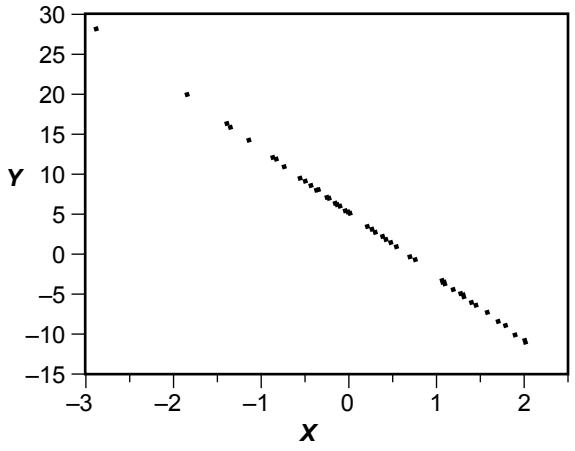


Figura 7.6 Una relación lineal negativa perfecta, $r = -1$.

En la figura 7.6 se muestra un diagrama de dispersión para $r = -1$.

Observa que cuando X aumenta, Y disminuye.

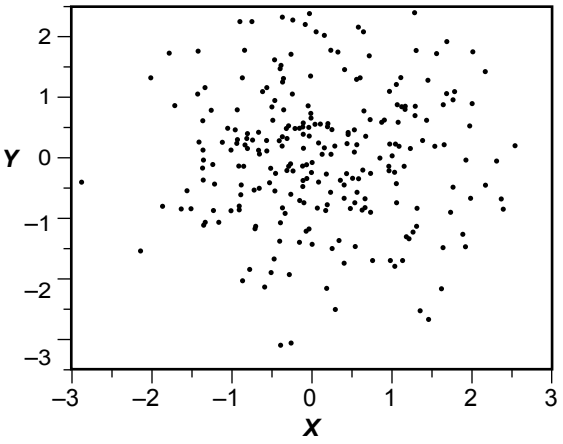


Figura 7.7 No existe relación entre las variables, $r = 0$.

En la figura 7.7 se muestra un diagrama de dispersión para $r = 0$. Observa que no existe relación entre X y Y.

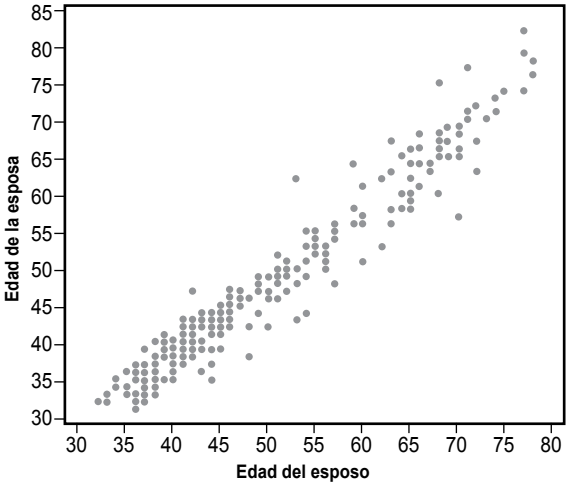


Figura 7.8 Diagrama de dispersión para las edades de los matrimonios, $r = 0.97$.

Con datos reales, no debes esperar valores de r exactamente iguales a –1, 0 o +1. Los datos de las edades de los matrimonios que se describieron en la sección "Introducción a los datos bivariados" tienen un $r = 0.97$, y se muestran en la figura 7.8.

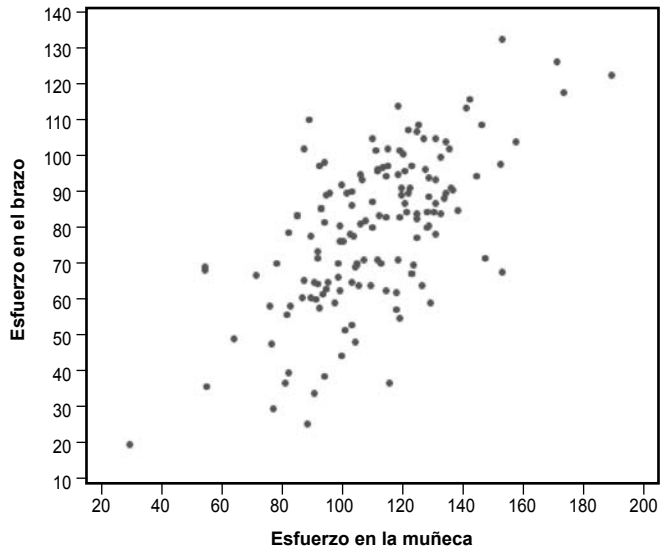
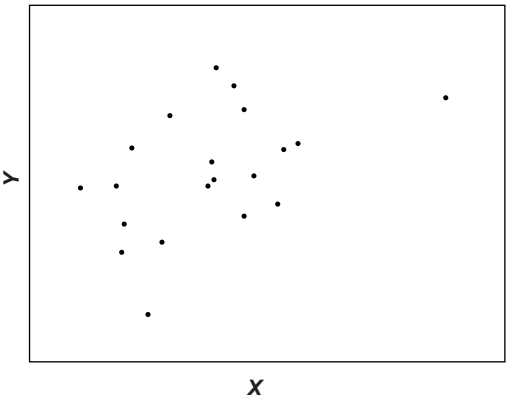


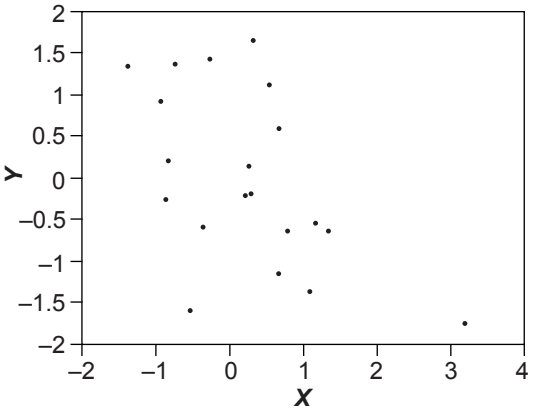
Figura 7.9 Diagrama de dispersión para el esfuerzo en muñeca y brazo, $r = 0.63$.

En la figura 7.9 se muestran los datos del esfuerzo en el puño y en el brazo, descritos en la sección "Introducción a los datos bivariados", con un $r = 0.63$.

PREGUNTAS



1. El diagrama de dispersión de la izquierda representa:
- *a*) una relación positiva.
- *b*) una relación negativa.
- *c*) no tiene relación.



2. El diagrama de dispersión de la izquierda representa:
- *a*) una relación positiva.
- *b*) una relación negativa.
- *c*) no tiene relación.

Propiedades del coeficiente de correlación

Una propiedad básica del coeficiente de correlación *r* es que sus valores posibles se establecen en el rango comprendido entre −1 a 1. Una correlación de −1 significa una relación lineal perfecta negativa; una correlación de 0 significa que no existe relación lineal, y una correlación de 1 significa una relación lineal perfecta positiva.

El coeficiente de correlación de Pearson es simétrico en el sentido de que la correlación de *X* con *Y* es la misma que la correlación entre *Y* con *X*. Por ejemplo, la correlación de peso con estatura es la misma que la correlación entre estatura y peso.

Una propiedad crítica del coeficiente de correlación *r* es que no se ve afectado por transformaciones lineales. Esto significa que multiplicar una variable por una constante y sumarle una constante no cambia la correlación de esta variable con otras. Por ejemplo, la correlación entre peso y estatura no depende de que ésta se mida en pulgadas, pies, metros, etcétera. En forma similar, por ejemplo, sumar 5 puntos a cada estudiante en su examen, no cambia la correlación del resultado de ese examen con el promedio obtenido en la preparatoria GPA.

PREGUNTAS

1. La correlación entre la temperatura y el número de barquillos vendidos es la misma sin importar si la temperatura se midió en °C o °F.

a) falso.

b) verdadero.

2. La correlación entre dos grupos de números es la misma que la correlación entre los logaritmos de los números de los dos grupos.

a) falso.

b) verdadero.

3. ¿Cuál de los siguientes números no es un valor posible del coeficiente de correlación?

a) −1.5.

b) −1.0.

c) 0.

d) 0.99.

4. ¿Cuál es mayor, la correlación entre estatura y peso, o la correlación entre peso y estatura?

a) Correlación entre peso y estatura.

b) Aproximadamente iguales.

c) Exactamente iguales.

d) Correlación entre estatura y peso.

Cálculo del coeficiente de correlación *r*

Hay algunas fórmulas que se usan para calcular el coeficiente de correlación. Algunas tienen un sentido más conceptual que otras, pero son fáciles de utilizar con fines de cálculo. Iniciemos con la fórmula que tiene mayor sentido conceptual.

Vamos a calcular la correlación entre las variables *X* y *Y*, que se muestran en la tabla 7.3. Empecemos por calcular la media de las *X* y restar este valor a todos los valores de *X*. A la nueva variable le llamaremos "*x*". De manera similar, se calcula la variable "*y*". Las variables *x* y *y* son las desviaciones respecto a sus medias

correspondientes. Observa que las medias de *x* y *y* son 0. Ahora creamos una nueva columna multiplicando *x* por *y*.

Antes de seguir con los cálculos, veamos por qué la suma de la columna *xy* revela la relación entre *X* y *Y*. Si no existiera relación entre *X* y *Y*, entonces los valores positivos de *X* tendrían la misma probabilidad de aparecer con un valor positivo o con un valor negativo de *Y*. Esto hace que se tenga la misma probabilidad de obtener valores positivos y negativos de *xy* y la suma debe ser pequeña. Por otro lado, considera la tabla 7.3, en la cual valores altos de *X* se asocian a valores altos de *Y,* y los valores bajos de *X* se asocian a valores bajos de *Y*. Puedes ver que los valores positivos de *x* están asociados a valores positivos de *y,* y los valores negativos de *x* están asociados a valores negativos de *y;* por tanto, en todos los casos, el producto de *xy* es positivo, dando por resultado un valor grande para la columna *xy*.

Finalmente, si hubiera una relación negativa, entonces los valores positivos de *x* estarían asociados con los valores negativos de *y,* y los valores negativos de *x* estarían asociados a los positivos de *y*, lo que conduciría a obtener valores negativos de *xy*.

	X	***Y***	***x***	***y***	***xy***	x^2	y^2
	1	4	−3	−5	15	9	25
	3	6	−1	−3	3	1	9
	5	10	1	1	1	1	1
	5	12	1	3	3	1	9
	6	13	2	4	8	4	16
Total	20	45	0	0	30	16	60
Media	4	9	0	0	6		

Tabla 7.3 Cálculo de *r.*

El coeficiente de correlación de Pearson, *r*, se diseñó de tal forma que la correlación entre estatura y peso sea la misma, no importando si la estatura se mide en metros, pies, pulgadas, etcétera. Para alcanzar esta propiedad, el coeficiente de correlación se calcula dividiendo la suma de la columna *xy* entre la raíz cuadrada del producto de la suma de la columna x^2 y la suma de la columna y^2 (Σy^2). La fórmula es:

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

Por tanto

$$r = \frac{30}{\sqrt{16(60)}} = \frac{30}{\sqrt{960}} = 9.968$$

Una fórmula alterna que evita el cálculo de las desviaciones es:

$$r = \frac{\sum XY - \dfrac{\sum X \sum Y}{n}}{\sqrt{\left(\sum X^2 - \dfrac{(\sum X)^2}{n}\right)}\sqrt{\left(\sum Y^2 - \dfrac{(\sum Y)^2}{n}\right)}}$$

PREGUNTAS

1. ¿Cuál es la desviación respecto a la media que le corresponde al dato igual a 6?

X	2	4	6

a) Respuesta _____.

2. ¿Cuál es la suma de xy?

X	2	4	6
Y	4	3	5

a) Respuesta _____.

3. ¿Qué efecto tiene en la correlación sumar 12 a cada dato de una de las variables?

a) La correlación puede aumentar o disminuir dependiendo de los datos.

b) La correlación se incrementa.

c) La correlación no muestra cambio.

4. ¿Cuál es la correlación entre las variables X y Y que se muestran en seguida?

X	13	8	8	6	9	11	9	11
Y	11	8	6	8	10	9	8	12

a) Respuesta _____.

Ley de la suma de las varianzas II

Recuerda que cuando las variables X y Y son independientes, la varianza de la suma o diferencia de X y Y viene dada por:

$$\sigma^2_{X\pm Y} = \sigma^2_X + \sigma^2_Y$$

Es decir "La varianza de X más (o menos) Y es igual a la varianza de X más la varianza de Y".

Cuando X y Y están correlacionadas, se debe usar la siguiente fórmula:

$$\sigma^2_{X\pm Y} = \sigma^2_X + \sigma^2_Y \pm 2\rho\sigma_X\sigma_Y$$

donde, ρ es la correlación entre X y Y en la población. Por ejemplo, si la varianza de la parte verbal del examen de admisión SAT fue de 10 000, la varianza de la parte de matemáticas fue de 11 000 y la correlación entre ambas pruebas fue de 0.50, entonces la varianza total del examen de admisión SAT (verbal + matemáticas) fue:

$$\sigma_{oral\,+\,mat.} = 10\,000 + 11\,000 + (2)(0.5)\sqrt{10\,000}\sqrt{11\,000}$$

la cual es igual a 31 488. La varianza de la diferencia es:

$$\sigma_{oral\,-\,mat.} = 10\,000 + 11\,000 - (2)(0.5)\sqrt{10\,000}\sqrt{11\,000}$$

la cual es igual a 10 512.

Si la varianza y la correlación se calculan para una muestra, se utiliza la siguiente notación para expresar la ley de la suma de las varianzas.

$$s^2_{X \pm Y} = s^2_X + s^2_Y \pm 2r\, s_X\, s_Y$$

PREGUNTAS

1. Si la varianza de un grupo A de datos es 100, la varianza de un grupo B es 225 y la correlación entre los grupos es 0.5, ¿cuál es la varianza de A + B?

a) Respuesta _____.

2. Si la varianza de un grupo A de datos es 100, la varianza de un grupo B es 225 y la correlación entre los grupos es 0.5, ¿cuál es la varianza de A – B?

a) Respuesta _____.

Introducción a la regresión lineal simple

En la regresión lineal simple, nosotros predecimos valores de una variable a partir de otra. A la variable que estamos prediciendo se le llama variable dependiente y nos referimos a ella como *Y*. La variable en la que estamos basando nuestras predicciones se llama variable independiente (o predictora) y nos referimos a ella como *X*. Cuando existe solamente una variable independiente, el método se conoce como regresión simple. En la regresión lineal simple los valores de *Y,* cuando se grafican como una función de *X,* se ajustan a una línea recta.

En la figura 7.10 se grafican los datos de la tabla 7.4. Se puede observar que existe una relación positiva entre la variable *X* y la variable *Y*. Si se estuviera tratando de predecir la variable *Y*, tomando como predictor a *X*, cuanto más alto sea el valor de *X*, más alto será el valor de *Y*.

X	*Y*
1.00	1.00
2.00	2.00
3.00	1.30
4.00	3.75
5.00	2.25

Tabla 7.4 Datos para ejemplo.

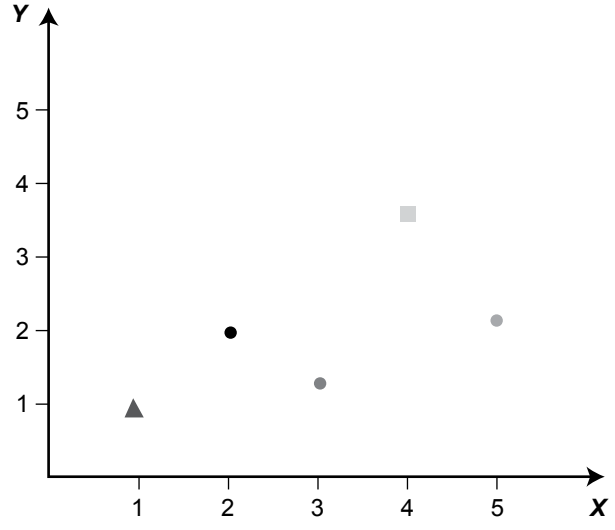


Figura 7.10 Diagrama de dispersión para los datos de ejemplo.

La regresión lineal consiste en encontrar la línea recta que mejor se ajuste a estos puntos. La línea que mejor se ajusta se llama línea de regresión. En la figura 7.11 la línea diagonal es la línea de regresión y representa los valores estimados o predichos de *Y*, para cada valor posible de *X*. Las líneas verticales, desde los puntos (que representan los valores observados) hasta la línea de regresión, representan los errores de predicción o de estimación. Como se puede observar, el tríangulo se encuentra muy cerca de la línea de regresión, por lo que su error de predicción es muy pequeño. En contraste, el cuadrado está más alejado de la línea de regresión; por tanto, su error de predicción es más grande.

El error de predicción para un punto dado es el valor del punto menos el valor estimado (el valor en la línea). En la tabla 7.5 se muestran los valores estimados ($\hat{Y}$) y los errores *de* estimación ($Y - \hat{Y}$). Por ejemplo, el primer punto tiene un valor *Y* de 1.00, y el valor estimado es 1.21; por tanto, el error de estimación es −0.21.

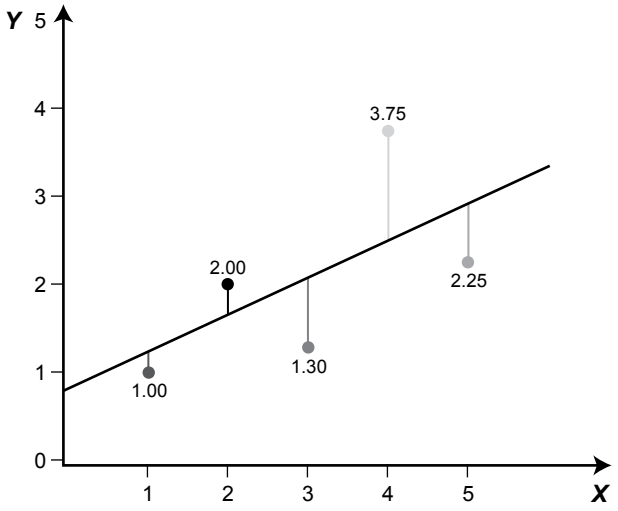


Figura 7.11 Diagrama de dispersión de los datos. Los puntos son los datos reales, la línea recta representa las predicciones; las líneas verticales entre los puntos y la línea recta representan los errores de predicción.

Puedes notar que no hemos especificado qué significa "línea de mejor ajuste". El criterio más utilizado para definir la línea de mejor ajuste es el que la define como la línea que minimiza la suma de los errores cuadrados de estimación. Éste es el criterio que fue utilizado para encontrar la línea en la figura 7.11. La última columna de la tabla 7.5 muestra el cuadrado de los errores de estimación. La suma de los cuadrados de los errores de la estimación que se muestra en la tabla 7.5 es el valor mínimo que se puede encontrar con cualquier línea de regresión.

X	*Y*	$\hat{Y}$	$Y - \hat{Y}$	$(Y - \hat{Y})^2$
1.00	1.00	1.210	–0.210	0.044
2.00	2.00	1.635	0.365	0.133
3.00	1.30	2.060	–0.760	0.578
4.00	3.75	2.485	1.265	1.600
5.00	2.25	2.910	–0.660	0.436

Tabla 7.5 Valores estimados ($\hat{Y}$) y los errores de estimación ($Y - \hat{Y}$).

La fórmula para la línea de regresión es:

$$\hat{Y} = bX + a$$

donde $\hat{Y}$ es el valor estimado o la predicción de Y, b es la pendiente de la línea y a es donde la línea intercepta al eje Y. La ecuación para la línea de la figura 7.11 es:

$$Y' = 0.425x + 0.785$$

Para $X = 1$,

$$Y' = (0.425)(1) + 0.785 = 1.21$$

Para $X = 2$,

$$\hat{Y} = (0.425)(2) + 0.785 = 1.64$$

A raíz de la llegada de las computadoras, la regresión lineal se calcula utilizando *software* estadístico. Sin embargo, los cálculos son relativamente fáciles y se describen aquí para quien esté interesado. Vamos a realizar los cálculos con los datos de la tabla 7.6. $\overline{X}$ es la media de X, $\overline{Y}$ es la media de Y, s_X es la desviación estándar de X, s_Y es la desviación estándar de Y, y r es la correlación entre X y Y.

$\overline{X}$	$\overline{Y}$	s_x	s_y	r
3	2.06	1.581	1.072	0.627

Tabla 7.6 Estadísticas para el cálculo de la línea de regresión.

