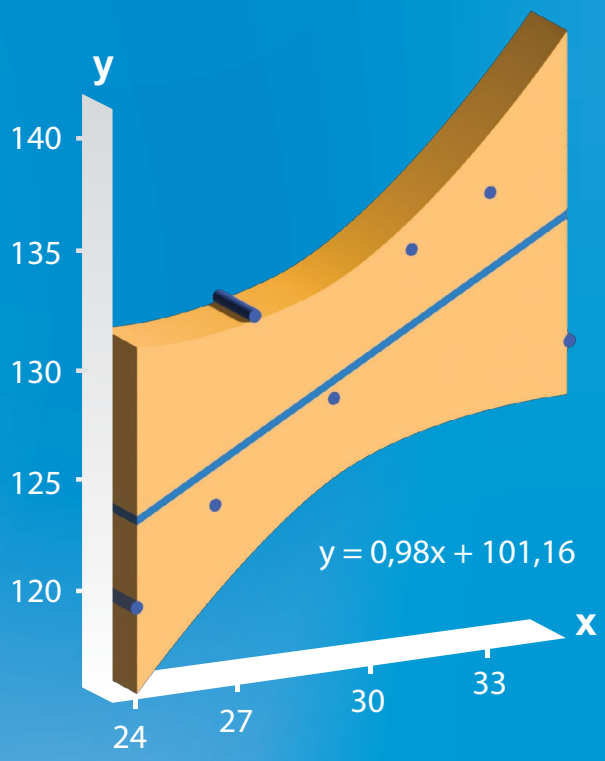
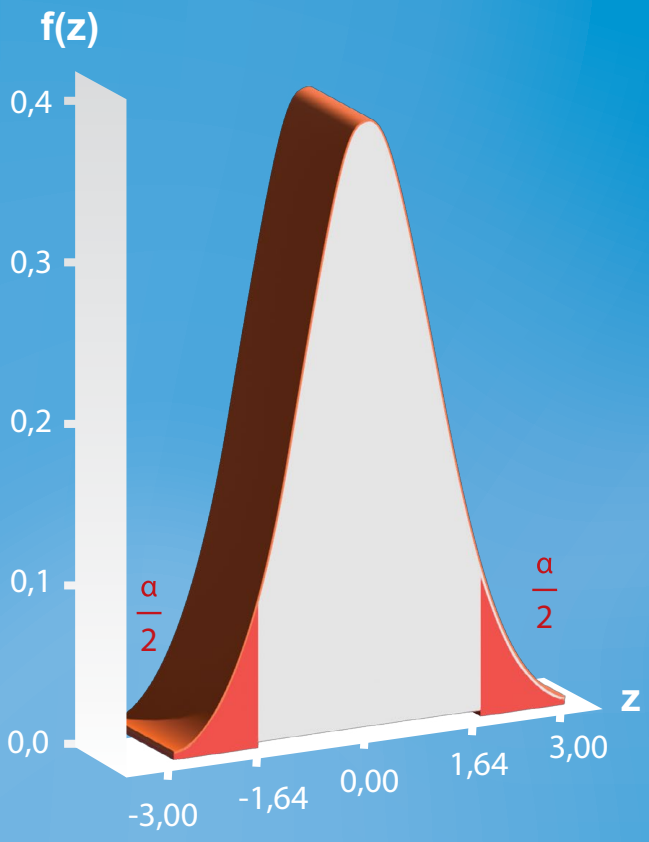
ELEMENTOS DE INFERENCIA ESTADÍSTICA CON R

José Andrey Zamora Araya

Eduardo Aguilar Fernández





euna

UNA
UNIVERSIDAD NACIONAL
COSTA RICA

ELEMENTOS DE INFERENCIA ESTADÍSTICA CON R

José Andrey Zamora Araya

Eduardo Aguilar Fernández

ELEMENTOS DE INFERENCIA ESTADÍSTICA CON R

Editorial Universidad Nacional
Heredia, Campus Omar Dengo
Costa Rica
Teléfono: 2562-6754 Fax: 2562-6761
Correo electrónico: euna@una.cr
Apartado postal: 86-3000 (Heredia, Costa Rica)

La Editorial Universidad Nacional (EUNA), es miembro del
Sistema Editorial Universitario Centroamericano (SEDUCA).