La pendiente b se puede calcular como sigue:

$$b = r \times \frac{s_y}{s_x}$$

y el intercepto, a, se puede calcular como

$$a = \bar{y} - b\bar{x}$$

Para nuestros datos,

$$b = 0.627 \times \frac{1.072}{1581} = 0.425$$

$$a = 2.06 - (0.425)\,(3) = 0.785$$

Observa que los cálculos se han desarrollado para estadísticos muestrales, en lugar de hacerse para parámetros de una población. Las fórmulas son las mismas; simplemente usa los valores de la media, desviación estándar y correlación poblacional.

Un ejemplo de aplicación

En un estudio sobre el número de ventas realizadas y los años de experiencia de los vendedores, consideramos si podemos predecir las ventas promedio de un vendedor, si conocemos los años de experiencia del mismo.

La figura 7.12 muestra un diagrama de dispersión de las ventas como una función de los años de experiencia. Puedes ver que en la figura se aprecia una fuerte relación positiva. La correlación es 0.78. La ecuación de la regresión es:

$$\text{Ventas} = (10.5)(\text{años de experiencia}) + 1.2$$

Por tanto, predecimos que un vendedor con una experiencia de 3 años puede obtener ventas de:
Ventas $= (10.5)(3) + 1.2 = 32.7$

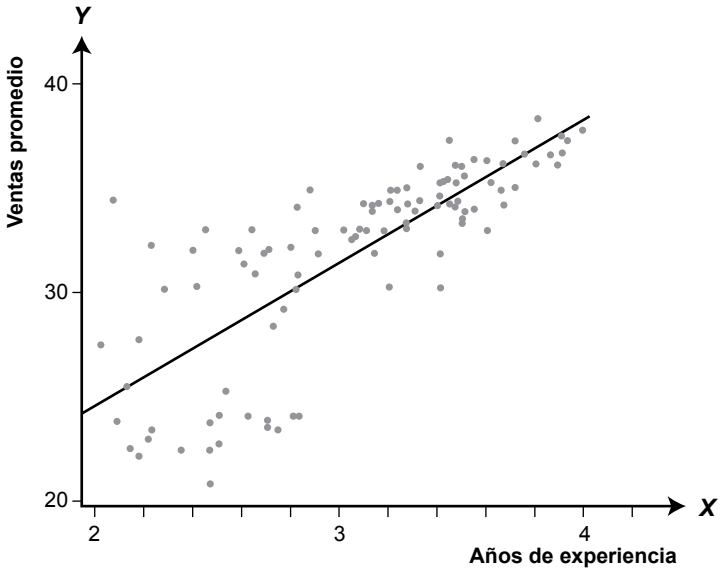


Figura 7.12 Ventas en función de los años de experiencia.

Suposiciones

Puedes sorprenderte, pero los cálculos mostrados en esta sección no se basan en ninguna suposición; desde luego, si la relación entre X y Y es no lineal, una función con diferente forma puede ajustarse mejor a los datos. La estadística inferencial para regresión se basa en varias suposiciones que se verán más adelante en este capítulo.

PREGUNTAS

1. La fórmula de la ecuación de regresión es $\hat{Y} = 3X - 2$. ¿Cuál sería el valor estimado de $\hat{Y}$, si $X = 4$?
2. Supón que es posible predecir el puntaje de una persona en un examen B, a partir del puntaje obtenido en un examen A. La ecuación de regresión es $B' = 2.3A + 9.5$. ¿Cuál sería el valor estimado de B, si una persona obtiene 40 puntos en el examen A?
3. Supón que una persona obtiene 32.5 puntos en el examen A y 92.25 en el B. Usando la misma ecuación de regresión de la pregunta anterior, ¿cuál sería el error de estimación?
4. ¿Cuál es el criterio que se usa generalmente para determinar la línea recta de mejor ajuste?

a) la recta que atraviesa la mayoría de los puntos.

b) la recta que tiene el mismo número de puntos arriba y abajo.

c) la línea que minimiza la suma de los errores cuadrados de estimación.

5. La media de X es igual a 3 y la media de Y es igual a 7. La recta de regresión para estimar $\hat{Y}$ a partir de X pasa necesariamente por el punto (3, 7).

a) verdadero.

b) falso.

6. Quieres tener la habilidad para predecir la medida del pie de las mujeres a partir de su estatura y para lograrlo recolectas información de tus compañeras de clase. La estatura media de las chicas en tu clase es 162.56 cm con una desviación estándar de 5.08 cm. La media del tamaño del pie es 20.32 cm con una desviación estándar de 2.54 cm. La correlación entre las dos variables es igual a 0.5. ¿Cuál es la pendiente de la recta de regresión?

7. Quieres tener la habilidad para predecir la medida del pie de las mujeres a partir de su estatura y para lograrlo recolectas información de tus compañeras de clase. La estatura media de las chicas en tu clase es 162.56 cm con una desviación estándar de 5.08 cm. La media del tamaño del pie es 20.32 cm con una desviación estándar de 2.54 cm. La correlación entre las dos variables es igual 0.5. ¿Cuál es el intercepto de la recta de regresión?

Particionando la suma de cuadrados

Un aspecto útil de la regresión es que se puede dividir la variación de Y en dos partes: la variación de los valores estimados y la variación de los errores de estimación. A la variación de Y se le llama suma de los cuadrados de Y y se define como la suma de los cuadrados de las desviaciones de Y respecto a su media. Cuando se trabaja con la población, la fórmula es:

$$SCY = \sum (Y - \mu_Y)^2$$

Donde SCY es la suma de los cuadrados de Y, Y es un valor individual de Y, μ_Y es la media de Y. Un ejemplo sencillo se muestra en la tabla 7.7. $\overline{Y}$ es 2.06 y la SCY es la suma de los valores de la tercera columna y es igual a 4.597.

Y	$Y - \mu_Y$	$(Y - \mu_Y)^2$
1	−1.06	−1.1236
2	−0.06	0.0036
1.3	−0.76	0.5776
3.75	1.69	2.8561
2.25	0.19	0.0361

Tabla 7.7 Ejemplo de *SCY.*

Cuando se realizan los cálculos tomando una muestra se debe utilizar la media de la muestra, representada con Y, en lugar de la media de la población.

$$SCY = \sum (Y - \mu_Y)^2$$

En algunos casos es conveniente usar fórmulas que utilicen las desviaciones en lugar de los datos originales. Las desviaciones son respecto a la media correspondiente. Es común que se utilicen letras minúsculas para representar las

desviaciones. De ahí que el valor y indica la diferencia entre Y y $\overline{Y}$. En la tabla 7.8 se muestra el uso de esta notación. Los datos son los mismos que los que se mostraron en la tabla 7.7.

Y	y	y^2
1	−1.06	−1.1236
2	−0.06	0.0036
1.3	−0.76	0.5776
3.75	1.69	2.8561
2.25	0.19	0.0361

Tabla 7.8 Ejemplo de *SCY* usando las desviaciones.

Los datos de la tabla 7.9 los habíamos usado en la sección "Introducción a los datos bivariados", y los copiamos aquí por comodidad. La columna X contiene los valores de la variable predictora, en la columna Y se muestran los valores de la variable dependiente. La tercera columna, y, contiene las diferencias entre los valores de la columna Y y la media de las Y.

	X	Y	y	y^2	Y'	y'	y'^2	$Y - Y'$	$(Y - Y')^2$
	1	1	−1.06	−1.1236	1.21	−0.85	0.7225	−0.21	0.044
	2	2	−0.06	0.0036	1.635	−0.425	0.1806	0.365	0.133
	3	1.3	−0.76	0.5776	2.06	0	0	−0.76	0.578
	4	3.75	1.69	2.8561	2.485	0.425	0.1806	1.265	1.6
	5	2.25	0.19	0.0361	2.91	0.85	0.7225	−0.66	0.436
Sumas	15	10.3	0	4.597	10.3	0	1.806	0	2.791

Tabla 7.9

La cuarta columna, y^2, es simplemente el cuadrado de la columna y. La columna Y', contiene los valores estimados de Y. En la sección "Introducción a los datos bivariados" encontramos que la ecuación de la regresión para estos datos era:

$$Y' = 0.425X + 0.785$$

Los valores de Y' fueron calculados con esta ecuación. La columna y' contiene las desviaciones de los valores de Y' respecto a su media y y'^2 es el cuadrado de esta columna. La penúltima columna $(Y - Y')$ contiene los valores observados de (Y) menos los valores estimados (Y'), errores de estimación. La última columna contiene el cuadrado de estos errores.

Estamos ahora en posición de particionar la *SCY*. Recuerda que la *SCY* es la suma de los cuadrados de las desviaciones respecto a su media. Es, por tanto, la suma de la columna y^2, que es igual a 4.597. La *SCY* puede ser dividida en dos partes: en la suma de los cuadrados de los valores estimados (*SSY'*) y en la suma de los cuadrados del error (*SCE*). La suma de los cuadrados de los valores estimados es la suma de las desviaciones al cuadrado de los valores estimados respecto a su media. En otras palabras, es la suma de la columna y^2 y es igual a 1.806. La suma de los cuadrados de error es la suma de los cuadrados de las diferencias entre los valores

observados y los estimados. Es, por tanto, la suma de la columna $(Y - Y')^2$ y es igual a 2.791. La suma se verifica:

$$SCY = SCY' + SCE$$

$$4.597 = 1.806 + 2.791$$

En la tabla 7.9 hay también otras particularidades que debemos notar. Observa que la suma de y y la suma de y' son ambas cero. Siempre serán cero, debido a que estas variables fueron calculadas restando a cada valor su media. También observa que la media de $(Y - Y')$ es 0, lo que indica que a pesar de que algunas Y's (estimaciones) son más altas que sus respectivas $Y's$ (observaciones) y otras son menores, la diferencia promedio es cero.

SCY es la variación total, *SCY′* es la variación explicada y la *SCE* es la variación no explicada. Por tanto, la proporción de la variación explicada puede ser calculada como:

$$\text{Proporción explicada} = SCY'/SCY$$

De manera similar, la proporción no explicada es:

$$\text{Proporción no explicada} = SCE/SCY$$

Existe una relación importante entre la proporción de la variación explicada y la correlación de Pearson: r^2 es la proporción de la variación explicada. Por tanto, si $r = 1$, entonces la proporción de la variación explicada es 1; si $r = 0$, entonces la proporción explicada es 0. Un último ejemplo, para $r = 0.4$, la proporción de la variación explicada es 0.16.

Debido a que la varianza se calcula dividiendo la variación entre N (para una población) o $n - 1$ (para una muestra), la relación escrita arriba en términos de la variación también se mantiene para la varianza. Por ejemplo:

$$\sigma^2_{total} = \sigma^2_{Y'} + \sigma^2_e$$

Donde el primer término es la varianza total, el segundo término es la varianza de Y' y el último término es la varianza de los errores de estimación $(Y - Y')$. De igual forma, r^2 es la proporción de la varianza explicada.

PREGUNTAS

1. Calcula la suma de cuadrados de *Y:* 2, 9, 11, 13, 15.

a) Respuesta _____.

2. Si la *SCY* es de 25.5 y la *SCY′* es de 18.3, ¿cuál es la *SCE*?

a) Respuesta _____.

3. Cuanto más grande sea la _____ más grande será la proporción de la variación explicada.

a) *SCY.*
b) *SCY′.*
c) *SCE.*
d) *Y.*

4. La proporción de la variación explicada es 0.3. Si la $SCY = 20$, ¿cuál será el valor de la *SCY*?

a) Respuesta _____.

5. Si $r = 0.84$, ¿qué proporción de la variación es explicada?

a) Respuesta _____.

Error estándar de regresión

En la figura 7.13 se muestran dos ejemplos de regresión. Puedes ver que, en la gráfica *A*, los puntos están más cerca de la línea que en la gráfica *B*. Por tanto, las predicciones realizadas con la gráfica *A* son más precisas que las realizadas con la gráfica *B*.

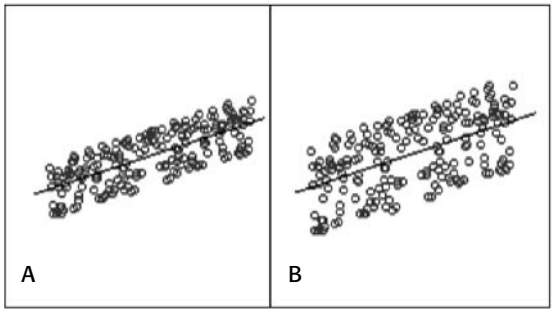


Figura 7.13 Diferentes precisiones en la predicción de la regresión.

El error estándar de la regresión es una medida de la precisión de las predicciones o estimaciones. Recuerda que la línea de regresión es la línea que minimiza la suma de las desviaciones al cuadrado de las estimaciones (también llamada suma de cuadrados del error). El error estándar de la regresión está relacionado con esta cantidad y se define como:

$$\sigma_{estimada} = \sqrt{\frac{\sum(Y - Y')^2}{n}}$$

Donde $\sigma_{estimada}$ es el error estándar de la regresión, Y es el valor observado, Y' es el valor estimado y n es el número de pares de valores. El numerador es la suma de las diferencias al cuadrado entre el valor estimado y el observado. Los datos de la tabla 7.10 son cinco valores de $X - Y$ de una población.

	X	Y	Y'	$Y - Y'$	$(Y - Y')^2$
	1	1	1.21	−0.21	0.044
	2	2	1.635	0.365	0.133
	3	1.3	2.06	−0.76	0.578
	4	3.75	2.485	1.265	1.6
	5	2.25	2.91	−0.66	0.436
Suma	15	10.3	10.3	0	2.791

Tabla 7.10 Datos de ejemplo.

La última columna muestra que la suma de los cuadrados de los errores de los valores estimados es 2.791. Por tanto, el error estándar de la regresión es:

$$\sigma_{estimada} = \sqrt{\frac{2.791}{5}} = 0.747$$

Existe una versión de la fórmula para el error estándar de la regresión, expresada en términos del coeficiente de correlación de Pearson:

$$\sigma_{estimada} = \sqrt{\frac{(1 - \rho^2)\ SCY}{n}}$$

donde ρ es el coeficiente de correlación de la población y *SCY* es,

$$SCY = \sum (Y - \mu_Y)^2$$

Para los datos de la tabla 7.10, $\mu_y = 10.30$, $SCY = 4.597$ y $r = 0.6268$; por tanto,

$$\sigma_{estimada} = \sqrt{\frac{(1 - 0.6268^2)(4.597)}{5}} = 0.747$$

El cual es igual al valor calculado anteriormente.

Se usan fórmulas similares cuando el error estándar de la regresión se calcula para una muestra. La única diferencia es que el denominador es $n - 2$ en lugar de n. La razón de que se utilice $n - 2$ en lugar de $n - 1$ es que se estiman dos parámetros (la pendiente y la intersección con el eje Y) para poder obtener la suma de los cuadrados. Se muestra a continuación la fórmula para calcular el error estándar de la regresión a partir de una muestra.

$$s_{estimada} = \sqrt{\frac{\sum (Y - Y')^2}{n - 2}}$$

$$s_{estimada} = \sqrt{\frac{2.791}{3}} = 0.964$$

$$s_{estimada} = \sqrt{\frac{(1 - r^2)SCY}{n - 2}}$$

PREGUNTAS

1. En una línea de regresión, si el error estándar de estimación es _____ las estimaciones serán más precisas.

a) mayor.

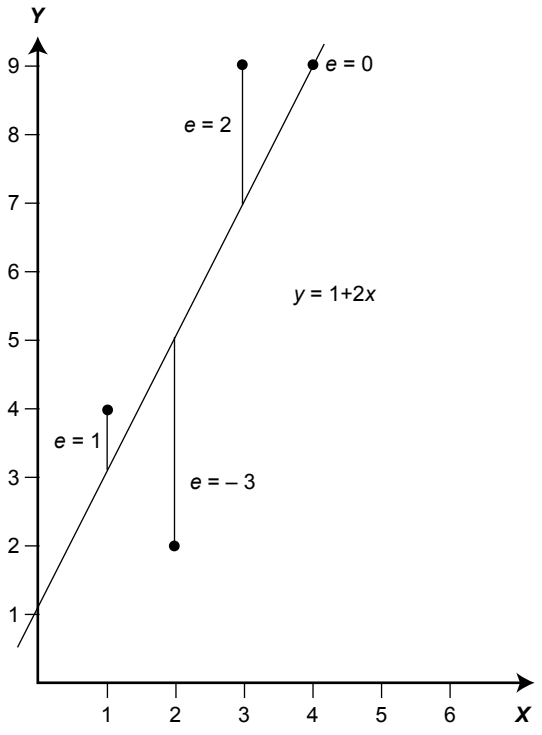
b) menor.

c) el error estándar de estimación no tiene relación con la precisión de las predicciones.

2. Se utilizó una línea de regresión para predecir *Y* a partir de *X* en una cierta población. En esta población, la $SCY = 50$, la correlación entre *X* y *Y* es 0.5 y la población es de 100. ¿Cuál es el error estándar de estimación?

3. Seleccionas una muestra de 10 alumnos de noveno semestre y quieres predecir la media de sus calificaciones, a partir de la media obtenida en octavo semestre. Determinas que la *SCE* es igual a 5.8. ¿Cuál es el error estándar de estimación?

4. La gráfica representa una línea para estimar *Y* a partir de *X*. Esta gráfica muestra el error de estimación para cada valor observado de *Y*. Utiliza esta información para calcular el error estándar de la regresión en esta muestra.



Estadística inferencial para *b* y *r*

En esta sección se muestra cómo realizar pruebas de significancia, así como el cálculo de los intervalos de confianza para la pendiente de la línea de regresión y el coeficiente de correlación de Pearson. Como podrás observar, si la pendiente de la regresión es significativamente diferente de cero, entonces el coeficiente de correlación es también significativamente diferente de cero.

Suposiciones

Aunque no se necesitó ninguna suposición para determinar la línea de mejor ajuste a los datos, sí se deben hacer algunas suposiciones para realizar inferencia estadística.

1. Linealidad: la relación existente entre las dos variables debe ser lineal.
2. Homocedasticidad: la varianza alrededor de la regresión lineal debe ser la misma para todos los valores de *X*. Una falta clara a esta suposición se muestra en la figura 7.14. Observa que los valores pronosticados para los estudiantes con promedios altos, en la preparatoria, son más precisos que los valores pronosticados para los promedios menores de los estudiantes. En otras palabras, los valores para los estudiantes con promedios altos en la preparatoria se encuentran más cerca de la línea de regresión, mientras que los valores para los promedios menores de los alumnos en la preparatoria se encuentran más alejados de ésta.
3. Los errores de predicción se distribuyen normalmente: esto significa que las desviaciones para la regresión lineal se distribuyen en forma normal. No significa que *X* o *Y* se distribuyan normalmente.

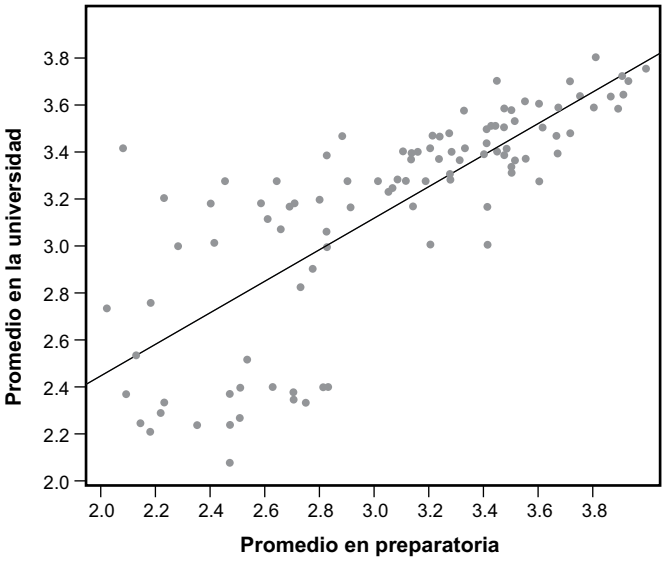


Figura 7.14 Diagrama de dispersión con ajuste a una línea recta del promedio de calificaciones de la preparatoria y de la universidad.

Prueba de significancia para la pendiente (*b*)

Recuerda la fórmula general para una prueba *t:*

$$t = \frac{\text{estadístico} - \text{valor hipotético}}{\text{estimación del error estándar del estadístico}}$$

En este caso, el estadístico es el valor de la pendiente b y su valor hipotético es 0. Los grados de libertad para esta prueba están dados por la expresión:

$$gl = n - 2$$

donde n es el número de pares de valores.

El error estimado estándar de b se calcula utilizando la siguiente fórmula:

$$s_b = \frac{s_{estimado}}{\sqrt{SCX}}$$

Donde s_b es el error estándar estimado de b, $s_{esimado}$ es el error estándar de regresión, SCX es la suma de las desviaciones al cuadrado de X respecto a su media, y se calcula como:

$$SCX = \sum (X - \mu_X)^2$$

donde μ_X es la media de X. Como se vio previamente, el error estándar de regresión puede calcularse mediante la siguiente fórmula:

$$s_{esimado} = \sqrt{\frac{(1 - r^2)SCY}{n - 2}}$$

La aplicación de estas fórmulas se ilustra utilizando los datos de la tabla 7.11. Estos datos ya fueron utilizados en la sección "Introducción a los datos bivariados". La columna X tiene los datos de la variable predictora y la columna Y los datos de la variable dependiente. La tercera columna, x, contiene las diferencias entre la columna X y su respectiva media, $\overline{X}$. La cuarta columna, x^2, es el cuadrado de la columna x.

La quinta columna, y, contiene la diferencia entre la columna Y y su media, $\overline{Y}$. La última columna, y^2, es simplemente el cuadrado de la columna y.

X	Y	x	x^2	y	y^2
1	1	−2	4	−1.06	−1.1236
2	2	−1	1	−0.06	−0.0036
3	1.3	0	0	−0.76	−0.5776
4	3.75	1	1	1.69	2.8561
5	2.25	2	4	0.19	0.0361
Suma 15	10.3	0	10	0	4.597

Tabla 7.11 Datos de ejemplo.

Los cálculos del error estándar de regresión (s_e) para estos datos se muestran en la sección error estándar de regresión y es igual a:

$$s_e = 0.964$$

SCX es la suma del cuadrado de las desviaciones de X respecto a su media. Es, por tanto, igual a la suma de la columna x_2, y es igual a:

$$SCX = 10.00$$

Ahora contamos con toda la información necesaria para calcular el error estándar de b:

$$s_b = \frac{0.964}{\sqrt{10}} = 0.305$$

Como se calculó previamente, la pendiente b es 0.425. Por tanto,

$$t = \frac{0.425}{0.305} = 1.39$$

$$gl = n - 2 = 5 - 2 = 3$$

El valor de p para una prueba de dos colas es aproximadamente igual a 0.2. Por tanto, la pendiente no es significativamente diferente de 0.

Intervalo de confianza para la pendiente

El método para calcular el intervalo de confianza para la pendiente de una población es muy similar a los métodos para calcular otros intervalos de confianza. Para calcular un intervalo de confianza al 95%, la fórmula es:

$$\text{límite inferior: } b - (t_{0.95})(s_b)$$

$$\text{límite superior: } b + (t_{0.95})(s_b)$$

Donde $t_{0.95}$ es el valor de t para calcular un intervalo de confianza al 95%.

Los valores de t que se utilizan en el intervalo de confianza pueden ser consultados en una tabla de distribución t. En la tabla 7.12 se muestran algunos valores de t para diferentes grados de libertad y niveles de confianza.

gl	0.95	0.99
2	4.303	9.925
3	3.182	5.841
4	2.776	4.604
5	2.571	4.032
8	2.306	3.355
10	2.228	3.169
20	2.086	2.845
50	2.009	2.678
100	1.984	2.626

Tabla 7.12 Una tabla pequeña de *t*.

Utiliza la tabla para encontrar el valor de t, necesario para calcular el intervalo de confianza.

Aplicando las fórmulas se obtiene:

$$\text{límite inferior: } 0.425 - (3.182)(0.305) = -0.55$$

$$\text{límite superior: } 0.425 + (3.182)(0.305) = 1.40$$

Prueba de significancia para la correlación

La fórmula para la prueba de significancia del coeficiente de correlación de Pearson se muestra a continuación:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

donde n es el número de pares de datos. Para nuestro ejemplo,

$$t = \frac{0.627\sqrt{5-2}}{\sqrt{1-0.627^2}} = 1.39$$

Nota que este valor es el mismo que se obtuvo en la prueba de b. Los grados de libertad se obtienen de la misma manera: $n - 2 = 3$.

Intervalo de confianza para la correlación

Existen varios pasos para calcular un intervalo de confianza para r.

Recuerda que en el capítulo 2, "Distribuciones muestrales" vimos que:

1. La distribución muestral del coeficiente de correlación de Pearson, r, es sesgada.
2. La transformación de Fisher de r a z' es normal.
3. $z' = 0.5 \ln[(1 + r)/(1 - r)]$.
4. z' tiene un error estándar de $\frac{1}{\sqrt{n-3}}$.

El cálculo del intervalo de confianza involucra los siguientes pasos:

1. Convertir r a z'. Para nuestros datos, $r = 0.627$ se transforma en $z' = 0.736$. Esto puede hacerse utilizando la fórmula que se indica en el punto 3.
2. Encontrar el error estándar de $z'(s_{z'})$. Para nuestro ejemplo, $n = 5$, por lo que $s_{z'} = 0.707$.

3. Calcula el intervalo de confianza en términos de z', utilizando la fórmula:

$$\text{límite inferior} = z' - (z_{0.95})(s_{z'})$$

$$\text{límite superior} = z' + (z_{0.95})(s_{z'})$$

Para nuestro ejemplo,

$$\text{límite inferior} = 0.736 - (1.96)\,(0.707) = -0.650$$

$$\text{límite superior} = 0.736 + (1.96)\,(0.707) = 2.122$$

4. Convertir el intervalo de z', a r.
Para nuestro ejemplo,

$$\text{límite inferior} = -0.57$$

$$\text{límite superior} = 0.97$$

El intervalo es muy amplio debido a que el tamaño de la muestra es muy pequeño.

PREGUNTAS

1. ¿Cuáles de las siguientes suposiciones se hacen al realizar inferencia estadística en regresión?

a) el error de estimación se distribuye en forma normal.

b) X se distribuye normalmente.

c) Y se distribuye normalmente.

d) la varianza alrededor de la línea de regresión es la misma para todos los valores de X.

e) la relación entre X y Y es lineal.

2. La pendiente de una línea de regresión es igual a 0.8 y el error estándar de la pendiente es 0.3. La muestra que se usó para calcular la línea de regresión fue de 12 elementos. Calcula un intervalo de confianza al 95% para la pendiente.

3. En una muestra de 20 elementos, la correlación entre dos variables es 0.5. Determina si la correlación es significativa con $\alpha = 0.05$.

4. Calcula el límite inferior de un intervalo de confianza al 95% para una correlación de 0.75 y un tamaño de muestra de 25.

Actividades

I. Contesta y resuelve los siguientes ejercicios para reafirmar los conceptos.

1. ¿Cuál es la ecuación de la línea de regresión? ¿Qué significa cada término?

2. La ecuación de una recta de regresión es $Y' = 2X + 9$.

a) ¿cuál sería el valor estimado de Y si X es igual a 6?

b) si el valor estimado de Y es 14, ¿cuál es el valor de X?

3. ¿Qué criterio se utiliza para decidir cuál es la recta de regresión de mejor ajuste?

4. ¿Qué mide el error estándar de estimación? ¿Cuál es la fórmula para calcular el error estándar de estimación?

5. En un análisis de regresión, la suma de cuadrados para los valores estimados es igual a 100, y la suma de cuadrados del error es 200. ¿Cuál es el valor de r^2?

a) en otro análisis de regresión, el 40% de la varianza es explicada. La suma de cuadrados totales es 1 000. ¿Cuál es la suma de cuadrados de los valores estimados?

X	*Y*
2	5
4	6
4	7
5	11
6	12

6. Para los datos *X, Y*, que se muestran al lado, calcula:

a) *r* y determina si es significativamente diferente de cero.

b) la pendiente de la línea de regresión y prueba si es significativamente diferente de cero.

c) un intervalo de confianza al 95% para la pendiente.

d) ¿qué suposiciones se necesitan hacer para realizar inferencias estadísticas en regresión lineal?

7. La correlación entre los años de educación y el salario para una muestra de 20 personas que trabajan en cierta compañía es 0.4.

a) ¿Es esta correlación estadísticamente significativa a un nivel del 0.05?

8. Se toma una muestra de valores *X* y *Y*, y se utiliza una línea de regresión para predecir *Y* a partir de *X*. Si la $SCY' = 300$, $SCE = 500$ y $n = 50$:

a) ¿cuál es el valor de: *SCY*?

b) ¿cuál es el error estándar de estimación?

c) ¿cuál es r^2?

9. Utilizando la regresión lineal, estima el valor de *Y* si *X* es igual a 43.

X	*Y*	*X*	*Y*
59	56	47	50
52	63	55	63
44	55	49	57
51	50	45	73
42	48	46	46
41	58	60	60
45	36	65	47
27	13	64	73
63	50	50	58
54	81	74	85
44	56	59	44
50	64		

10. La ecuación de la línea de regresión para estimar el número de horas que pasan viendo televisión los niños (*Y*), a partir del número de horas que pasan viendo televisión sus padres es $Y' = 4 + 1.2X$.

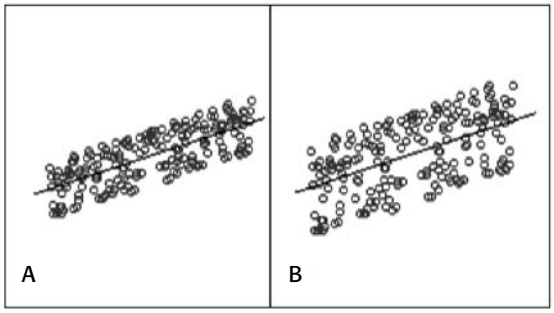
a) si el error estándar de *b* es 0.4, ¿es la pendiente estadísticamente significante a un nivel del 0.05?

b) si la media de *X* es 8, ¿cuál es la media de *Y*?

11. Con los datos proporcionados en la siguiente tabla, calcula la ecuación de la recta de regresión para estimar *Y* a partir de *X*.

μ_x	μ_y	s_x	s_y	*r*
10	12	2.5	3	−0.6

a) ¿Quién tiene un error estándar de estimación mayor, *A* o *B*?



12. Falso/Verdadero: Si la pendiente en un análisis de regresión lineal simple es estadísticamente significativa, entonces el coeficiente de correlación también tiene que ser significativo.

13. Falso/Verdadero: Si la pendiente de la regresión entre X y Y es mayor en la población 1 que en la población 2, la correlación necesariamente debe ser mayor en la población 1. ¿Por qué sí o por qué no?

14. Falso/Verdadero: Si la correlación es 0.8, entonces se explica el 40% de la varianza.

15. Falso/Verdadero: Si el valor real de Y es 31, pero se estima en 28, entonces el error de estimación es 3.

II. Resuelve los siguientes ejercicios de aplicación

1. Se hizo una encuesta a una muestra de 10 estudiantes de tercer semestre de la licenciatura en Administración del grupo 1301 del semestre 2009-1, y se encontraron los siguientes datos:

Estudiante	Estatura (m)	Peso (kg)
1	1.50	48
2	1.54	50
3	1.60	52
4	1.72	70
5	1.80	72
6	1.50	50
7	1.61	57
8	1.54	54
9	1.63	80
10	1.70	62

a) traza un diagrama de dispersión para estos datos.

b) ¿qué indica este diagrama acerca de la relación entre las dos variables?

c) traza una recta que pase por los datos, para aproximar una relación lineal entre la estatura y el peso.

d) aplica el método de mínimos cuadrados para plantear la ecuación estimada de regresión.

e) predice el peso de un estudiante que mide 1.75 metros.

f) calcula el error estándar en la regresión.

g) calcula el coeficiente de correlación y el coeficiente de determinación e interprétalos.

2. Los datos siguientes muestran las ventas (en miles de cajas) y los costos de un anuncio publicitario para la televisión (en millones de pesos) para 7 marcas principales de refrescos.

Marca	Gastos de publicidad ($)	Ventas de cajas(miles)
Coca-Cola	13.0	19.3
Pepsi-Cola	9.4	13.8
Sprite	6.4	8.4
Diet Coke	5.7	5.5
7-Up	4.2	5.9
Jarritos	2.9	5.3
Boing	1.6	2.5

a) dibuja el diagrama de dispersión, ¿qué parece indicar este diagrama acerca de la relación entre las dos variables?

b) traza una recta que pase por los datos, para aproximar una relación lineal entre los gastos del anuncio y las ventas.

c) aplica el método de los cuadrados mínimos para plantear la ecuación estimada de regresión.

d) predice las ventas para una marca que decida gastar 7 millones de pesos en un anuncio publicitario.

e) calcula el error estándar en la regresión.

f) calcula el coeficiente de determinación y correlación e interprétalos.

3. La *Revista del Consumidor* publicó en su número 381 de noviembre de 2008 la siguiente información acerca del uso de los teléfonos celulares:

Año	Usuarios que compran tiempo aire (en miles de usuarios)
2000	1 628
2001	1 784
2002	2 006
2003	2 029
2004	2 508
2005	3 268
2006	4 035
2007	5 199

a) traza un diagrama de dispersión para estos datos.

b) aplica el método de mínimos cuadrados para plantear la ecuación estimada de regresión.

c) traza una recta que pase por los datos, para aproximar una relación lineal entre la estatura y el peso.

d) predice cuántos usuarios comprarán tiempo aire para su teléfono celular para 2010.

e) calcula el error estándar en la regresión.

f) calcula el coeficiente de correlación y el coeficiente de determinación e interprétalos.

4. Un vendedor de Century 21 desea establecer la relación entre el tiempo en meses que están a la venta los departamentos antes de lograr su venta y el precio pedido por ellos. Los datos de una muestra de 9 departamentos se muestran a continuación:

Meses en venta	6.5	7.0	8.6	12.1	9.0	9.5	8.6	10.6	15.0
Precio pedido (en miles de pesos)	800	1 000	990	1 250	1 400	1 100	990	990	1 250

a) traza un diagrama de dispersión para estos datos.

b) aplica el método de mínimos cuadrados para plantear la ecuación estimada de regresión.

c) determina cuánto tiempo se tardará en vender un departamento que cuesta 1 500 000 pesos.

d) calcula el error estándar en la regresión.

e) calcula el coeficiente de correlación y el coeficiente de determinación e interprétalos.

5. El Organismo Operador de agua en el Municipio de Cuautitlán Izcalli, OPERAGUA, quiere conocer la relación entre el consumo mensual domiciliario de agua y el tamaño de las familias; toma una muestra de 10 familias elegidas al azar y encuentra los siguientes datos:

Metros cúbicos consumidos	**65**	**120**	**130**	**43**	**140**	**90**	**180**	**64**	**79**	**92**
Miembros de la familia	2	7	9	4	12	6	9	3	3	4

a) traza un diagrama de dispersión para estos datos.

b) aplica el método de mínimos cuadrados para plantear la ecuación estimada de regresión.

c) establece cuántos metros cúbicos al mes consumirá una familia que tiene 10 miembros.

d) calcula el error estándar en la regresión.

e) calcula el coeficiente de correlación y el coeficiente de determinación e interprétalos.

6. El gerente de una mueblería FAMSA quiere conocer la relación de las ventas logradas por un vendedor en dos años, por lo que toma una muestra de 8 vendedores que lograron convertirse en el Vendedor del Mes y encontró los siguientes datos:

Unidades vendidas en 2007	170	133	86	161	112	133	136	82
Unidades vendidas en 2006	99	95	50	80	92	88	130	100

a) traza un diagrama de dispersión para estos datos.

b) aplica el método de mínimos cuadrados para plantear la ecuación estimada de regresión.

c) calcula el error estándar en la regresión.

d) calcula el coeficiente de correlación y el coeficiente de determinación e interprétalos.

7. Según el INEGI, los nacimientos registrados en el país en 2007 fueron:

Mes de registro	Nacimientos registrados
1 enero	220 670
2 febrero	211 330
3 marzo	213 299
4 abril	270 819
5 mayo	225 298
6 junio	205 572
7 julio	211 180
8 agosto	249 626
9 septiembre	220 666
10 octubre	241 529
11 noviembre	211 857
12 diciembre	173 237

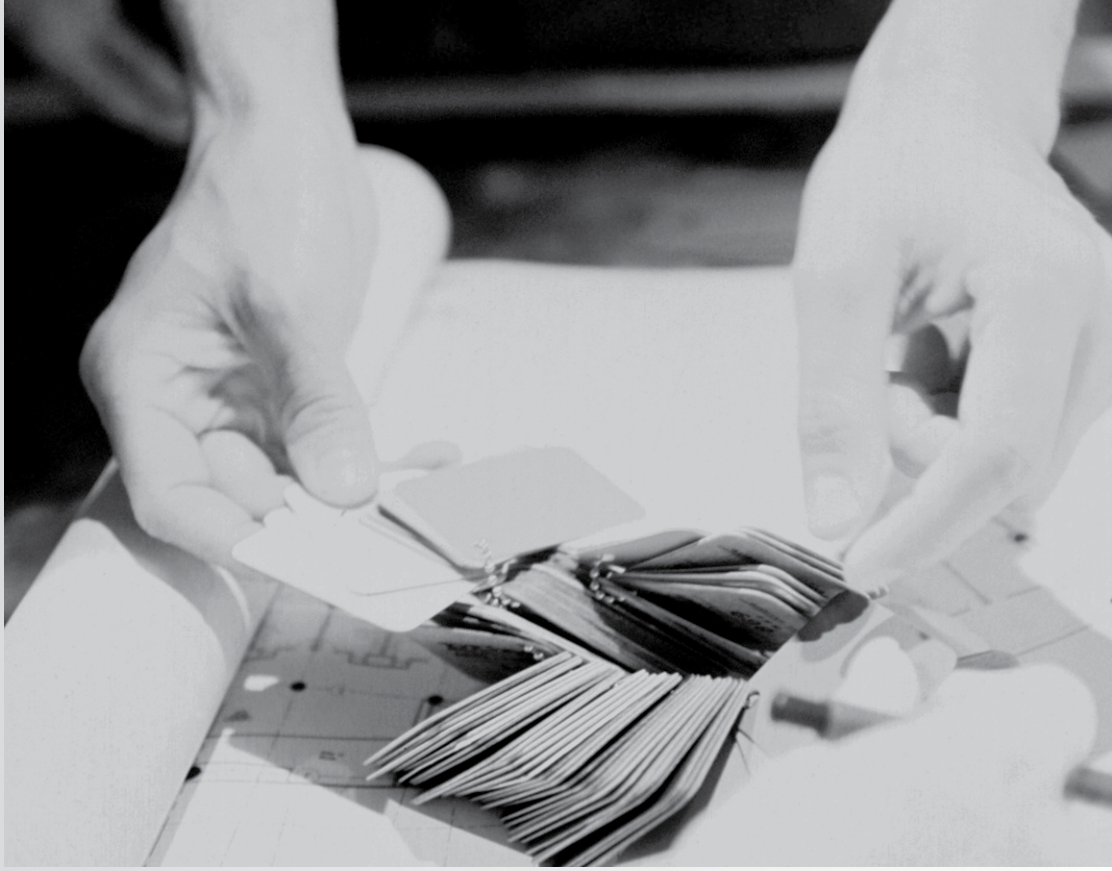
a) aplica el método de mínimos cuadrados para plantear la ecuación estimada de regresión.

b) el INEGI reportó que en julio de 2007 se registraron 211 180 nacimientos, utiliza la ecuación obtenida y predice cuántos debieron registrarse en ese mes, compara resultados y obtén tus conclusiones.

c) estima cuántos nacimientos se registraron en enero de 2008.

d) calcula el error estándar en la regresión.

e) calcula el coeficiente de correlación y el coeficiente de determinación e interprétalos.



8 Chi cuadrada

Chi cuadrada es una distribución particularmente útil en estadística. La primera sección describe los aspectos básicos de la distribución. Las dos secciones siguientes cubren las pruebas estadísticas más comunes en las que se usa la distribución Chi cuadrada. La sección "Tablas de una sola clasificación" muestra cómo se usa la distribución Chi cuadrada para probar la diferencia entre las frecuencias teóricas esperadas y las observadas. La sección "Tablas de contingencia" muestra cómo se usa la prueba de Chi cuadrada para probar la asociación entre dos variables nominales.

Distribución Chi cuadrada

La distribución de Chi cuadrada es la distribución de la suma de cuadrados de las desviaciones normales estandarizadas. Los grados de libertad de la distribución son igual al número de desviaciones que se sumaron. Por tanto, Chi cuadrada con un grado de libertad, que se escribe como $x^2(1)$, es simplemente la distribución de una sola desviación cuadrada normal. El área bajo la distribución de Chi cuadrada entre cero y 4 es la misma área bajo la distribución normal estándar entre cero y 2, ya que 4 es 2^2.