Producción editorial: Marianela Camacho Alfaro
marianela.camacho.alfaro@una.cr
Diseño de cubierta: Programa de Publicaciones e Impresiones de la UNA.

519.54
Z25e
Zamora Araya José Andrey
Elementos de inferencia estadística con R / José Andrey Zamora Araya, Eduardo Aguilar Fernández. - Primera edición. - Heredia, Costa Rica : EUNA, 2022.
1 recurso en línea (189 páginas) : archivo, de texto, PDF
ISBN 978-9977-65-647-2
1. INFERENCIA ESTADÍSTICA 2. MUESTREO 3. ESTIMACIÓN 4 HIPÓTESIS 5 MODELOS I. Aguilar Fernández, Eduardo II. Título

A Ana Maricela e Irene Isabel, quienes me motivan cada día. A Cecilia y Teodorico, con especial cariño.

Eduardo Aguilar Fernández

A mi esposa Rosibel y mi hija Natasha, las amo, sin ustedes nada de lo que hago sería posible.

José Andrey Zamora Araya

Contenido

Índice de figuras

Índice de tablas

Prefacio

El presente documento tiene por objetivo mostrar el uso de comandos básicos de R para el abordaje de temas relacionados con la Estadística Inferencial. El material está desarrollado con fin de brindar apoyo a estudiantes y docentes en el manejo de temáticas relacionadas con la disciplina. Además, puede ser de utilidad para otras personas que requieran de textos de consulta sobre el conocimiento de determinados aspectos que comprende la Estadística Inferencial.

Para una mejor comprensión del texto es importante tener conocimientos básicos de R relacionados con el uso de paquetes, vectores, funciones, bases de datos, cálculo de medidas estadísticas, así como en la utilización de operadores matemáticos. Ejemplos del uso de estos objetos pueden observarse en Aguilar y Zamora (2020). También es necesario tener conocimiento de temas de Estadística Descriptiva así como de conceptos, axiomas y teoremas fundamentales de la teoría de Probabilidad Clásica.

La obra está constituida por seis capítulos. El primero de ellos trata sobre las distribuciones de muestreo para la media y la proporción, de tal forma que la persona lectora reconoce los principales elementos teóricos que fundamentan esta temática y se fomenta el manejo de comandos básicos para realizar operaciones habituales relacionadas con el cálculo de probabilidades en general.

En el segundo capítulo se desarrollan las bases para la estimación estadística cuya aplicación práctica se visualiza en el tercer capítulo, donde se abordan aspectos relacionados con la estimación de intervalos confianza para uno o dos parámetros.

El cuarto capítulo explica los principales aspectos sobre la teoría de la decisión estadística o prueba de hipótesis sobre un parámetro y el cálculo de los distintos estadísticos de prueba que requieren los análisis. El quinto capítulo trata aspectos de las pruebas de hipótesis aplicados a inferencias relacionadas con dos parámetros. En ambos capítulos se muestran representaciones gráficas de la zona de rechazo y no rechazo definidas para un contraste, además se explica con detalle diferentes argumentos de las funciones utilizadas para la realización de las distintas pruebas.

El sexto capítulo muestra las principales funciones que permiten medir el nivel de asociación y modelar la relación lineal entre dos variables estadísticas. También se describen funciones que permiten la representación gráfica de la relación, así como la comprobación de diferentes supuestos que deben cumplirse para la determinación del modelo.