Considera el siguiente problema: tienes una muestra de dos datos seleccionados de una distribución normal estándar, el cuadrado de cada dato y la suma de los cuadrados. ¿Cuál es la probabilidad de que la suma de estos dos cuadrados sea igual o mayor a 6? La respuesta se puede dar utilizando la distribución Chi cuadrada con dos grados de libertad. Utilizando la calculadora de Chi cuadrada, encontramos que la probabilidad de encontrar un valor de Chi cuadrada (con 2 gl) igual o mayor a 6 es 0.050.

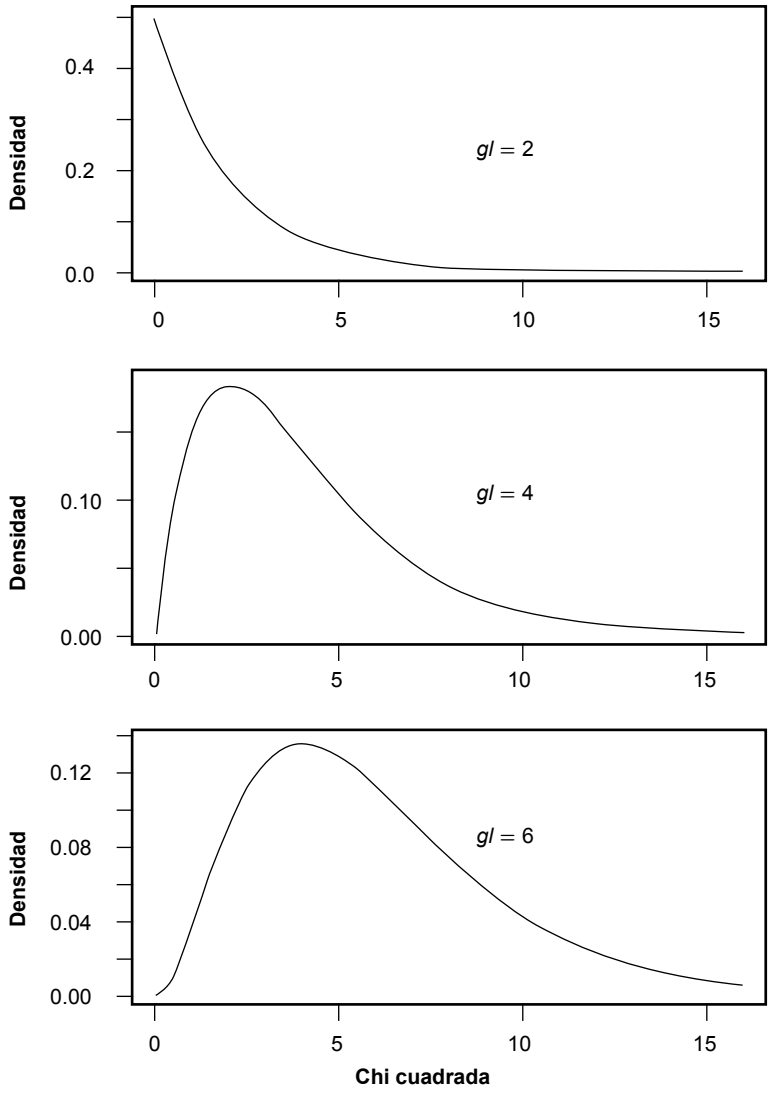


Figura 8.1 Distribución Chi cuadrada con 2, 4 y 6 grados de libertad.

La media de la distribución Chi cuadrada es igual a sus grados de libertad. La distribución Chi cuadrada es una distribución con sesgo positivo, y este sesgo se hace menor conforme aumentan los grados de libertad. Al incrementarse los grados de libertad, la distribución Chi cuadrada se aproxima a una distribución normal. En la figura 8.1 se muestra la función de densidad para la distribución de Chi cuadrada para tres grados de libertad. Observa cómo el sesgo disminuye al aumentar los grados de libertad.

La distribución Chi cuadrada es muy importante porque muchas pruebas estadísticas se aproximan a la distribución Chi cuadrada. Las dos pruebas más comunes que se realizan utilizando la distribución Chi cuadrada son las pruebas de las diferencias entre los valores teóricos esperados y las frecuencias observadas (tablas de una clasificación) y las pruebas de relación entre variables categóricas (tablas de contingencia). Existen otras pruebas del uso de la distribución Chi cuadrada que van más allá de lo tratado en este texto.

PREGUNTAS

1. Supón que tienes una muestra de 12 elementos seleccionados de una población normal; tienes los cuadrados de cada dato y la suma de los cuadrados. ¿Cuántos grados de libertad tendrá una distribución Chi cuadrada que corresponde a esta suma?

2. ¿Cuál es la media de una distribución Chi cuadrada con 8 grados de libertad?

3. ¿Cuál distribución Chi cuadrada se aproxima a una distribución normal?

a) una distribución Chi cuadrada con gl = 0.

b) una distribución Chi cuadrada con gl = 1.

c) una distribución Chi cuadrada con gl = 2.

d) una distribución Chi cuadrada con gl = 10.

4. Supón que tienes una muestra de 3 elementos seleccionados de una población normal; tienes los cuadrados de cada dato y la suma de los cuadrados. ¿Cuál es la probabilidad de que la suma de esos tres datos al cuadrado sea mayor o igual a 9?

Tablas de una sola clasificación

La distribución Chi cuadrada se usa para probar si los datos observados difieren significativamente de los valores teóricos esperados. Por ejemplo, para un dado legal, la probabilidad de obtener cualquiera de los 6 resultados posibles en un lanzamiento es 1/6. En la tabla 8.1 se muestran los resultados obtenidos al lanzar el dado 36 veces. Como se puede ver, en la tabla 8.1 algunos resultados ocurren más frecuentemente que otros, por ejemplo el "3" se obtuvo 9 veces, mientras el "4" sólo se obtuvo 2 veces. ¿Son estos datos consistentes con la hipótesis de que el dado es legal? Naturalmente, no esperaríamos que las frecuencias en la muestra de los 6 posibles resultados sean iguales, ya que existen diferencias debidas al azar. Por tanto, encontrar que las frecuencias son diferentes no significa que el dado no sea legal. Una forma de probar si el dado es legal es realizando una prueba de significancia. La hipótesis nula es que el dado es legal. Esta hipótesis se prueba calculando la probabilidad de obtener frecuencias tan diferentes o más que las obtenidas en la muestra respecto a la distribución uniforme de frecuencias. Si la probabilidad es suficientemente baja, entonces la hipótesis nula puede ser rechazada.

Resultado	Frecuencia
1	8
2	5
3	9
4	2
5	7
6	5

Tabla 8.1 Resultados de 36 lanzamientos de un dado

El primer paso para realizar la prueba de significancia consiste en calcular la frecuencia esperada para cada resultado, si la hipótesis nula fuera cierta. Por ejemplo, la frecuencia esperada de "1" es 6, ya que la probabilidad de obtener un "1" en un lanzamiento es 1/6 y en 36 lanzamientos la frecuencia esperada será:

$$\text{Frecuencia esperada} = (1/6)(36) = 6$$

Observa que las frecuencias esperadas son esperadas únicamente en un sentido teórico. En realidad, no "esperamos" que las frecuencias observadas sean exactamente iguales a las "frecuencias esperadas".

Los cálculos continúan como sigue: si representamos con E la frecuencia esperada y con O la observada, calculamos para cada resultado:

$$\frac{(E-O)^2}{E}$$

En la tabla 8.2 se muestran estos cálculos:

Resultado	Frecuencia esperada	Frecuencia observada	$\frac{(E-O)^2}{E}$
1	6	8	0.667
2	6	5	0.167
3	6	9	1.5
4	6	2	2.667
5	6	7	0.167
6	6	5	0.167

Tabla 8.2 Cálculo de Chi cuadrada.

Ahora sumamos todos los valores de la columna 4:

$$\sum \frac{(E-O)^2}{E} = 5.333$$

La distribución muestral de

$$\sum \frac{(E-O)^2}{E}$$

se distribuye aproximadamente como una Chi cuadrada con $k - 1$ grados de libertad, donde k es el número de categorías. Para nuestro ejemplo, el estadístico de prueba es:

$$\chi_5^2 = 5.333$$

Es decir, el valor de Chi cuadrada calculada con 5 grados de libertad es 5.333.

Utilizando la tabla de Chi cuadrada, se determina que la probabilidad de obtener un valor de Chi cuadrada igual o mayor a 5.333 es mayor a 0.3. Por tanto, la hipótesis nula que asegura que el dado es legal no puede ser rechazada.

Esta prueba de Chi Cuadrada puede usarse para probar otras desviaciones entre frecuencias observadas y esperadas. El ejemplo siguiente muestra una prueba aplicada al caso de la aplicación de un examen de admisión para la universidad y el promedio en preparatoria.

En la primera columna de la tabla 8.3 se muestran diferentes intervalos de puntajes obtenidos en el examen de admisión. En la segunda columna se muestran las proporciones de la distribución normal para cada intervalo. Las frecuencias esperadas (E) se calculan multiplicando el número total de puntos, 105, por la proporción. La última columna muestra los puntos observados para cada intervalo. Puedes observar que las frecuencias observadas presentan gran variación respecto a las esperadas. Observa que si los puntajes se distribuyeran en forma normal únicamente habría un puntaje de 35 entre –1 y 0, y sin embargo se observaron 60.

Rango	Proporción	E	O
Más de 1	0.159	16.695	19
0 a 1	0.341	35.805	17
−1 a 0	0.341	35.805	60
Menos de −1	0.159	16.695	9

Tabla 8.3 Puntaje observado y esperado en el examen de admisión.

La prueba para determinar si los valores observados se desvían significativamente de los esperados se calcula de la manera siguiente:

$$\chi^2_3 = \sum \frac{(E-O)^2}{E} = 30.09$$

El subíndice 3 indica que hay 3 grados de libertad. Como antes, los grados de libertad son el número de resultados menos 1. Utilizando la tabla de la distribución Chi cuadrada se determina que $p < 0.001$ para este valor de Chi cuadrada. Por tanto, la hipótesis nula que asegura que los datos se distribuyen en forma normal debe ser rechazada.

PREGUNTAS

1. Compras una bolsa de 40 caramelos de 4 colores. Tienes curiosidad de saber si los colores tienen la misma probabilidad de aparecer en la bolsa o cuáles colores son los más probables. Si los 4 colores tienen la misma probabilidad, ¿cuántos caramelos de cada color esperas que hubiera en la bolsa?
2. Supón que abres la bolsa y encuentras 8 rojos, 5 verdes, 12 naranjas y 15 azules. En la prueba de la hipótesis nula, que asegura que los colores de los caramelos ocurren con la misma frecuencia, ¿cuál es el valor de Chi cuadrada?
3. Supón que tienes una perinola con 8 caras, y en cada cara está impreso un número. Quieres determinar si alguno de estos números tiene mayor probabilidad de ocurrir respecto a los otros. Giras 200 veces la perinola y registras la frecuencia de los resultados. En una prueba de Chi cuadrada para los datos anteriores, ¿cuál es el valor "*p*" de la probabilidad?

Tablas de contingencia

Esta sección muestra el uso de la distribución Chi cuadrada para realizar la prueba de significancia entre variables nominales. Por ejemplo, en la tabla 8.4, se muestran los datos de un estudio entre el número de robos por día y tiendas de autoservicio.

	Tienda de autoservicio				
Número de robos	**Walmart**	**Soriana**	**Comercial Mexicana**	**Bodega Aurrerá**	**Total**
1-5	15	24	25	29	93
>5	7	14	8	23	52
Total	22	38	33	52	145

Tabla 8.4 Frecuencias del número de robos por día y la tienda de autoservicio.

El problema que se debe resolver es si hay una relación significativa entre el número de robos por día y las tiendas de autoservicio. El primer paso es calcular las frecuencias esperadas para cada celda, suponiendo que no existe relación entre las variables (hipótesis nula). Estas frecuencias esperadas se calculan a partir de los totales como sigue: empezaremos calculando la frecuencia esperada para la combinación Walmart /0 – 5 robos. Nota que 22 de 145 días hubo robos en Walmart. La proporción de robos en esa tienda es por tanto de 0.1517. Si no hubiera relación entre la tienda de autoservicio y el número de robos por día, esperaríamos

que 0.1517 de los robos se hacen en tiendas Walmart. Debido a que en 93 de los casos hubo robos de (1 – 5), esperaríamos que (0.1517) (93) = 14.11 de los días hubo entre (1 – 5) robos y se presentaron en tiendas Walmart. En forma similar, esperaríamos (0.1517) (52) = 7.89 de los días hubo más de 5 robos y se presentaron en tiendas Walmart. En general, la frecuencia esperada para una celda en el renglón I y la columna j es igual a:

$$E_{ij} = \frac{T_i \times T_j}{T}$$

Donde E_{ij} es la frecuencia esperada para la celda *i, j*, T_i es el total del renglón *i*, T_j es el total de la columna *j*, y T es el número total de observaciones. Para la celda Walt Mart /0 – 5 robos, $i = 1, j = 1, T_i = 93, T_j = 22$ y $T = 145$. En la tabla 8.5 se muestran las frecuencias esperadas (en paréntesis) para cada celda.

	Tienda de autoservicio				
Número de robos	**Walmart**	**Soriana**	**Comercial Mexicana**	**Bodega Aurrerá**	**Total**
1-5	15	24	25	29	93
	(14.11)	(24.37)	(21.17)	(33.35)	
>5	7	14	8	23	15
	(7.89)	(13.63)	(11.83)	(18.65)	
Total	22	38	33	52	145

Tabla 8.5 Frecuencias observadas y esperadas para número de robos por día y la tienda de autoservicio.

Se calcula el valor de Chi cuadrada como sigue:

$$\chi_c^2 = \sum \frac{(E - O)^2}{E} = 3.687$$

Los grados de libertad son igual a $(r-1)(c-1)$, donde r es el número de renglones y c es el número de columnas. Para nuestro ejemplo, los grados de libertad son $(2-1)(4-1) = 3$. Usando la tabla de la distribución Chi cuadrada, se determina que el valor de la probabilidad de obtener un valor de Chi cuadrada igual o mayor a 3.687 con 3 grados de libertad es menor a 0.20. Por tanto, la hipótesis nula de no relación (independencia) entre el número de robos por día y tiendas de autoservicio debe aceptarse.

Una suposición importante en la prueba de Chi cuadrada de independencia es que cada sujeto debe aportar información a una celda solamente. Por tanto, la suma de las frecuencias de todas las celdas debe ser igual al número de sujetos que participaron en el experimento. Considera un experimento en el que cada uno de 16 estudiantes intenta resolver dos problemas de probabilidad. En la tabla 8.6 se muestran los datos.

	Problema 1	**Problema 2**
Resolvió	10	4
No resolvió	6	12

Tabla 8.6 Solución a los problemas de probabilidad.

Para estos datos, no sería válido el uso de la prueba de Chi cuadrada, ya que cada estudiante aporta información a dos celdas: a una celda sí resolvió el problema 1, y a otra sí resolvió el problema 2. El total de las frecuencias en las celdas es 32, mientras que el total de estudiantes participantes es únicamente 16.

La fórmula de Chi cuadrada produce un estadístico que es solamente aproximado a una distribución Chi cuadrada. Con el fin de que esta aproximación sea adecuada, el número total de elementos debe ser de 20 o más. Algunos autores afirman que debe hacerse la corrección por continuidad siempre que una frecuencia esperada sea menor de 5.

La corrección por continuidad, cuando se aplica a tablas de contingencia 2×2, se conoce como corrección de Yates.

PREGUNTAS

1. Un estudiante está interesado en saber la relación entre el género y la carrera que se estudia en la Universidad. Selecciona una muestra de hombres y de mujeres del campus y les pregunta si el área de su especialidad es Ciencias Químico Biológicas (CQB), Ciencias Administrativas (CA) o Ciencias Físico Matemáticas (CFM). Los resultados se muestran en la siguiente tabla. ¿Cuántas mujeres se esperan que estudien Ciencias Administrativas?

Género	**Área de especialidad**		
	CQB	CA	CFM
Femenino	10	14	10
Masculino	11	8	4

2. Se realizó una prueba de Chi cuadrada para determinar si existe una relación entre el género y la especialidad. ¿Qué valor de Chi cuadrada se obtiene?

3. Algunos autores mencionan que la corrección por continuidad debe usarse cuando se tiene una tabla de contingencia con:

a) sólo 4 celdas en total.

b) una frecuencia esperada en una celda menor a 5.

c) la frecuencia en algunas celdas es mucho mayor que en otras.

4. Supongamos que un experimentador le pregunta a un grupo de 60 participantes si se asustan al ver una película de terror. Después de preguntar, el experimentador les pide ver una película de terror y les vuelve a preguntar al final de ésta si sintieron miedo. Los datos experimentales aparecen en la siguiente tabla. ¿Puede el experimentador usar una prueba de Chi cuadrada para determinar si después de ver la película de terror aumentó el número de personas que contestaban sentir miedo al ver películas de terror?

a) sí

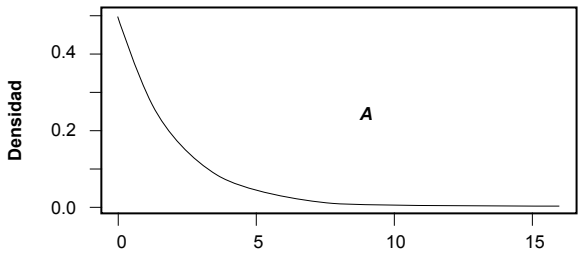
b) no

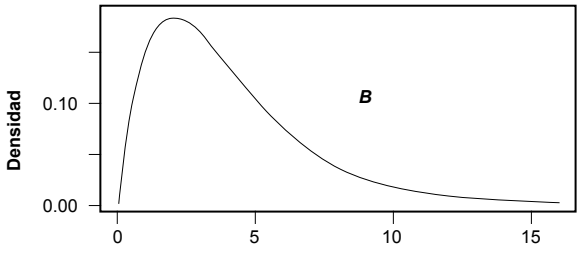
	Miedo	
	Sí	No
Antes	25	35
Después	37	23

Actividades

I. Contesta y resuelve los siguientes ejercicios para reafirmar los conceptos.

1. ¿Cuál de las dos distribuciones que se muestran en seguida (*A* o *B*) tiene mayores grados de libertad?





2. A 11 personas se les da a probar dos sabores de helado y se les pregunta cuál prefieren. Dos personas contestan que el primero, y nueve prefieren el segundo. ¿Es válido utilizar la prueba de Chi cuadrada para determinar si la diferencia entre las proporciones de preferencias es significante? ¿Por qué sí o por qué no?

3. Se sospecha que un dado está cargado. Se lanza 25 veces y se obtienen los siguientes resultados:

Resultado	Frecuencia
1	9
2	4
3	1
4	8
5	3
6	0

Realiza una prueba de significancia para probar si el dado está cargado. *a*) ¿Cuántos grados de libertad tiene el valor de Chi cuadrada? *b*) ¿Cuál es el valor de *p*?

4. Se realiza un experimento para estudiar la relación entre el hábito de fumar y la incontinencia urinaria. De 322 personas que presentaban incontinencia, 113 eran fumadores, 51 ex fumadores y 158 nunca habían fumado. 284 personas no presentaban incontinencia y 68 de ellas eran fumadores, 23 ex fumadores y 193 nunca habían fumado. *a*) Elabora una tabla para mostrar estos datos. *b*) Calcula la frecuencia esperada para cada celda. *c*) Realiza una prueba de significancia para averiguar si existe relación entre la incontinencia y el hábito de fumar. ¿Cuál es el valor de Chi cuadrada? ¿Cuál es el valor de *p*? *d*) ¿Cuáles son tus conclusiones?

5. Supón que en un despliegue de euforia en tu escuela un grupo de estudiantes de segundo año organizó una rifa gratis. Afirman que pusieron boletos con todos los nombres de todos los estudiantes de la escuela en la tómbola y sacaron 36 boletos. De los ganadores, 6 eran estudiantes del primer año, 14 eran de nuevo ingreso y 7 eran estudiantes del último año. Los resultados no te

parecen tan aleatorios. Piensas que es un poco sospechoso que los estudiantes de segundo año, que organizaron la rifa, también ganaron más premios. Tu escuela está compuesta de 30% de estudiantes de primer año, 25% de segundo año, 25% de nuevo ingreso y 20% de estudiantes del último año. *a*) ¿Cuáles son las frecuencias esperadas de los ganadores de cada clase? *b*) Realiza una prueba de significancia para determinar si los ganadores de los premios fueron distribuidos en todas las clases como se esperaba de acuerdo con los porcentajes de estudiantes en cada grupo. Calcula el valor de Chi cuadrada y el valor de *p. c*) ¿Cuáles son tus conclusiones?

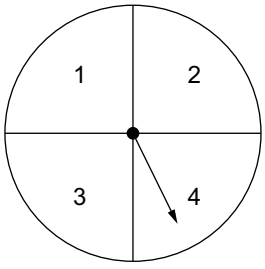
6. Algunos padres de los jugadores de una liga menor de béisbol piensan que se está presentando una tendencia. Creen que existe relación entre el número que porta cada jugador y su posición. Deciden registrar sus observaciones. Los datos hipotéticos aparecen en la siguiente tabla. Realiza una prueba Chi cuadrada para determinar si se verifica la sospecha de los padres acerca de que hay una relación entre el número del jugador y la posición correcta. ¿Cuál es el valor de Chi cuadrada? ¿Cuál es el valor de *p*?

	Bases	**Jardineros**	**Picher**	**Total**
0-9	12	5	5	22
10-19	5	10	2	17
20+	4	4	7	15
Total	21	19	14	54

7. Falso/Verdadero: Una distribución Chi cuadrada con 2 grados de libertad tiene una media más grande que una distribución Chi cuadrada con 12 grados de libertad.

8. Falso/Verdadero: Una distribución Chi cuadrada por lo general se usa para determinar si hay una relación significante entre dos variables continuas.

9. Falso/Verdadero: Imagina que quieres determinar si la aguja que se muestra en la figura de abajo es sesgada. Haces girar la aguja 50 veces y registras cuántas veces la flecha cae en cada sección. Podrías rechazar la hipótesis nula al nivel del 0.05 y determinar si la aguja tiene sesgo, si el valor de la Chi cuadrada que calculaste fue igual o mayor a 7.82.



10. Un estudio comparó a miembros de una clínica médica que presentaron quejas con una muestra aleatoria de miembros que no se quejaron. El estudio dividió a los quejosos en dos subgrupos: aquellos que presentaron quejas sobre el tratamiento médico y aquellos que presentaron quejas no médicas. A continuación se presentan los datos sobre el número total en cada grupo y el número que abandonó voluntariamente la clínica médica. Organiza los datos

en una tabla de doble entrada y analiza los datos con el fin de determinar si hay relación entre queja (ninguna queja, queja médica, queja no médica) y dejar la clínica (sí o no).

	Sin queja	Queja médica	Queja no médica
Total	743	199	440
Izquierdo	22	26	28

11. Imagina que crees que hay una relación entre el color de los ojos de las personas y el lugar donde prefieren sentarse en una sala de conferencias grande. Decides recolectar datos de una muestra aleatoria de personas individuales y realizar una prueba de Chi cuadrada de independencia. ¿Qué categorías deben aparecer en tu tabla de doble entrada? Usa la información para construir la tabla, y asegúrate de especificar los niveles para cada categoría.

12. Un geólogo recolectó manualmente muestras de piedra caliza en un área particular. Realizó una evaluación cualitativa, tanto de textura como de color, y obtuvo los resultados que se muestran a continuación. ¿Hay indicios suficientes para concluir que hay asociación entre la textura y el color en estas piedras? Explica tu respuesta.

	Color		
Textura	**Brillo**	**Medio**	**Oscuro**
Fino	4	20	8
Medio	5	23	12
Grueso	21	23	4

13. Supón que a los estudiantes de tu universidad se les pidió identificaran sus preferencias políticas (PRI, PAN, PRD, Independientes) y el sabor de helado que más les gusta (chocolate, vainilla, o fresa). Supón que sus respuestas se presentan en la siguiente tabla de contingencia.

	Chocolate	Vainilla	Fresa	Total
PRI	26	43	13	82
PAN	45	12	8	65
PRD	9	13	4	
Total		68	25	173

a) ¿qué proporción de personas prefieren helado de chocolate?

b) ¿qué proporción de personas tienen preferencia por el PRD?

c) ¿qué proporción de personas con preferencia por el PRD les gusta más el helado de chocolate?

d) ¿qué proporción de personas prefieren helado de chocolate y tienen preferencia por el PRD?

e) analiza los datos para determinar si existe relación entre las preferencias políticas y las preferencias al sabor del helado.

14. Una empresa de acreditación educativa recolectó datos acerca de la eficiencia terminal en una universidad. Entre 2 332 hombres, 1 343 no se habían titulado de la universidad, y entre 959 mujeres, 441 no se habían graduado.

a) organiza los datos en una tabla de contingencia para estudiar la relación entre género y titulación.

b) identifica una prueba apropiada para analizar la relación entre género y titulación. Realiza la prueba y redacta tus conclusiones.

II. Resuelve los siguientes ejercicios de aplicación

1. El gerente de ventas de una agencia de autos Nissan asegura que el color preferido por los clientes en las camionetas *X-trail* tiene la misma probabilidad de ser seleccionado por los clientes. La agencia vendió 15 camionetas de color plata, 10 de color blanco y 20 de color rojo. Supongamos que queremos utilizar un nivel de significancia de 0.05 para probar la aseveración de que los tres colores son igualmente probables de ser seleccionados por los clientes.

a) ¿cuál es la hipótesis nula?

b) ¿cuál es la frecuencia esperada para cada una de las tres categorías?

c) ¿cuál es el valor del estadístico de prueba?

d) ¿cuál es el valor crítico?

e) ¿qué concluyes acerca de la aseveración dada?

2. Una tienda Telcel recabó los siguientes datos e indican las frecuencias observadas del número de veces que un cliente pregunta por cuatro modelos de teléfono celular con características y precios similares: 15, 24, 14, 19. Supongamos que queremos utilizar un nivel de significancia de 0.05 para probar la aseveración de que los cuatro modelos tienen la misma predilección por los clientes. ¿Cuál es tu conclusión?

3. Un profesor de Estadística desea saber si el juego de azar Melate está balanceado y todos los números aparecen con la misma frecuencia; consultó los resultados de los últimos 200 sorteos, calculó el estadístico de la prueba: $x^2_{calculada} = 49.113$. Determina si es confiable este juego de azar con un nivel de significancia del 10%.

4. Arroja un dado 72 veces y demuestra que no está cargado con un nivel de significancia del 5%.

a) Compara tus resultados con los obtenidos por tus compañeros y discute los resultados obtenidos del estadístico de prueba.

5. La aseguradora ING de México seleccionó al azar reportes de choques de automóviles que se registran en el municipio de Cuautitlán Izcalli. Los resultados se incluyen en la siguiente tabla:

Día	Lunes	Martes	Miércoles	Jueves	Viernes	Sábado	Domingo
Número de choques	14	15	21	10	34	27	19

a) utiliza un nivel de significancia de 0.05 para comprobar si los choques de los automovilistas ocurren con la misma frecuencia en los diferentes días de la semana. ¿Cuál es tu conclusión?

6. Un juez del registro civil de Atizapán de Zaragoza desea saber si el número de niños que son registrados ocurre con la misma frecuencia para cualquier día de la semana. Se obtuvieron los registros de los niños elegidos al azar del registro civil; los resultados se presentan en la siguiente tabla:

Día de registro	Lunes	Martes	Miércoles	Jueves	Viernes
Número de niños	48	32	33	37	50

a) ¿cuál es tu conclusión utilizando un nivel $\alpha = 0.05$?

7. Un supervisor desea saber si los obreros que faltan al trabajo se presentan en igual número de veces para los tres turnos que tiene la fábrica; obtuvo los datos de las tarjetas de asistencia de los obreros de los últimos dos meses y los resumió en la siguiente tabla:

	1er. turno 6:00-14:00 horas	2do. turno 14:00-22:00 horas	3er. turno 22:00-6:00 horas
Obreros que faltaron al trabajo	46	54	80

a) ¿cuál es tu conclusión utilizando un nivel $\alpha = 0.01$?

8. El gerente de una cadena de tintorerías selecciona al azar los reportes de prendas dañadas de los últimos tres años. Los resultados se presentan en la siguiente tabla:

Mes	Ene.	Feb.	Mar.	Abr.	May.	Jun.	Jul.	Ago.	Sep.	Oct.	Nov.	Dic.
Número de prendas dañadas	15	14	6	8	12	3	10	9	8	10	12	19

a) utiliza un nivel de significancia de 0.05 para probar que el número de prendas dañadas que ocurren en los diferentes meses se presentan con la misma frecuencia.

9. El número de clientes atendidos por cajera en un centro comercial, en las llamadas "cajas rápidas" en un determinado periodo elegido al azar, se presenta a continuación:

Número de caja	1	2	3	4	5	6	7	8
Número de clientes atendidos	25	30	19	21	26	34	25	20

a) utiliza un nivel de significancia de 0.05 para probar que las cajeras tienen la misma habilidad para atender a los clientes.

10. Un vendedor de pantalones de mezclilla afirma que el número de prendas que vende es igual sin importar la marca; un ayudante que estudia la licenciatura desea probar la afirmación de su patrón, obtiene información de las ventas del último semestre y encuentra los siguientes datos:

Marca	Furor	Levi´s	Tommy	Zapa	Diesel
Pantalones vendidos	20	38	25	15	22

a) ¿cuál es la conclusión con $\alpha = 0.05$?

11. Un vendedor de pantalones de mezclilla afirma que el número de prendas que vende se presenta en una proporción de 1:3:3:5 según la marca; se obtiene información de las ventas del último año y encuentra los siguientes datos:

Marca	Furor 1	Tommy 3	Diesel 3	Levi´s 5
Pantalones vendidos	20	42	50	94

a) ¿cuál es la conclusión con $\alpha = 0.05$?

12. La panificadora Bimbo acaba de desarrollar un nuevo panqué sabor a naranja. Como una prueba de mercado pone a la venta el producto en cuatro zonas en las que dividió al Distrito Federal y registra el número de unidades vendidas. Los datos aparecen en la siguiente tabla:

Zona	Norte	Sur	Este	Oeste
Unidades vendidas	990	1 280	840	1 120

a) los datos demuestran que el producto es igualmente aceptado en todas las zonas en que se dividió el D.F. ¿Cuál es la conclusión con $\alpha = 0.01$?

13. Se realizó una encuesta entre los estudiantes de tercer semestre de la licenciatura en Administración para saber el tipo de cerveza que prefieren. Los datos se presentan en la siguiente tabla:

Género	Tipo de cerveza		
	Oscura	Clara	Light
Masculino	40	22	15
Femenino	28	42	18

a) determina con $\alpha = 0.05$ si género y tipo de cerveza son independientes.

14. Se hizo un estudio para determinar si existe dependencia entre el género de los consumidores y la marca de cigarros de los fumadores actuales (se considera como fumador actual quienes han fumado cigarros al menos un día en los 30 días previos a la encuesta). Realiza la prueba con $\alpha = 0.01$.

Género	**Marca**		
	Marlboro	Camel	Otra
Masculino	71	25	14
Femenino	56	21	23

a) determina con $\alpha = 0.05$ si la independencia entre las variables indicadas.

15. Las creencias asociadas al uso de tabaco que estimulan su consumo se presentan en la siguiente tabla:

Sexo	**Creencia**	
	Más amigos	Son más atractivos
Masculino	24	13
Femenino	56	21

a) determina con $\alpha = 0.05$ la independencia entre las variables indicadas.

16. Un vendedor estrella de seguros de ING de México asegura que el tipo de seguro que vende se presenta en una proporción de 1:3:6; se obtiene información de las ventas que realizó en el último trimestre y se resumió en la siguiente tabla:

Tipo de seguro	**Casa**	**Gastos médicos mayores**	**Automóvil**
	1	3	6
Número de ventas	14	28	68

a) determina si la afirmación del vendedor es correcta, con $\alpha = 0.01$.

17. El dueño de un lavado de autos ubicado en la zona esmeralda de Chiluca clasifica los coches a los que da servicio en tres categorías: chicos, medianos y grandes para fijar el precio que deben pagar sus clientes; debido al aumento de clientes que llevan camionetas SUV a servicio (autos grandes), está pensando en ampliar sus instalaciones y comprar equipo especial para este tipo de auto. La decisión de llevar a cabo el proyecto se va a tomar únicamente si verifica que el tipo de auto a los que da servicio se presenta en una proporción de 1:2:3; consulta las notas del último mes y observa lo siguiente:

Tipo de auto	Chico	Mediano	Grande
	1	3	6
Número de servicios	125	190	205

a) ¿cuál debe ser la decisión del dueño, con un nivel de significancia de 0.01?

18. Se hizo un estudio sobre la preferencia de los centros comerciales y la distancia en kilómetros de residencia de los clientes; los resultados se muestran a continuación:

Distancia de residencia	Centro comercial preferido			
	Bodega Aurrerá	Comercial Mexicana	Soriana	Walmart
0 a menos de 2 km	62	50	27	40
2 a menos de 5 km	120	72	12	18
5 km o más	78	40	8	48

a) utiliza esta información para verificar si el centro comercial preferido por un cliente es independiente de la distancia de su residencia, realiza la prueba con $\alpha = 0.01$.

19. El gerente de compras recibió un reporte del porcentaje de tornillos defectuosos que se encontraron recientemente. Recaba información de sus proveedores y el tamaño de tornillo que utiliza para fabricar cierto componente para automóvil; los datos obtenidos se presentan a continuación:

Proveedor	Tamaño de tornillo defectuoso		
	AA	AB	AC
Aceros nacionales	27	52	12
Tornillos especiales	38	46	13
Tornimex	19	10	30

a) realiza una prueba de hipótesis de independencia entre las dos variables indicadas con $\alpha = 0.05$.

20. Los datos de una encuesta que se realizó a los votantes en Cuautitlán, México, se presentan a continuación.

Nivel de estudios	Partido por el que votó en en las últimas elecciones municipales			
	PAN	PRI	PRD	OTRO
Básico	40	55	27	10
Media superior	110	72	32	12
Licenciatura	80	60	18	13

a) realiza una prueba de hipótesis de independencia entre las dos variables indicadas con $\alpha = 0.05$.

21. Un estudio de mercado para conocer el tipo de regalo que compran los clientes para el día de San Valentín en un centro comercial arrojó los siguientes datos:

Género	Tipo de regalo			
	Ropa	Perfume	Joya	Flores
Masculino	18	32	42	48
Femenino	27	38	30	21

a) realiza una prueba de hipótesis de independencia entre las dos variables indicadas con $\alpha = 0.01$.

22. Un hotel desea saber cómo consideran el servicio los clientes. Aplica una encuesta y encuentra los siguientes datos:

Servicio	Días de estancia		
	1-3	4-6	7 o más
Malo	20	14	7
Bueno	27	12	9
Excelente	15	17	14

23. Realiza una prueba de hipótesis y determina la independencia entre las variables indicadas con $\alpha = 0.05$.

24. El gerente de una pizzería recaba información sobre las ventas del último mes y la resume en la siguiente tabla:

Pizza	Día de la semana		
	Lun. - Mar.	Mié. - Jue.	Vier. - Sáb.
Hawaiana	10	21	40
Peperoni	8	15	20
Especial (ingrediente extra)	5	18	28

a) realiza una prueba de hipótesis y determina la independencia entre las variables indicadas con $\alpha = 0.01$.

25. Los obreros de una fábrica obtienen renumeraciones según la categoría que alcanzan después de realizar una prueba; se toma una muestra y se obtienen los siguientes datos:

Categoría	Escolaridad		
	Primaria	Secundaria trunca	Secundaria
A	12	64	24
B	27	38	30
C	32	40	42

a) realiza una prueba de hipótesis y determina la independencia entre las variables indicadas con $\alpha = 0.01$.

26. El Palacio de Acero quiere conocer el comportamiento de los vendedores (número de ventas), según el tipo de tienda en que son asignados, utilizando los datos muestrales de la siguiente tabla:

Vendedor	Tienda	Outlet
A	40	56
AA	51	64
AB	70	89

a) realiza una prueba de independencia entre el tipo de tienda y el vendedor con $\alpha = 0.01$.

Segunda parte
Laboratorio

Laboratorio de inferencia estadística: preguntas

Distribuciones muestrales

Demo básico

1. Supón que tenemos una urna que contiene muchas esferas, cada una con un número. Los números van de 0 a 32, y hay un número igual de esferas para cada número. Un estudiante va a realizar un experimento de acuerdo con el siguiente procedimiento: selecciona 5 esferas de la urna, calcula la media de los números, registra el resultado y vuelve a introducir las esferas. Repite el proceso 499 veces y, por tanto, tiene 500 promedios. Elabora una distribución de frecuencias para las 500 medias. En este caso, el tamaño de la muestra es...

a) Respuesta _____.

2. Supón que tenemos una urna que contiene muchas esferas, cada una con un número. Los números van de 0 a 32 y hay un número igual de esferas para cada número. Un estudiante va a realizar un experimento de acuerdo con el siguiente procedimiento: selecciona 5 esferas de la urna, calcula la media de los números, registra el resultado y vuelve a introducir las esferas. Repite el proceso 499 veces y, por tanto, tiene 500 promedios. Elabora una distribución de frecuencias para las 500 medias. En este caso, el número de muestras es...

a) Respuesta_____.

3. Hay el mismo número de esferas con el número 0, con el número 1, etc. Ya que los números están distribuidos uniformemente en la urna, a la distribución se le conoce como distribución uniforme. Esta distribución tiene un rango de valores de 0 a 32 y una media de 16. Con las medias de 100 muestras se construye una segunda distribución. ¿Cómo será esta distribución comparada con la distribución uniforme original?

a) la segunda distribución no es una distribución uniforme (y difiere más de lo que se podría esperar por variaciones debidas al azar).

b) la media de la segunda distribución está cerca de la media de la distribución uniforme.

c) la desviación estándar de la segunda distribución es menor que la de la distribución uniforme.

4. La media de la distribución muestral de la media:

a) es la media de la población.

b) depende de la forma de la distribución de la población.

5. La media de la distribución muestral del rango:

a) es igual al rango en la población.

b) es mayor que el rango en la población.

c) es menor que el rango en la población.

6. La distribución muestral del rango:

a) es simétrica.

b) tiene sesgo positivo (sesgo a la derecha).

c) tiene sesgo negativo (sesgo a la izquierda).

Demo de tamaño de la muestra

1. Supón que tenemos una urna que contiene muchas esferas, cada una con un número entero. Los números van de 0 a 32, y hay un número igual de esferas para cada número. Hay el mismo número de esferas con el número 0, con el número 1, etc. Los números están distribuidos uniformemente en la urna, y la media es 16. Selecciona 2 esferas de la urna, y calcula la media de los dos números; registra el resultado y vuelve a introducir las esferas en la urna. Ahora selecciona al azar 10 esferas y calcula la media de los números. ¿Qué tan probable es que las dos medias sean iguales?

a) muy probable.

b) muy improbable.

c) aproximadamente 50% de probabilidad.

2. Supón que tenemos una urna que contiene muchas esferas, cada una con un número entero. Los números van de 0 a 32, y hay un número igual de esferas para cada número. Hay el mismo número de esferas con el número 0, con el número 1, etc. Los números están distribuidos uniformemente en la urna, y la media es 16. Selecciona 10 esferas de la urna, y calcula la media de los números; registra el resultado y vuelve a introducir las esferas en la urna. Ahora selecciona al azar 25 esferas y vuelve a calcular la media. ¿Cuál media es más probable que sea menor de 10?

a) la media de las 10 esferas es ligeramente más probable.

b) la media de las 10 esferas es la mayoría de las veces más probable.

c) la media de las 25 esferas es ligeramente más probable.

d) la media de las 25 esferas es la mayoría de las veces más probable.

e) ambas tienen aproximadamente la misma probabilidad.

3. Supón que tenemos una urna que contiene muchas esferas, cada una con un número entero. Los números van de 0 a 32, y hay un número igual de esferas para cada número. Hay el mismo número de esferas con el número 0, con el número 1, etc. Los números están distribuidos uniformemente en la urna, y la media es 16. Selecciona 2 esferas de la urna, y calcula la media de los números, registra el resultado y vuelve a introducir las esferas en la urna. Ahora selecciona al azar 25 esferas y vuelve a calcular la media. ¿Cuál media es más probable que esté cerca de 16, el promedio de todos los números en la urna?

a) la media de las 2 esferas.

b) la media de las 10 esferas.

c) ambas tienen aproximadamente la misma probabilidad.

Demo teorema del límite central

1. ¿Cuál es el sesgo de una distribución normal?

a) Respuesta_____.