Se espera que la estructura del documento permita a la persona lectora la comprensión de las diferentes temáticas así como de las herramientas que el software proporciona. Se plantearon diferentes ejemplos en cada sección, los cuales se desarrollan con detalle de modo que permitan tener una idea clara del uso de las funciones y sus respectivos argumentos. Además de los ejemplos resueltos, se dejan a la persona lectora ejercicios que le permitirán reforzar su aprendizaje sobre el uso de comandos. Para su elaboración se empleó la versión 4.0.4 (R Core Team, 2021) y la interface gráfica RStudio, versión 1.1.383 (RStudio Team, 2015). Es importante mencionar que esta obra viene a complementar el material elaborado por Aguilar y Zamora (2020), el cual es de libre acceso y muestra el uso del software R en temas propios de la Estadística descriptiva.

Asimismo, el documento no pretende incorporar la amplia variedad de opciones que brinda el entorno R para la aplicación de diferentes técnicas de inferencia estadística, pues su objetivo principal es brindar apoyo a aquellas personas que desean obtener conocimiento del uso básico de R como herramienta para el cálculo de diferentes medidas estadísticas que sustentan la realización de inferencias.

Finalmente, se desea agradecer a todas las personas que hicieron posible la elaboración de este documento, en especial a estudiantes de la carrera de Bachillerato y Licenciatura en Enseñanza de la Matemática de la Universidad Nacional, pues sus observaciones han sido de gran ayuda para la retroalimentación del material.

Los autores

Enero, 2022

Heredia, Costa Rica

Capítulo 1

Distribuciones de muestreo

Cuando se hace un estudio, las personas investigadoras realizan análisis estadísticos, con el fin de conocer las características de interés de la población en estudio. Si la población es pequeña, y es viable recolectar información de todas las unidades estadísticas, lo más apropiado es realizar una enumeración total.

No obstante, en muchas ocasiones, no es posible o viable realizar una enumeración total, en cuyo caso la persona investigadora dispone de varios métodos para obtener información sobre las características de la población de interés, entre ellos están la simulación, el diseño de experimentos y el muestreo (Ugarte, Militino, y Arnholt, 2008).

El muestreo es la forma más usual de recopilar información sobre la población, debido a que economiza tiempo y dinero. Además, si se realiza de manera apropiada, puede brindar información precisa acerca de las principales características de la población.

Entre los tipos de muestreo aleatorios que más se utilizan están: el simple al azar, el sistemático, el estratificado y el de conglomerados.

En este primer capítulo se hace referencia a las distribuciones de muestreo y la revisión de algunos conceptos importantes para su desarrollo.

Definición 1.1 Variable aleatoria

Sea $(\Omega, \mathbb{F}, P)$ un espacio de probabilidad. Una función

$$X : \Omega \to \mathbb{R}$$

es una variable aleatoria si se cumple que para cualquier intervalo I en $\mathbb{R}$ en el conjunto $\omega : X(\omega) \in I$ es un evento, es decir, está en $\mathbb{F}$.

Definición 1.2 Función de distribución de una variable aleatoria

Sea $(\Omega, \mathbb{F}, P)$ un espacio de probabilidad y $X : \Omega \to \mathbb{R}$ una variable aleatoria. Se llama función de distribución de la variable aleatoria X a la función F definida por:

$$F(X = x) = \mathbb{P}(\{\omega : X(\omega) \leq x\}) = \mathbb{P}(X \leq x)$$

.

En algunas ocasiones, para resaltar que F es la función de distribución de X, se escribe F_X en lugar de F.

Definición 1.3 Parámetro

Un parámetro θ_X, es una función de la distribución de probabilidad f de una variable aleatoria (va) X. Esto significa que el parámetro θ_X puede denotarse por $t(f)$, donde t denota la función aplicada a f. Cada θ_X es obtenido aplicando algún procedimiento numérico t, a la función de distribución de probabilidad f.

Los parámetros son los que caracterizan a las distribuciones de probabilidad y por lo general son desconocidos. En la estadística clásica los parámetros se consideran fijos, es decir, son tratados como constantes, mientras que en la estadística bayesiana, se tiene un modelo que determina la probabilidad de observar diferentes valores de una variable X, bajo diferentes valores de los parámetros.