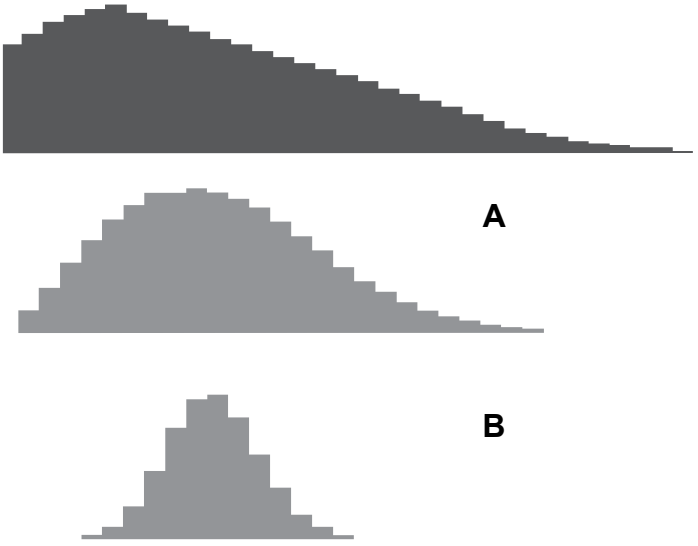
2. ¿Cuál de las siguientes afirmaciones acerca del efecto del tamaño de la muestra es verdadera?

a) el tamaño de la muestra afecta al sesgo de la distribución muestral de la media cuando la distribución de la población es normal.

b) sin tomar en cuenta la forma de la distribución de la población, la distribución muestral de la media se aproxima a una distribución normal, conforme se incrementa el tamaño de la muestra.

c) a medida que el tamaño de la muestra aumenta, la distribución muestral de la media se aproxima a la distribución de la población.

3. La gráfica superior muestra la distribución de la población. La distribución muestral de la media para $n = 10$ se muestra en



a) figura A.

b) figura B.

4. Si la distribución muestral de la media para $n = 5$ tiene una varianza igual a 50, ¿cuál sería la varianza de la distribución muestral de la media para $n = 10$?

a) Respuesta_____.

5. Si la distribución de la población tiene una desviación estándar igual a 10, ¿cuál sería la desviación estándar de la distribución muestral de la media para $n = 4$?

a) Respuesta_____.

Estimación

Simulación de sesgo y variabilidad

1. Si una población normal tiene una media de 20 y una desviación estándar de 5, ¿cuál sería la media de la distribución muestral de la media para $n = 10$?

a) Respuesta_____.

2. Para una distribución sesgada, la media de la distribución muestral de la media es:

a) la media de la población.

b) ligeramente menor que la media de la población.

c) ligeramente mayor que la media de la población.

3. Para la distribución normal, ¿es la mediana de la muestra, un estimador insesgado de la mediana de la población?

a) sí.

b) no.

4. Para una distribución con un sesgo positivo, ¿es la mediana de la muestra un estimador insesgado de la mediana de la población?

a) sí.

b) no.

5. La distribución muestral de la varianza:

a) tiene sesgo negativo.

b) está distribuida normalmente.

c) tiene sesgo positivo.

6. La media de la distribución muestral de la varianza, calculada con n en el denominador, es:

a) menor que la varianza de la población.

b) igual que la varianza de la población.

c) mayor que la varianza de la población.

7. Supón que se usa la fórmula con n en el denominador. Si la varianza de la población es 16, ¿cuál es la media de la distribución muestral de la varianza, para $n = 5$?

a) Respuesta _____.

8. Para una distribución normal, ¿cuál tiene menor variabilidad muestral, la media o la mediana?

a) la media.

b) la mediana.

c) son casi iguales.

9. Para una distribución normal, ¿cuál es aproximadamente la relación del error estándar de la mediana al error estándar de la media, cuando el tamaño de la muestra es 25?

a) Respuesta _____.

10. Para la siguiente distribución, ¿cuál tiene menor variabilidad muestral, la media o la mediana?

a) la media.

b) la mediana.

c) son casi iguales.

Simulación de intervalos de confianza

1. ¿Qué proporción de los intervalos de confianza para la media al 95% no contienen a la media de la población?

a) Respuesta _____.

2. ¿Cuál es más ancho, un intervalo de confianza al 95 o al 99%?

a) Respuesta _____.

3. ¿En qué forma el tamaño de la muestra afecta la confianza de que el intervalo contenga a la media de la población?

a) cuanto más grande sea el tamaño de la muestra, más grande será la confianza de que el intervalo contenga a la media de la población.

b) cuanto más grande sea el tamaño de la muestra, más pequeña será la confianza de que el intervalo contenga a la media de la población.

c) el tamaño de la muestra no afecta la confianza que se tiene en que el intervalo contiene a la media de la población.

4. ¿Es posible que un intervalo de 95% contenga a la media de la población, cuando un intervalo de 99% no la contiene?

a) sí.

b) no.

Prueba de medias

Demo de la distribución *t*

1. ¿Cuál distribución es la distribución *t*?

a) la distribución azul.

b) la distribución roja.

2. Con el fin de abarcar el 95% de la distribución, a partir de la media, ¿tienes que recorrer más distancia en ambas direcciones,

a) en una distribución *t*?

b) en una distribución normal?

Simulación de robustez

1. Cuando la hipótesis nula es verdadera, todas las suposiciones de la prueba son verdaderas y usas un nivel de significancia de 0.05, ¿cuál es la probabilidad de que la hipótesis nula sea rechazada?

a) Respuesta _____.

2. Cuando la hipótesis nula es verdadera, todas las suposiciones de la prueba son verdaderas y usas un nivel de significancia de 0.05, si realizas 100 simulaciones, ¿la hipótesis nula será rechazada en exactamente 5 simulaciones?

a) sí.

b) no.

3. Con un tamaño de muestra pequeña (5 por grupo) y 2 poblaciones ligeramente sesgadas, la probabilidad del error Tipo I, cuando se utiliza un nivel de 0.05, es:

a) 0.05.

b) ligeramente inferior a 0.05.

c) ligeramente superior a 0.05.

4. Con un tamaño de muestra pequeña (5 por grupo) y 2 poblaciones muy sesgadas, la probabilidad del error Tipo I, cuando se utiliza un nivel de 0.05, es:

a) 0.05.

b) ligeramente inferior a 0.05.

c) ligeramente superior a 0.05.

5. Con un tamaño de muestra pequeña (5 por grupo) y 2 poblaciones que difieren grandemente en sus desviaciones estándar (una es 3 veces más grande que la otra), la probabilidad del error tipo I, cuando se utiliza un nivel de 0.05, es:

a) menor a 0.04.

b) mayor a 0.06.

c) entre 0.04 y 0.06.

6. Con un tamaño de muestra moderadamente pequeño (20 por grupo) y 2 poblaciones que difieren grandemente en sus desviaciones estándar (una es 3 veces más grande que la otra), la probabilidad del error Tipo I, cuando se utiliza un nivel de 0.05, es:
   a) menor a 0.04.
   b) mayor a 0.06.
   c) entre 0.04 y 0.06.
7. Hay una circunstancia que conduce a obtener un valor alto de la probabilidad de cometer error Tipo I.
   a) se violan las suposiciones de normalidad y sesgo.
   b) los tamaños de muestra son diferentes. La muestra más pequeña procede de una población sesgada y la muestra más grande procede de una distribución normal.
   c) los tamaños de muestra son diferentes. La muestra más pequeña procede de una población con una varianza grande y la muestra más grande procede de una población con una varianza pequeña.
   d) los tamaños de muestra son diferentes. La muestra más pequeña procede de una población con una varianza pequeña y la muestra más grande procede de una población con una varianza grande.

Simulación de la prueba *t* para muestras correlacionadas

1. Una desviación estándar grande de las diferencias da como resultado:
   a) un valor grande de *t*.
   b) un valor pequeño de *t*.
2. Una correlación grande entre los tratamientos o condiciones, da como resultado:
   a) una desviación estándar grande de las diferencias.
   b) una desviación estándar pequeña de las diferencias.

Potencia

Demo de potencia 1

1. La potencia es mayor para un nivel de:
   a) 0.01.
   b) 0.05.
   c) 0.10.
2. ¿Qué aumenta más la potencia?
   a) un aumento en el tamaño de la muestra de 10 a 20.
   b) un aumento en el tamaño de la muestra de 20 a 30.
3. A medida que el tamaño de la muestra se incrementa, la potencia:
   a) se incrementa rápidamente al principio y lentamente al final.
   b) se incrementa linealmente.
   c) no se ve afectada.
4. A medida que se incrementa el tamaño de muestra, la probabilidad del error Tipo I:
   a) aumenta.
   b) disminuye.
   c) aumenta y después disminuye.

5. Si se incrementa la diferencia entre la media de la población y la media hipotética, entonces:

a) se incrementa la potencia.

b) disminuye la potencia.

6. ¿Cuál prueba es más potente?

a) La prueba de una cola.

b) La prueba de dos colas.

7. A mayor desviación estándar, mayor potencia.

a) falso.

b) verdadero.

8. ¿Cuáles de las siguientes afirmaciones son verdaderas?

a) aumentar el tamaño de la muestra, incrementa la potencia.

b) aumentar el tamaño de la muestra, disminuye la probabilidad de cometer error Tipo I.

c) cuanto menor sea la diferencia entre la media de la población y la media hipotética, es menor la potencia.

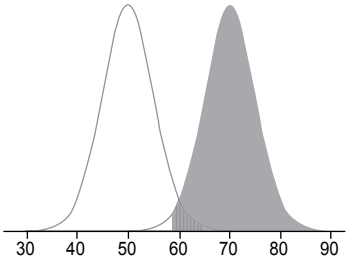
d) una prueba a un nivel de 0.05, tiene mayor potencia que una prueba a un nivel de 0.01.

Demo de potencia 2

1. La gráfica muestra la potencia para una prueba *z* de una cola, cuando la hipótesis nula asegura que la media de la población es 50. La distribución en rojo es la distribución muestral de la media suponiendo que la hipótesis nula es cierta. La distribución en azul es la distribución muestral de la media suponiendo que la media de la población es 70. Una media de la muestra por arriba de 58 se declara significativamente diferente a 50, a un nivel de 0.05. El área sombreada en la distribución roja es:

a) la probabilidad del error Tipo I.

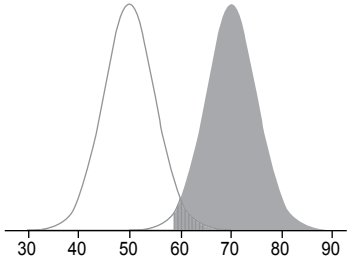
b) la potencia.



2. La gráfica muestra la potencia para una prueba *z* de una cola, cuando la hipótesis nula asegura que la media de la población es 50. La distribución en rojo es la distribución muestral de la media suponiendo que la hipótesis nula es cierta. La distribución en azul es la distribución muestral de la media suponiendo que la media de la población es 70. Una media de la muestra por arriba de 58 se declara significativamente diferente a 50, a un nivel de 0.05. El área sombreada en la distribución azul es:

a) la probabilidad del error Tipo I.

b) la potencia.



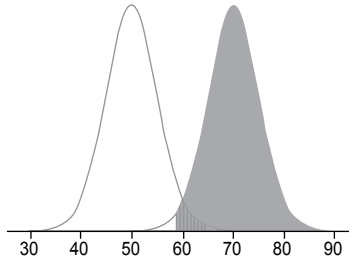
3. La gráfica muestra la potencia para una prueba *z* de una cola, cuando la hipótesis nula asegura que la media de la población es 50. La distribución en rojo es la distribución muestral de la media suponiendo que la hipótesis nula es cierta. La distribución en azul es la distribución muestral de la media suponiendo que la media de la población es 70. Una media de la muestra por arriba de 58 se declara significativamente diferente a 50, a un nivel de 0.05. Si la distribución azul tiene una media de 75, en lugar de 70, entonces:

a) la probabilidad del error Tipo I se incrementa.

b) la potencia se incrementa.

c) las dos distribuciones no se traslapan.

d) el punto crítico para declarar significancia se incrementa.

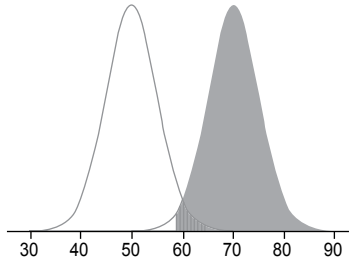


4. La gráfica muestra la potencia para una prueba *z* de una cola, cuando la hipótesis nula asegura que la media de la población es 50. La distribución en rojo es la distribución muestral de la media suponiendo que la hipótesis nula es cierta. La distribución en azul es la distribución muestral de la media suponiendo que la media de la población es 70. Una media de la muestra por arriba de 58 se declara significativamente diferente a 50, a un nivel de 0.05. Si la desviación estándar se reduce, entonces:

a) la probabilidad del error Tipo I se incrementa.

b) la potencia se incrementa.

c) el punto crítico para declarar significancia se incrementa.



Correlación y regresión

Demo de ajuste lineal

1. El criterio usual para determinar la mejor línea de ajuste es:

a) la suma de los errores cuadrados de la predicción o estimación.

b) la suma de las desviaciones absolutas.

c) La línea que atraviesa la mayoría de los puntos.

2. En la gráfica, el error de estimación para el punto especificado por la flecha se representa por:

a) la longitud de la línea roja horizontal.

b) la longitud de la línea vertical azul.

c) la línea que minimiza los errores cuadrados es por lo general la misma línea que minimiza los errores absolutos.

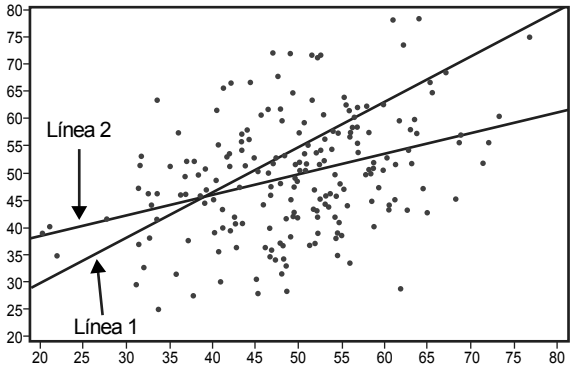
Demo de la línea de regresión

1. Cuanto menor sea el error estándar de regresión, es mejor el ajuste.

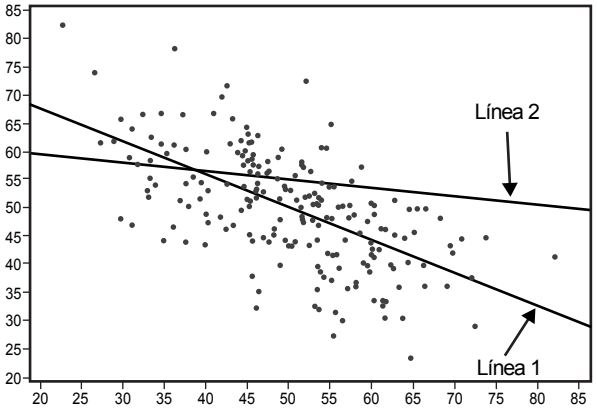
a) falso.

b) verdadero.

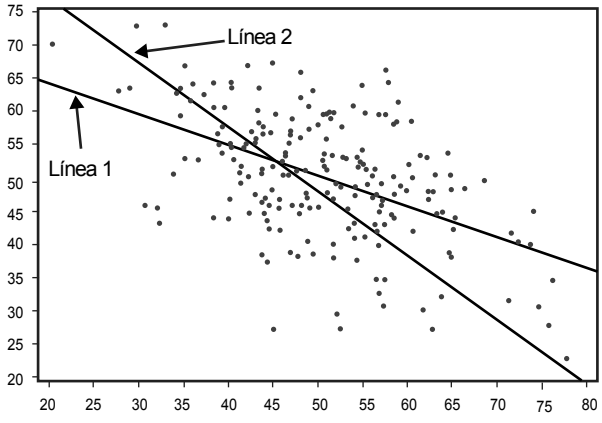
2. En la gráfica, ¿cuál de las dos líneas muestra mejor ajuste?
a) línea 1.
b) línea 2.



3. En la gráfica, ¿cuál de las dos líneas muestra mejor ajuste?
a) línea 1.
b) línea 2.



4. En la gráfica, ¿cuál de las dos líneas muestra mejor ajuste?
a) línea 1.
b) línea 2.



Chi cuadrada

Demo de pruebas de distribuciones

1. Cuando la distribución hipotética es la distribución normal, las frecuencias esperadas son mayores en la parte media de la distribución en comparación con las frecuencias esperadas en las colas:
   a) falso.
   b) verdadero.
2. Con 100 observaciones seleccionadas de una distribución normal, es muy probable que la hipótesis nula que asegura que la distribución es uniforme sea rechazada a un nivel de 0.01:
   a) falso.
   b) verdadero.
3. Cuando se prueban las desviaciones respecto a una distribución, la hipótesis nula es que todas las frecuencias observadas en la muestra deben ser igual a las esperadas:
   a) falso.
   b) verdadero.
4. Si la verdadera distribución es uniforme y la hipótesis nula asegura que la distribución es normal, entonces:
   a) las frecuencias observadas en las colas serán menores que las frecuencias esperadas.
   b) las frecuencias observadas deberán ser mayores en la parte media y menores en las colas.
   c) la mayor diferencia entre las frecuencias observadas y las esperadas ocurrirá en las colas.

Simulación de la tabla de contingencia 2×2

1. Cuando se usa la prueba de Chi cuadrada, la probabilidad de cometer el error Tipo I es igual al nivel de significancia.
   a) falso.
   b) verdadero.
2. Normalmente, la prueba de Chi cuadrada produce más el error Tipo I de lo que indica el nivel de significancia.
   a) falso.
   b) verdadero.
3. La corrección de Yates, por lo general, da como resultado acercar el valor de la probabilidad del error Tipo I, al nivel de significancia.
   a) falso.
   b) verdadero.
4. Cuando la hipótesis nula es falsa, la corrección de Yates conduce a rechazar la hipótesis nula más frecuentemente que la prueba de Chi cuadrada incorrecta.
   a) falso.
   b) verdadero.

Laboratorio de inferencia estadística: respuestas

Distribuciones muestrales

Demo básico

1. **R.** 5. El tamaño de la muestra es el número de datos incluidos en una muestra. En este caso, el estudiante seleccionó 5 esferas de la urna cada vez, y calculó la media de la muestra. Por tanto, el tamaño de la muestra es cinco.
2. **R.** 500. El tamaño de la muestra es el número de datos incluidos en una muestra. En este caso, el estudiante seleccionó 5 esferas de la urna, cada vez y calculó la media de la muestra. Por tanto, el tamaño de la muestra es 5. El estudiante repitió este proceso 499 veces más, por lo que el número de muestras es 500.
3. **R.** Todas las respuestas son correctas.
4. **R.** La media de la distribución muestral de la media es igual a la media de la población sin importar la forma de la distribución de la población.
5. **R.** La media de la distribución muestral del rango es menor que el rango en la población. El rango en la población es el mayor rango posible que se puede determinar en una muestra. La media del rango en una muestra será menor que el de la población.
6. **R.** La distribución muestral del rango tiene sesgo negativo. Con $n = 5$, presenta un sesgo muy pronunciado.

Demo de tamaño de la muestra

1. Aunque no es totalmente imposible, es muy poco probable que las dos medias sean iguales.
2. El tamaño de la muestra afecta la dispersión (variabilidad) de la distribución muestral de la media. Un tamaño de muestra pequeña produce una distribución con mayor dispersión. Por tanto, la media de la muestra más pequeña es más probable que se aleje más del valor medio de 16. Para este caso, la respuesta es que la media de las 10 esferas es la mayoría de las veces más probable.
3. El tamaño de la muestra afecta la dispersión (variabilidad) de la distribución muestral de la media. Para un tamaño de muestra grande, las medias de las muestras se agrupan más cerca del valor de la media de la distribución. En este caso, el promedio de las 25 esferas (tamaño de muestra mayor) es más probable que esté más cerca de 16.

Demo Teorema del Límite Central

1. Una distribución tiene sesgo si una de sus colas es más larga que la otra. Una distribución normal es simétrica respecto a su media, por tanto no tiene sesgo.
2. La distribución muestral de la media se distribuye en forma normal para cualquier tamaño de muestra, si la distribución de la población es normal. La distribución se aproxima a una distribución normal conforme se aumenta el tamaño de la muestra. Ésta es una parte muy importante del Teorema del Límite Central.
3. **R.** Figura B. Con una muestra de tamaño 10, la distribución muestral de la media se aproxima a una distribución normal y tiene menor dispersión.
4. La varianza de la distribución muestral de la media es inversamente proporcional al tamaño de la muestra. Por tanto, al aumentar al doble el tamaño de la muestra, la varianza debe ser la mitad, es decir, 25.
5. La varianza de la distribución muestral de la media es igual a la varianza dividida entre n. Para este ejemplo, la varianza de la población es igual a 100 (desviación estándar al cuadrado). Por tanto, la varianza de la distribución muestral de la media es $100/4 = 25$, entonces la desviación estándar (error estándar de la media) es igual a 5.

Estimación

Simulación de sesgo y variabilidad

1. **R.** 20. La media es un estimador insesgado. Por tanto, la media de la distribución muestral de la media, es igual a la media de la población.
2. **R.** La media de la población. La media de la muestra es un estimador insesgado de la media de la población sin importar la forma de la distribución.
3. **R.** Sí. La media de la distribución muestral de la mediana es la mediana de la población para distribuciones muestrales.
4. **R.** No. La media de la distribución muestral de la mediana es mayor que la mediana de la población para distribuciones con sesgos positivos.
5. **R.** Sesgo positivo. Tiene un gran sesgo positivo. Sin embargo, para tamaños de muestra grande el sesgo se reduce.
6. **R.** Mayor que la varianza de la población. La varianza de la muestra sobrestima la varianza de la población. Por tanto, la media de la distribución muestral será mayor que la varianza de la población.
7. **R.** 20. El estimador es sesgado, ya que sobrestima la varianza de la población en $n/n - 1$, que para este caso es $5/4 = 1.25$. Por tanto, la media de la distribución muestral es $(1.25)(16) = 20$. Ten presente que en esta simulación únicamente nos aproximamos a la distribución muestral; por tanto, se puede mostrar un valor ligeramente diferente.
8. La media.
9. La simulación muestra que es aproximadamente igual a 1.36.
10. **R.** La media. La media tiene menos variabilidad muestral que la mediana, aun cuando las distribuciones sean muy sesgadas.

Simulación de intervalos de confianza

1. **R.** 0.05. 95% de los intervalos contienen a la media, por lo que $1 - 0.95 = 0.05$ no la contiene.
2. **R.** El intervalo al 99%. Es más ancho el intervalo en el que mayor confianza se tiene que contenga a la media de la población.
3. **R.** el tamaño de la muestra no afecta la confianza de que el intervalo contenga a la media de la población. Tamaños de muestra grande disminuyen el ancho de los intervalos de confianza, pero no cambian el nivel de la confianza.
4. **R.** No. No es posible porque un intervalo al 99% incluye al intervalo al 95%.

Prueba de medias

Demo de la distribución *t*

1. **R.** Azul. La distribución *t* tiene más densidad en las colas que la distribución normal.
2. **R.** La distribución *t*. La distribución *t* tiene más densidad en las colas, que la distribución normal, por lo que se tiene que recorrer más camino a partir de la media para abarcar el 95% del área.

Simulación de robustez

1. **R.** 0.05. Cuando las suposiciones se cumplen, la probabilidad de cometer el error Tipo I es igual al nivel de significancia.

2. **R.** No. La probabilidad de que para una simulación se rechace la hipótesis nula es 0.05. Sin embargo, debido a que existe variación atribuida al azar, es improbable que exactamente en 5 simulaciones sea rechazada.
3. **R.** Es ligeramente inferior a 0.05. En este caso, la prueba es conservadora y mantiene la probabilidad del error Tipo I por debajo del nivel de 0.05.
4. **R.** Es ligeramente inferior a 0.05. En este caso, la prueba es conservadora y mantiene la probabilidad del error Tipo I, por debajo del nivel de 0.05. Esto es cierto aun para poblaciones muy sesgadas.
5. **R.** Mayor a 0.06. Para este caso, la prueba tiene un ligero sesgo positivo, por lo que la probabilidad del error Tipo I se encuentra entre 0.06 y 0.07.
6. **R.** Entre 0.04 y 0.06. En este caso, la prueba tiene un ligerísimo sesgo positivo, por lo que la probabilidad del error Tipo I se encuentra entre 0.05 y 0.06.
7. **R.** Los tamaños de muestra son diferentes. La muestra más pequeña procede de una población con una varianza grande y la muestra más grande procede de una población con una varianza pequeña.

Simulación de la prueba *t* para muestras correlacionadas

1. **R.** Un valor pequeño de *t*. La desviación estándar de las diferencias, cuando se divide entre la raíz cuadrada de *n*, es el error estándar de la media de las diferencias. Este error estándar es el denominador en la prueba de *t*, por lo que una desviación estándar pequeña da como resultado un valor grande de *t*.
2. **R.** Una pequeña desviación estándar de las diferencias. La ley de la suma de las varianzas indica que cuanto mayor sea la correlación, menor será la varianza (y, por tanto, la desviación estándar) de las diferencias.

Potencia

Demo de potencia 1

1. **R.** 0.10. Cuanto menos riguroso sea el nivel de significancia, más alta será la potencia.
2. **R.** Un incremento en el tamaño de la muestra de 10 a 20. El incremento es mayor para muestras pequeñas.
3. **R.** Se incrementa rápidamente al principio y lentamente al final.
4. **R.** Ninguno. El tamaño de la muestra no afecta la probabilidad del error Tipo I.
5. **R.** Aumenta la potencia. Las diferencias mayores son más fáciles de detectar.
6. **R.** La prueba de una cola. La prueba de una sola cola es más potente, siempre y cuando la dirección del efecto sea correcta.
7. **R.** Falso. Una desviación estándar grande da como resultado una menor potencia.
8. **R.** Verdadero, Falso, Verdadero, Verdadero.

Demo de potencia 2

1. **R.** La probabilidad del error Tipo I. El área sombreada en la distribución roja es la probabilidad de obtener una media que se declare significativamente diferente de 50, cuando la hipótesis nula es verdadera.
2. **R.** La potencia. El área sombreada en la distribución azul es la probabilidad de obtener una media que se declare significativamente diferente de 50, cuando la media de la población es 70.
3. **R.** Aumenta la potencia. La probabilidad del error Tipo I no aumenta, ya que es función de la distribución roja. La potencia se incrementa debido a que una mayor área de la distribución azul se encuentra a la derecha del punto crítico. Las

distribuciones se traslapan menos, pero se siguen traslapando. El punto crítico para declarar significancia no se calcula en función de la distribución azul.

4. **R.** La potencia aumentaría y el punto crítico se reduciría. La probabilidad del error Tipo I no es una función de la desviación estándar. La potencia se incrementaría debido a que el traslape de las distribuciones podría disminuir y el punto crítico disminuiría. El punto crítico se reduciría, porque se necesita una menor distancia a partir de la media de la distribución roja, para declarar significancia.

Correlación y regresión

Demo de ajuste lineal

1. **R.** La suma de los errores cuadrados de la estimación. La línea de mejor ajuste es la que minimiza la suma de los errores cuadrados de predicción.
2. **R.** La longitud de la línea vertical azul. Los errores siempre son las distancias verticales de los puntos a la línea. Muestran la desviación entre los valores observados y estimados.
2. **R.** Falso. Las líneas por lo general son similares, pero idénticas sólo bajo muy raras circunstancias.

Demo de la línea de regresión

1. **R.** Verdadero. El error estándar de regresión es el promedio de los errores cuadrados de estimación. El mejor ajuste será el que tenga errores más pequeños.
2. **R.** Línea 1.
3. **R.** Línea 2.
4. **R.** Línea 1.

Chi cuadrada

Demo de las pruebas de distribuciones

1. **R.** Verdadero. La distribución normal tiene mayor densidad en la parte media en comparación con la densidad en las colas.
2. **R.** Verdadero. Las distribuciones son muy diferentes; por tanto, con una muestra de 100 la hipótesis nula casi siempre se rechaza.
3. **R.** Falso. La hipótesis nula se formula para valores de la población.
4. La mayor diferencia entre las frecuencias observadas y esperadas ocurre en las colas. La distribución uniforme y la distribución normal son muy diferentes sobre todo en las colas. Las frecuencias observadas se obtuvieron muestreando una distribución uniforme y las esperadas una normal.

Simulación de la tabla de contingencia 2×2

1. **R.** Falso. La prueba sólo es aproximada, ya que la probabilidad de cometer el error Tipo I puede ser algo diferente al nivel de significancia.
2. **R.** Falso. Aunque en algunas ocasiones es verdadero, por lo general es falso.
3. **R.** Falso. Por lo general, la probabilidad de cometer el error Tipo I con la corrección es menor al nivel de significancia.
4. **R.** Falso. La corrección es una prueba conservadora y por tanto conduce a menos rechazos.

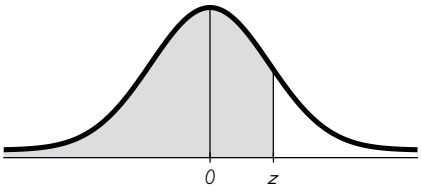
Tablas de distribución

Tabla de Distribución chi Cuadrada. Valores Críticos que corresponden a la cola superior

Grados de libertad	*p*													
	0.001	0.01	0.02	0.05	0.10	0.20	0.30	0.50	0.70	0.80	0.90	0.95	0.98	0.99
1	10.83	6.63	5.41	3.84	2.71	1.64	1.07	0.45	0.15	0.06	0.02	0.00	0.00	0.00
2	13.82	9.21	7.82	5.99	4.61	3.22	2.41	1.39	0.71	0.45	0.21	0.10	0.04	0.02
3	16.27	11.34	9.84	7.81	6.25	4.64	3.66	2.37	1.42	1.01	0.58	0.35	0.18	0.11
4	18.47	13.28	11.67	9.49	7.78	5.99	4.88	3.36	2.19	1.65	1.06	0.71	0.43	0.30
5	20.52	15.09	13.39	11.07	9.24	7.29	6.06	4.35	3.00	2.34	1.61	1.15	0.75	0.55
6	22.46	16.81	15.03	12.59	10.64	8.56	7.23	5.35	3.83	3.07	2.20	1.64	1.13	0.87
7	24.32	18.48	16.62	14.07	12.02	9.80	8.38	6.35	4.67	3.82	2.83	2.17	1.56	1.24
8	26.12	20.09	18.17	15.51	13.36	11.03	9.52	7.34	5.53	4.59	3.49	2.73	2.03	1.65
9	27.88	21.67	19.68	16.92	14.68	12.24	10.66	8.34	6.39	5.38	4.17	3.33	2.53	2.09
10	29.59	23.21	21.16	18.31	15.99	13.44	11.78	9.34	7.27	6.18	4.87	3.94	3.06	2.56
11	31.26	24.72	22.62	19.68	17.28	14.63	12.90	10.34	8.15	6.99	5.58	4.57	3.61	3.05
12	32.91	26.22	24.05	21.03	18.55	15.81	14.01	11.34	9.03	7.81	6.30	5.23	4.18	3.57
13	34.53	27.69	25.47	22.36	19.81	16.98	15.12	12.34	9.93	8.63	7.04	5.89	4.77	4.11
14	36.12	29.14	26.87	23.68	21.06	18.15	16.22	13.34	10.82	9.47	7.79	6.57	5.37	4.66
15	37.70	30.58	28.26	25.00	22.31	19.31	17.32	14.34	11.72	10.31	8.55	7.26	5.98	5.23
16	39.25	32.00	29.63	26.30	23.54	20.47	18.42	15.34	12.62	11.15	9.31	7.96	6.61	5.81
17	40.79	33.41	31.00	27.59	24.77	21.61	19.51	16.34	13.53	12.00	10.09	8.67	7.26	6.41
18	42.31	34.81	32.35	28.87	25.99	22.76	20.60	17.34	14.44	12.86	10.86	9.39	7.91	7.01
19	43.82	36.19	33.69	30.14	27.20	23.90	21.69	18.34	15.35	13.72	11.65	10.12	8.57	7.63
20	45.31	37.57	35.02	31.41	28.41	25.04	22.77	19.34	16.27	14.58	12.44	10.85	9.24	8.26
21	46.80	38.93	36.34	32.67	29.62	26.17	23.86	20.34	17.18	15.44	13.24	11.59	9.91	8.90
22	48.27	40.29	37.66	33.92	30.81	27.30	24.94	21.34	18.10	16.31	14.04	12.34	10.60	9.54
23	49.73	41.64	38.97	35.17	32.01	28.43	26.02	22.34	19.02	17.19	14.85	13.09	11.29	10.20
24	51.18	42.98	40.27	36.42	33.20	29.55	27.10	23.34	19.94	18.06	15.66	13.85	11.99	10.86
25	52.62	44.31	41.57	37.65	34.38	30.68	28.17	24.34	20.87	18.94	16.47	14.61	12.70	11.52

Tabla de Distribución *t*. Valores Críticos que corresponden a la cola superior

Grados de libertad	*p*											
	0.4	0.25	0.1	0.05	0.025	0.01	0.005	0.0025	0.002	0.0015	0.001	0.0005
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.321	159.153	212.205	318.309	636.619
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	15.764	18.216	22.327	31.599
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	8.053	8.891	10.215	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	5.951	6.435	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.030	5.376	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	4.317	4.524	4.800	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.207	4.442	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	3.991	4.199	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	3.835	4.024	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	3.716	3.892	4.144	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	3.497	3.624	3.789	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.550	3.706	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.489	3.639	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.438	3.583	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.395	3.535	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.358	3.494	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.326	3.459	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.298	3.428	3.610	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.273	3.401	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.251	3.376	3.552	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.231	3.355	3.527	3.819
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.214	3.335	3.505	3.792
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.198	3.318	3.485	3.768
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.183	3.302	3.467	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.170	3.287	3.450	3.725
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.158	3.274	3.435	3.707
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.147	3.261	3.421	3.690
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.047	3.136	3.250	3.408	3.674
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.038	3.127	3.239	3.396	3.659
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.118	3.230	3.385	3.646
31	0.256	0.682	1.309	1.696	2.040	2.453	2.744	3.022	3.109	3.221	3.375	3.633
32	0.255	0.682	1.309	1.694	2.037	2.449	2.738	3.015	3.102	3.212	3.365	3.622
33	0.255	0.682	1.308	1.692	2.035	2.445	2.733	3.008	3.094	3.204	3.356	3.611
34	0.255	0.682	1.307	1.691	2.032	2.441	2.728	3.002	3.088	3.197	3.348	3.601
35	0.255	0.682	1.306	1.690	2.030	2.438	2.724	2.996	3.081	3.190	3.340	3.591
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.055	3.160	3.307	3.551
60	0.254	0.679	1.296	1.671	2.000	2.390	2.660	2.915	2.994	3.094	3.232	3.460
∞	0.253	0.675	1.282	1.645	1.961	2.327	2.577	2.809	2.880	2.970	3.092	3.293



Puntuaciones *z* POSITIVAS

TABLA A-2 (*continuación*) Área acumulativa desde la IZQUIERDA

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495 *	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949 *	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.50 y mayores	.9999									

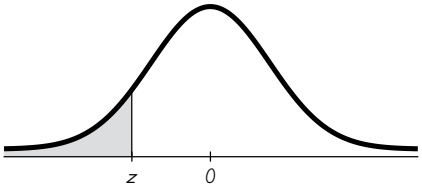
NOTA: Para valores de *z* por encima de 3.49, utilice 0.9999 para el área.

*Utilice estos valores comunes que resultan por interpolación:

Puntuación *z*	Área
1.645	0.9500
2.575	0.9950

Valores críticos comunes

Nivel de confianza	Valor crítico
0.90	1.645
0.95	1.96
0.99	2.575



Puntuaciones *z* NEGATIVAS

TABLA A-2 Distribución normal estándar (*z*): Área acumulativa desde la IZQUIERDA

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.50 y menores	.0001									
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	* .0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	* .0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

NOTA: Para valores de *z* por debajo de −3.49, utilice 0.0001 para el área.

*Utilice estos valores comunes que resultan por interpolación:

Puntuación *z*	Área
−1.645	0.0500
−2.575	0.0050

Formulario

INFERENCIA ESTADÍSTICA

DISTRIBUCIÓN DE LA MEDIA ARITMÉTICA EN EL MUESTREO

Para calcular el
Número de muestras posibles de tamaño n sin reposición

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

DONDE

$\binom{N}{n}$: Núm. de muestras posibles

N: Tamaño de la población

n: Tamaño de la muestra

Para calcular el
Número de muestras posibles de tamaño n con reposición

$$N^n$$

DONDE

N: Tamaño de la población
n: Tamaño de la muestra

Teorema

$$\mu = \mu_{\bar{x}}$$

DONDE

μ: Media poblacional
$\mu_{\bar{x}}$: Media de la distribución muestral (media de medias)

Para calcular la
Media de la distribución muestral

$$\mu_{\bar{x}} = \frac{\sum_{i=1}^{n} \bar{x}}{\binom{N}{n}}$$

DONDE

$\mu_{\bar{x}}$: Media de la distribución muestral
$\bar{x}$: Media muestral
$\binom{N}{n}$: Núm. de muestras posibles

Para calcular el
Error estándar de la media en la distribución

$$\sigma_{\bar{x}} = \sqrt{\frac{\sum_{i=1}^{n} \left(\bar{x} - \mu_{\bar{x}}\right)^2}{\binom{N}{n}}}$$

DONDE

$\sigma_{\bar{x}}$: Error estándar para la media muestral
$\bar{x}$: Media muestral
$\binom{N}{n}$: Núm. de muestras posibles
$\mu_{\bar{x}}$: Media de la distribución muestral

Teorema ***Para calcular el error estándar para la media (población infinita)*** $$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$	**DONDE** $\sigma_{\bar{x}}$: Error estándar para la media σ: Desviación estándar poblacional n: Tamaño de la muestra
Teorema ***Para calcular el error estándar para la media (población finita)*** $$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$	**DONDE** $\sigma_{\bar{x}}$: Error estándar para la media σ: Desviación estándar poblacional n: Tamaño de la muestra N: Tamaño de la población
Factor de corrección para poblaciones finitas $$\sqrt{\frac{N-n}{N-1}}$$	**DONDE** N: Tamaño de la población n: Tamaño de la muestra

INFERENCIA ESTADÍSTICA

DISTRIBUCIÓN DE LA PROPORCIÓN EN EL MUESTREO

Teorema

$$P = P_{\bar{p}} \quad \cong \quad \pi = P_{\bar{p}}$$

DONDE

$P = \pi$: Proporción poblacional de éxitos
$P_{\bar{p}}$: Proporción de la distribución muestral

Para calcular la
Proporción de la distribución muestral

$$P_{\bar{p}} = \frac{\sum_{i=1}^{n} \bar{p}}{\binom{N}{n}}$$

DONDE

$P_{\bar{p}}$: Proporción de la distribución muestral
$\bar{p}$: Proporción muestral de éxitos
$\binom{N}{n}$: Núm. de muestras posibles

Para calcular el
Error estándar para la proporción en la distribución muestral

$$\sigma_{\bar{p}} = \sqrt{\frac{\sum_{i=1}^{n} (\bar{p} - P_{\bar{p}})^2}{\binom{N}{n}}}$$

DONDE

$\sigma_{\bar{p}}$: Error estándar para la proporción de la distribución muestral
$\bar{p}$: Proporción muestral de éxito
$P_{\bar{p}}$: Proporción de la distribución muestral
$\binom{N}{n}$: Núm. de muestras posibles

Teorema

Para calcular el error estándar para la proporción (población infinita)

$$\sigma_{\bar{p}} = \sqrt{\frac{P \cdot Q}{n}} \quad \cong \quad \sigma_{\bar{p}} = \sqrt{\frac{\pi(1 - \pi)}{n}}$$

$$P = \pi$$

$$Q = 1 - \pi$$

DONDE

$\sigma_{\bar{p}}$: Error estándar para la proporción
P: Proporción poblacional de éxito
Q: Proporción poblacional de fracaso
n: Tamaño de la muestra

Teorema ***Para calcular el error estándar para la proporción (población finita)*** $$\sigma_{\bar{p}} = \sqrt{\frac{P \cdot Q}{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$ $$\sigma_{\bar{p}} = \sqrt{\frac{\pi(1-\pi)}{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$	**DONDE** $\sigma_{\bar{p}}$: Error estándar para la proporción P: Proporción poblacional de éxito Q: Proporción poblacional de fracaso n: Tamaño de la muestra N: Tamaño de la población
Fórmula para ***Determinar la distancia de la proporción muestral $\bar{p}$ a partir de la proporción poblacional*** $$z = \frac{\left(\bar{p} \pm \frac{1}{2n}\right) - P_{\bar{p}}}{\sigma_{\bar{p}}}$$	**DONDE** z: Núm. de desviaciones estándar de $\bar{p}$ respecto a la proporción de la distribución muestral $\bar{p}$: Proporción muestral de éxito $P_{\bar{p}}$: Proporción de la distribución muestral n: Tamaño de la muestra $\sigma_{\bar{p}}$: Error estándar para la proporción
Factor de corrección de continuidad $$\frac{1}{2n}$$	**DONDE** n: Tamaño de la muestra

INFERENCIA ESTADÍSTICA

ESTIMACIÓN

Para calcular el
Error en la estimación para la media cuando se conoce σ

$$E = z_{\alpha/2} \cdot \sigma_{\bar{x}}$$

DONDE

E: Error en la estimación
α: Nivel de significancia
$\sigma_{\bar{x}}$: Error estándar para la media
z: Puntuación normal estándar z