Si X es una variable aleatoria, un ejemplo de parámetro para X es la media aritmética de una población finita. La media aritmética de la variable aleatoria X de una población se denota por μ_X, o bien, $E[X]$. De esta manera se tiene que

$$\theta_X = t(f) = E[X] = \mu_X = \int_{-\infty}^{\infty} x f(x) dx$$

si X es una variable continua y f es la función de distribución de X.

Definición 1.4 Estimador

Sean $X_1, X_2, \cdots, X_n$ los valores de una variable X en los n elementos que forman parte de una muestra aleatoria extraída de una población. Un estimador de un parámetro particular de la población es una función de los valores $X_1, X_2, \cdots, X_n$ que se utiliza para aproximar o estimar el valor desconocido del parámetro.

Dado que un estimador es cualquier función de una muestra S, entonces puede denotarse por $T_X = t(\mathbf{S})$. Note que el estimador o estadístico T_X de θ_X, puede también ser denotado por $\hat{\theta}_X$.

Un ejemplo de estimador es la media aritmética de una muestra, esto pues,

$$T_X = t(\mathbf{S}) = \bar{x} = \frac{\sum_{i=1}^{n} X_i}{n}$$

es una función que permite, a partir de una muestra aleatoria $\mathbf{X} = X_1, X_2 \ldots, X_n$, calcular una estimación para el parámetro $\mu_X = E[X]$.

Definición 1.5 Estimación

Se llama estimación a la medida descriptiva obtenida en una muestra aleatoria extraída de una población.

Los estimadores son medidas que cambian de una muestra a otra, es decir, son variables. De esta forma, si $S_1, S_2, \cdots, S_k$ representan k muestras aleatorias del mismo tamaño de una población dada, la media aritmética de cada muestra estaría dada por $\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_k$, medidas que pueden diferir entre sí, dado que las muestras son distintas. Considerando que $\bar{x}$ puede cambiar en virtud de la muestra, entonces la media aritmética es una variable aleatoria y se denota por $\bar{X}$.

Es importante aclarar que estimación y estimador no tienen el mismo significado, ya que un estimador $\hat{\theta}_X$, de un parámetro θ_X es una función de la muestra, mientras que una estimación es un valor numérico de un estimador, obtenido a partir de los datos de la muestra.

1.1. Distribuciones de muestreo

Definición 1.6 Distribución de muestreo

Se llama distribución de muestreo de un estadístico a la **distribución de probabilidad** de los valores posibles del estadístico que resulta cuando se extraen repetidamente de la población muestras de tamaño n.

Agunos ejemplos de distribuciones de muestreo son:

1. Distribución del promedio muestral.
2. Distribución de la variancia muestral.
3. Distribución de la desviación estándar muestral.
4. Distribución de la proporción muestral.

Definición 1.7 Valor esperado de un estimador

El valor esperado de la variable aleatoria $\hat{\theta}$ se denota por $E[\hat{\theta}]$ y se define, en caso que $\hat{\theta}$ sea una variable aleatoria discreta, por

$$E[\hat{\theta}] = \sum_{i=1}^{k} \hat{\theta}_i p(\hat{\theta}_i)$$

En caso que $\hat{\theta}$ sea una variable aleatoria continua, se define por

$$E[\hat{\theta}] = \int_{-\infty}^{\infty} \hat{\theta} f(\hat{\theta}) d\hat{\theta}$$

1.1.1. Distribución de muestreo para la media

Considere inicialmente que se seleccionan, aleatoriamente, a 15 estudiantes de una población universitaria, con el fin de calcular el promedio de estatura del estudiantado. Si se repitiera este proceso, cinco veces, sería poco probable que el promedio de estas cinco muestras fuera el mismo, pues el promedio puede variar de muestra en muestra.

Sin embargo, los promedios muestrales son usados para estimar el valor desconocido de la media poblacional, ¿qué tan buena es esta aproximación?

Para la valoración de la exactitud de las estimaciones, provenientes de un estadístico, se usa la distribución de probabilidad asociada, con todos los posibles valores que puede tomar el estadístico. Es decir, se considera la distribución de muestreo del estadístico.