Para calcular el
Error en la estimación para la media cuando no se conoce σ
Si $n > 30$

$$E = z_{\alpha/2} \cdot S_{\bar{x}}$$

DONDE

n: Tamaño de la muestra
E: Error en la estimación
z: Puntuación normal estándar z
α: Nivel de significancia
$S_{\bar{x}}$: Error estándar para la media

Para calcular el
Error en la estimación para la media cuando no se conoce σ
Si $n \leq 30$

$$E = t_{\alpha/2} \cdot S_{\bar{x}}$$

DONDE

n: Tamaño de la muestra
E: Error en la estimación
α: Nivel de significancia
$S_{\bar{x}}$: Error estándar para la media
t: Puntuación "t" de student

Para determinar los grados de libertad para la distribución "t" de student

$$g.l. = n - 1$$

DONDE

$g.l.$: Grados de libertad
n: Tamaño de la muestra

Para calcular el
Error estándar para la media cuando se desconoce σ (población infinita)

$$S_{\bar{x}} = \frac{S}{\sqrt{n}}$$

DONDE

$S_{\bar{x}}$: Error estándar para la media
S: Desviación estándar muestral
n: Tamaño de la muestra

Para calcular el
Error estándar para la media cuando se desconoce σ (población finita)

$$S_{\bar{x}} = \frac{S}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

DONDE

$S_{\bar{x}}$: Error estándar para la media
S: Desviación estándar muestral
n: Tamaño de la muestra
N: Tamaño de la población

Para calcular los
Intervalos de confianza para la media

$$\bar{x} - E \leq \mu \leq \bar{x} + E$$

DONDE

$\bar{x}$: Media muestral
E: Error en la estimación
μ: Media poblacional

Para calcular el
Error en la estimación para la proporción

$$E = z_{\alpha/2} \cdot \sigma_{\bar{p}}$$

DONDE

E: Error en la estimación
z: Puntuación normal estándar
α: Nivel de significancia
$\sigma_{\bar{p}}$: Error estándar para la proporción

Para calcular
Intervalos de confianza para la proporción

$$\bar{p} - E \leq \pi \leq \bar{p} + E$$

$$\bar{p} - E \leq P \leq \bar{p} + E$$

DONDE

$\bar{p}$: Proporción muestral de éxito
E: Error en la estimación
π: Proporción poblacional
x: Número de éxitos en la muestra
n: Tamaño de la muestra

Para calcular la
Proporción muestral de éxitos

$$\bar{p} = \frac{x}{n}$$

DONDE

$\bar{p}$: Proporción muestral de éxito
x: Núm. de éxitos en la muestra
n: Tamaño de la muestra

Para calcular el
Error estándar de la proporción (población infinita)

$$\sigma_{\bar{p}} = \sqrt{\frac{\bar{p} \cdot \bar{q}}{n}}$$

DONDE

$\sigma_{\bar{p}}$: Error estándar para la proporción
$\bar{p}$: Proporción muestral de éxito
$\bar{q}$: Proporción muestral de fracaso
n: Tamaño de la muestra

Para calcular el

Error estándar de la proporción (población finita)

$$\sigma_{\bar{p}} = \sqrt{\frac{\bar{p} \cdot \bar{q}}{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

DONDE

$\sigma_{\bar{p}}$: Error estándar para la proporción
$\bar{p}$: Proporción muestral de éxito
$\bar{q}$: Proporción muestral de fracaso
n: Tamaño de la muestra
N: Tamaño de la población

NOTAS

INFERENCIA ESTADÍSTICA

CÁLCULO DEL TAMAÑO DE MUESTRA ADECUADO

Para calcular el ***Tamaño de la muestra adecuado para la media (población infinita)*** $$n = \left(\frac{\sigma \cdot z_{\alpha/2}}{E}\right)^2$$	**DONDE** n: Tamaño de la muestra σ: Desviación estándar poblacional z: Puntuación normal estándar E: Error en la estimación α: Nivel de significancia
Para calcular el ***Tamaño de la muestra adecuado para la media (población finita)*** $$n = \frac{z_{\alpha/2}^2 \cdot \sigma^2 \cdot N}{(N-1) \cdot E^2 + z_{\alpha/2}^2 \cdot \sigma^2}$$	**DONDE** n: Tamaño de la muestra σ: Desviación estándar poblacional z: Puntuación normal estándar E: Error en la estimación N: Tamaño de la población α: Nivel de significancia
Para calcular el ***Tamaño de la muestra para proporciones (población infinita)*** $$n = \frac{P \cdot Q \cdot Z_{\alpha/2}^2}{E^2}$$	**DONDE** n: Tamaño de la muestra P: Proporción poblacional de éxito Q: Proporción poblacional de fracaso z: Puntuación normal estándar E: Error en la estimación α: Nivel de significancia
Para calcular el ***Tamaño de la muestra para proporciones (población finita)*** $$n = \frac{z_{a/2}^2 \cdot P \cdot Q \cdot N}{(N-1) \cdot E^2 + z_{a/2}^2 P \cdot Q}$$	**DONDE** n: Tamaño de la muestra P: Proporción poblacional de éxito Q: Proporción poblacional de fracaso z: Puntuación normal estándar E: Error en la estimación N: Tamaño de la población α: Nivel de significancia

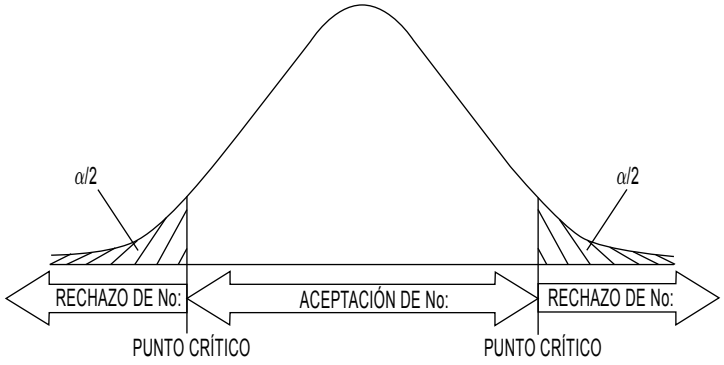
INFERENCIA ESTADÍSTICA

PRUEBA DE HIPÓTESIS PARA LA MEDIA

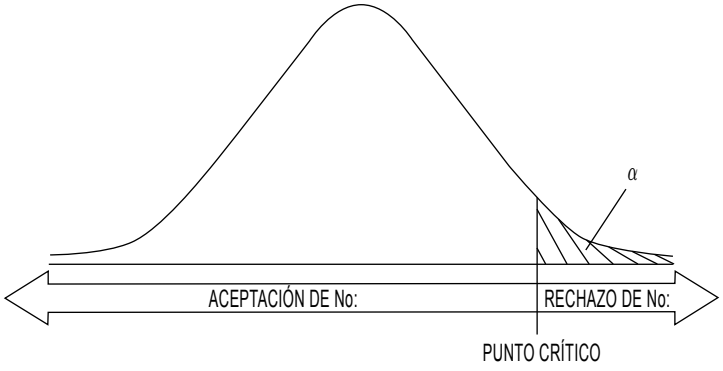
Prueba bilateral $H_0: \mu = 0$ $H_1: \mu \neq 0$	**DONDE** H_0: Hipótesis nula μ: Media poblacional H_1: Hipótesis alternativa
Prueba es unilateral por la izquierda $H_0: \mu \geq 0$ $H_1: \mu < 0$	**DONDE** H_0: Hipótesis nula μ: Media poblacional H_1: Hipótesis alternativa
Prueba es unilateral por la derecha $H_0: \mu \leq 0$ $H_1: \mu > 0$	**DONDE** H_0: Hipótesis nula μ: Media poblacional H_1: Hipótesis alternativa

GRÁFICAS: ÁREAS DE ACEPTACIÓN Y RECHAZO SEGÚN EL TIPO DE PRUEBA

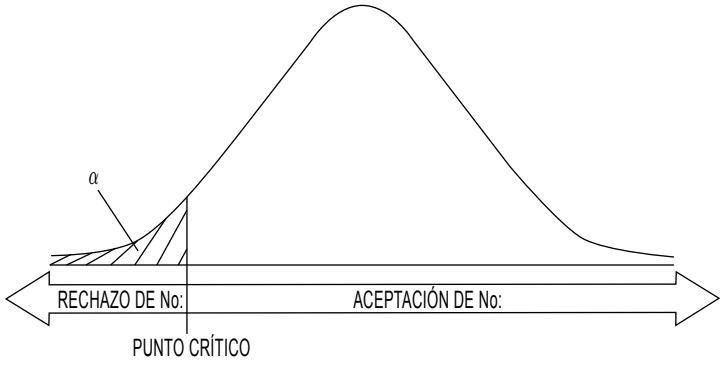
BILATERAL O (DE DOS COLAS)



UNILATERAL O (DE UNA COLA) A LA DERECHA



UNILATERAL O (DE UNA COLA) A LA IZQUIERDA



INFERENCIA ESTADÍSTICA

PRUEBAS ESTADÍSTICAS ADECUADAS

Pruebas estadística cuando se conoce σ sin importar el tamaño de la muestra

$$z_c = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

DONDE

z_c: z calculada
$\bar{x}$: Media muestral
μ: Media poblacional
$\sigma_{\bar{x}}$: Error estándar para la media

Pruebas estadísticas cuando se desconoce σ
Si $n >$ **30**

$$z_c = \frac{\bar{x} - \mu}{S_{\bar{x}}}$$

DONDE

n: Tamaño de la muestra
z_c: z calculada
$\bar{x}$: Media muestral
μ: Media poblacional
$S_{\bar{x}}$: Error estándar para la media

Para pruebas estadísticas cuando se desconoce σ
Si $n \leq$ **30**

$$t_c = \frac{\bar{x} - \mu}{S_{\bar{x}}}$$

DONDE

n: Tamaño de la muestra
t_c: t calculada
$\bar{x}$: Media muestral
μ: Media poblacional
$S_{\bar{x}}$: Error estándar para la media

INFERENCIA ESTADÍSTICA

PRUEBA DE HIPÓTESIS PARA LA PROPORCIÓN

Prueba bilateral $H_0: \pi = 0$ $H_1: \pi \neq 0$	**DONDE** H_0: Hipótesis nula H_1: Hipótesis alternativa π: Proporción poblacional
Prueba unilateral por la izquierda $H_0: \pi = 0$ $H_1: \pi < 0$	**DONDE** H_0: Hipótesis nula H_1: Hipótesis alternativa π: Proporción poblacional
Prueba unilateral por la derecha $H_0: \pi = 0$ $H_1: \pi > 0$	**DONDE** H_0: Hipótesis nula H_1: Hipótesis alternativa π: Proporción poblacional
Para calcular el ***Valor de z_c*** ***(Prueba Estadística Adecuada)*** $z_c = \dfrac{\overline{p} - \pi}{\sigma_{\overline{p}}}$	**DONDE** z_c: z calculada $\overline{p}$: Proporción muestral de éxito π: Proporción poblacional de éxito $\sigma_{\overline{p}}$: Error estándar para la proporción

Para calcular el
Error estándar para la proporción

Población infinita

$$\sigma_{\bar{p}} = \sqrt{\frac{\pi(1-\pi)}{n}}$$

Población finita

$$\sigma_{\bar{p}} = \sqrt{\frac{\pi(1-\pi)}{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

DONDE

$\sigma_{\bar{p}}$: Error estándar para la proporción
π: Proporción poblacional de éxito
$1-\pi$: Proporción poblacional de fracaso
n: Tamaño de la muestra
N: Tamaño de la población

NOTAS

INFERENCIA ESTADÍSTICA

PRUEBA DE HIPÓTESIS PARA LA DIFERENCIA DE DOS MEDIAS

Prueba es bilateral $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$	**DONDE** H_0: Hipótesis nula H_1: Hipótesis alternativa μ_1: Media 1 μ_2: Media 2
Prueba unilateral por la derecha $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$	**DONDE** H_0: Hipótesis nula H_1: Hipótesis alternativa μ_1: Media 1 μ_2: Media 2
Prueba unilateral por la izquierda $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 < \mu_2$	**DONDE** H_0: Hipótesis nula H_1: Hipótesis alternativa μ_1: Media 1 μ_2: Media 2
Prueba estadística cuando σ es conocida sin importar el tamaño de las muestras $z_c = \dfrac{\bar{x}_1 - \bar{x}_2}{\sigma_{\bar{x}_1 - \bar{x}_2}}$	**DONDE** z_c: z calculada $\bar{x}_1$: Media muestral de la población 1 $\bar{x}_2$: Media muestral de la población 2 $\sigma_{\bar{x}_1 - \bar{x}_2}$: Error estándar para la diferencia entre dos medias
Para calcular el ***Error estándar para la diferencia entre dos medias cuando σ es conocida*** $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\dfrac{\sigma_2^1}{n_1} + \dfrac{\sigma_2^2}{n_2}}$	**DONDE** $\sigma_{\bar{x}_1 - \bar{x}_2}$: Error estándar para la diferencia entre dos medias σ_1^2: Varianza de la población 1 σ_2^2: Varianza de la población 2 n_1: Tamaño de la muestra 1 n_2: Tamaño de la muestra 2
Prueba estadística cuando σ no se conoce y $n_1 + n_2 > 30$ $z_c = \dfrac{\bar{x}_1 - \bar{x}_2}{S_{\bar{x}_1 - \bar{x}_2}}$	**DONDE** z_c: z calculada $\bar{x}_1$: Media muestral de la población 1 $\bar{x}_2$: Media muestral de la población 2 $S_{\bar{x}_1 - \bar{x}_2}$: Error estándar para la diferencia entre dos medias

Para calcular el

Error estándar para la diferencia entre dos medias cuando σ es desconocida y n > 30

$$S_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{S_2^1}{n_1} + \frac{S_2^2}{n_2}}$$

DONDE

$S_{\bar{x}_1 - \bar{x}_2}$: Error estándar para la diferencia entre dos medias
S_1^2: Varianza de la población 1
S_2^2: Varianza de la población 2
n_1: Tamaño de la muestra 1
n_2: Tamaño de la muestra 2

Prueba estadística cuando se desconoce σ y $n_1 + n_2 \leq 30$

$$t_c = \frac{\bar{x}_1 - \bar{x}_2}{S_{\bar{x}_1 - \bar{x}_2}}$$

DONDE

t_c: t calculada
$\bar{x}_1$: Media muestral de la población 1
$\bar{x}_2$: Media muestral de la población 2
$S_{\bar{x}_1 - \bar{x}_2}$: Error estándar para la diferencia entre dos medias

---

Para calcular el

Error estándar para la diferencia entre dos medias cuando se desconoce σ y n ≤ 30

$$S_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

(muestreo independiente)

DONDE

$S_{\bar{x}_1 - \bar{x}_2}$: Error estándar para la diferencia entre dos medias
S_1^2: Varianza de la población 1
S_2^2: Varianza de la población 2
n_1: Tamaño de la muestra 1
n_2: Tamaño de la muestra 2

Para calcular los

Grados de libertad para la distribución "t" de student cuando el muestreo es independiente

$$g.l. = n_1 + n_2 - 2$$

DONDE

$g.l.$: Grados de libertad
n_1: Tamaño de la muestra 1
n_2: Tamaño de la muestra 2

NOTAS

INFERENCIA ESTADÍSTICA

PRUEBA DE HIPÓTESIS PARA DOS PROPORCIONES

Prueba bilateral $H_0: \pi_1 = \pi_2$ $H_1: \pi_1 \neq \pi_2$	**DONDE** H_0: Hipótesis nula H_1: Hipótesis alternativa π_1: Proporción de la población 1 π_2: Proporción de la población 2
Prueba unilateral por la izquierda $H_0: \pi_1 = \pi_2$ $H_1: \pi_1 < \pi_2$	**DONDE** H_0: Hipótesis nula H_1: Hipótesis alternativa π_1: Proporción de la población 1 π_2: Proporción de la población 2
Prueba unilateral por la derecha $H_0: \pi_1 = \pi_2$ $H_1: \pi_1 > \pi_2$	**DONDE** H_0: Hipótesis nula H_1: Hipótesis alternativa π_1: Proporción de la población 1 π_2: Proporción de la población 2
Para calcular z_c ***(Prueba estadística adecuada)*** $z_c = \dfrac{\overline{P}_1 - \overline{P}_2}{\sigma_{\overline{p}_1 - \overline{p}_2}}$	**DONDE** z_c: z calculada $\overline{p}_1$: Proporción de éxito de la muestra 1 $\overline{p}_2$: Proporción de éxito de la muestra 2 $\sigma_{\overline{p}_1 - \overline{p}_2}$: Error estándar combinado para la diferencia entre dos proporciones

Para calcular el

Error estándar combinado para la diferencia entre dos proporciones

$$\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{\overline{p}_1 \bullet \overline{q}_1}{n_1} + \frac{\overline{p}_2 \bullet \overline{q}_2}{n_2}}$$

DONDE

$\sigma_{\bar{p}_1 - \bar{p}_2}$: Error estándar combinado para la diferencia entre dos proporciones

$\overline{p}_1$: Probabilidad de éxito de la muestra 1

$\overline{p}_2$: Probabilidad de éxito de la muestra 2

$\overline{q}_1$: Probabilidad de fracaso de la muestra 1

$\overline{q}_2$: Probabilidad de fracaso de la muestra 2

n_1: Tamaño de la muestra 1

n_2: Tamaño de la muestra 2

NOTAS

INFERENCIA ESTADÍSTICA

PRUEBA DE JI CUADRADA
χ^2

(PRUEBAS DE INDEPENDENCIA)
Para realizar la prueba estadística

$$\chi_c^2 = \sum_{i=1}^{n} \frac{(F.O. - F.E)^2}{F.E}$$

DONDE

χ_c^2: Ji cuadrada calculada
F.O.: Frecuencia observada
F.E.: Frecuencia esperada

Para el cálculo de la frecuencia esperada en tablas de contingencia

$$F.E. = \frac{(Total\ de\ columna)(Total\ de\ renglón)}{Gran\ total\ o\ n}$$

Para el cálculo de grados de libertad en tablas de contingencia

$$g.l. = (\#\ de\ renglones - 1)\ (\#\ de\ columnas - 1)$$

NOTAS

INFERENCIA ESTADÍSTICA

REGRESIÓN Y CORRELACIÓN LINEAL SIMPLE

Para determinar la
Ecuación de mejor ajuste (método de mínimos cuadrados)

$$\hat{y} = a + bx$$

DONDE

$\hat{y}$: y estimada
a: Ordenada al origen
b: Pendiente muestral
x: Variable independiente
(a y b): son constantes o coeficientes de regresión

Para calcular
La pendiente

$$b = \frac{\sum x\,y - n\,\bar{x}\,\bar{y}}{\sum x^2 - n\,\bar{x}^2}$$

DONDE

b: Pendiente
x: Variable independiente
y: Variable dependiente
$\bar{x}$: Media de la variable independiente
$\bar{y}$: Media de la variable dependiente
n: Tamaño de la muestra

Para calcular la
Ordenada al origen

$$a = \bar{y} - b\bar{x}$$

DONDE

a: Ordenada al origen
$\bar{y}$: Media de la variable dependiente
b: Pendiente muestral
$\bar{x}$: Media de la variable independiente

Para calcular el
Error estándar de estimación

$$S_e = \sqrt{\frac{\sum y^2 - (a \cdot \sum y) - (b \cdot \sum xy)}{n - 2}}$$

DONDE

S_e: Error estándar de estimación
x: Variable independiente
y: Variable dependiente
n: Tamaño de la muestra
a: Ordenada al origen
b: Pendiente muestral

Para calcular el
Coeficiente muestral de determinación

$$r^2 = \frac{(a \cdot \sum y) + (b \cdot \sum xy) - n \cdot \bar{y}^2)}{\sum y^2 - (n \cdot \bar{y}^2)}$$

DONDE

r^2: Coeficiente muestral de determinación
a: Ordenada al origen
b: Pendiente
x: Variable independiente
y: Variable dependiente
$\bar{y}$: Media de la variable dependiente

Para calcular el
Coeficiente de correlación

$$r = \sqrt{r^2}$$

DONDE

r: Coeficiente de correlación
r^2: Coeficiente muestral de determinación

Para calcular el
Error estándar del coeficiente de regresión

$$S_b = \frac{S_e}{\sqrt{\sum x^2 - n(\bar{x})^2}}$$

DONDE

S_b: Error estándar del coeficiente de regresión
S_e: Error estándar de estimación
n: Tamaño de muestra
x: Variable independiente
$\bar{x}$: Media de la variable independiente