Definición 1.8 Distribución muestral para la media

Una distribución de muestreo para la media muestral está dada por la distribución de probabilidad para los valores posibles del estadístico $\bar{x}$ que resulta de extraer repetidamente de la población muestras de tamaño n.

Ejemplo 1.1 Considere una población de tamaño $N = 5$ elementos cuyas observaciones para una variable X determinada están dadas por $\{1, 2, 3, 4, 5\}$. Construya una distribución de muestreo para la media poblacional de la variable X considerando muestras de tamaño $n = 2$ ordenadas y con repeticiones.

Solución

Primeramente, considérese todas las muestras de tamaño $n = 2$ que pueden obtenerse de la población dada.

x	1	2	3	4	5
1	(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)
2	(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)
3	(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)
4	(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)
5	(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)

Luego se determina el promedio de cada muestra.

$\bar{X}$	1	2	3	4	5
1	1,0	1,5	2,0	2,5	3,0
2	1,5	2,0	2,5	3,0	3,5
3	2,0	2,5	3,0	3,5	4,0
4	2,5	3,0	3,5	4,0	4,5
5	3,0	3,5	4,0	4,5	5,0

Para la distribución de probabilidad se toma cada valor promedio y se determina la probabilidad de ocurrencia. En este caso se tiene que:

$\bar{X}$	1,0	1,5	2,0	2,5	3,0	3,5	4,0	4,5	5,0
$\mathbb{P}(\bar{X})$	$\frac{1}{25}$	$\frac{2}{25}$	$\frac{3}{25}$	$\frac{4}{25}$	$\frac{5}{25}$	$\frac{4}{25}$	$\frac{3}{25}$	$\frac{2}{25}$	$\frac{1}{25}$

En el caso de la media muestral, el valor esperado de la variable aleatoria $\bar{X}$ se denota por $E[\bar{X}]$ y se define, en caso que $\bar{X}$ sea una variable aleatoria discreta, por

$$E[\bar{X}] = \sum_{i=1}^{k} \bar{x}_i p(\bar{x}_i)$$

En caso que $\bar{X}$ sea una variable aleatoria continua, está dado por

$$E[\bar{X}] = \int_{-\infty}^{\infty} \bar{x} f(\bar{x}) d\bar{x}$$

Ejemplo 1.2 Determine el valor esperado de la distribución del Ejemplo 1.1.

Solución

$\bar{X}$	1,0	1,5	2,0	2,5	3,0	3,5	4,0	4,5	5,0
$\mathbb{P}(\bar{X})$	$\frac{1}{25}$	$\frac{2}{25}$	$\frac{3}{25}$	$\frac{4}{25}$	$\frac{5}{25}$	$\frac{4}{25}$	$\frac{3}{25}$	$\frac{2}{25}$	$\frac{1}{25}$
$\bar{X}\,\mathbb{P}(\bar{X})$	$\frac{1}{25}$	$\frac{3}{25}$	$\frac{6}{25}$	$\frac{10}{25}$	$\frac{15}{25}$	$\frac{14}{25}$	$\frac{12}{25}$	$\frac{9}{25}$	$\frac{5}{25}$

De esta manera se tiene que $E[\bar{X}] = 3$.

Teorema 1.1 Teorema del límite central. Muestras con reemplazo

Considere todos los valores de una variable aleatoria X con media μ_X y variancia finita σ_X^2. Si de esta población se extraen todas las muestras aleatorias posibles con reemplazo de tamaño n, la distribución de los promedios $\bar{X}$ obtenidos es aproximadamente normal con media denotada por $\mu_{\bar{X}}$ tal que $\mu_{\bar{X}} = \mu_X$ y variancia denotada por $\sigma_{\bar{X}}^2$, donde $\sigma_{\bar{X}}^2 = \dfrac{\sigma_X^2}{n}$, es decir:
Sean $X_1, X_2, \ldots X_n$ variables aleatorias independientes e idénticamente distribuidas (**iid**) con promedio μ_X y variancia σ_X^2, entonces:

1. $E[\bar{X}] = \mu_{\bar{X}} = \mu_X$
2. $Var[\bar{X}] = \sigma_{\bar{X}}^2 = \dfrac{\sigma_X^2}{n}$

Teorema 1.2 Teorema del límite central para muestreo sin reemplazo

Considere todos los valores de una variable aleatoria X con media μ_X y variancia finita σ_X^2. Si de esta población se extraen todas las muestras aleatorias posibles sin reemplazo de tamaño n, la distribución de los promedios $\bar{X}$ obtenidos es aproximadamente normal con media con media denotada por $\mu_{\bar{X}}$ tal que $\mu_{\bar{X}} = \mu_X$ y variancia denotada por $\sigma_{\bar{X}}^2$, donde $\sigma_{\bar{X}}^2 = \dfrac{\sigma_X^2}{n} \cdot \dfrac{N-n}{N-1}$, es decir:
Sean $X_1, X_2, \ldots X_n$ variables aleatorias idénticamente distribuidas con promedio μ_X y variancia σ_X^2, entonces:

1. $E[\bar{X}] = \mu_{\bar{X}} = \mu_X$
2. $Var[\bar{X}] = \dfrac{\sigma_X^2}{n} \cdot \dfrac{N-n}{N-1}$

La raíz cuadrada de la variancia del estimador es la desviación estándar de la variable, también conocida como error estándar del estimador.

Ejemplo 1.3 Verifique los resultados del teorema del límite central con los datos del Ejemplo 1.1.

Ejemplo 1.4 Suponga que $X_1, X_2, \ldots X_n$ son variables aleatorias **iid** y que $X_i \sim Unif(a = 2, b = 5)$ para $1 \leq i \leq n$. Compruebe que $\bar{X} \sim N\left(\mu_{\bar{X}} = \frac{a+b}{2}, \sigma^2_{\bar{X}} = \frac{(b-a)^2}{12n}\right)$.

Solución

Considérese inicialmente una muestra de tamaño 100.

```
> set.seed(1002)
> du <- runif(n = 100, min = 2, max = 5)
```

La Figura 1.1 muestra una simulación de una distribución uniforme, para $n = 100$.

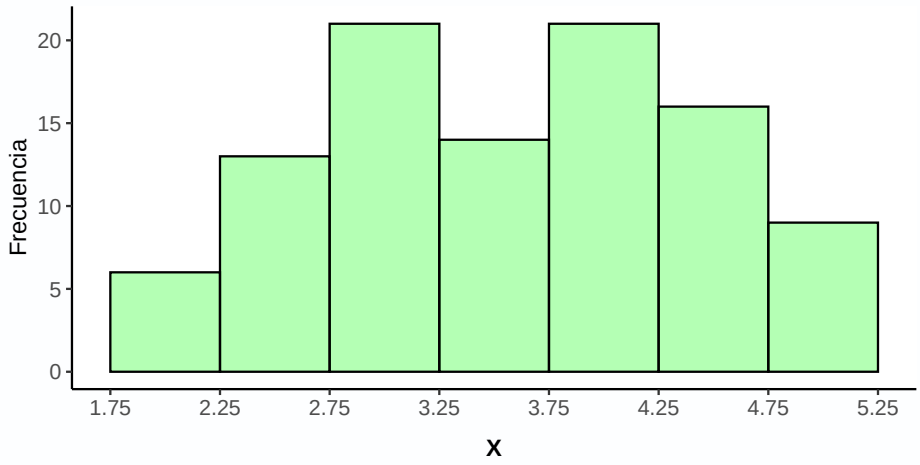


Figura 1.1: Distribución de la variable X, con $X \sim unif(2, 5)$

Fuente: Elaboración propia

Nota: el código de R para obtener la Figura 1.1 puede verse en el Anexo 1.

Ahora se simulan, con ayuda de R, 1000 muestras de tamaño 100, para una variable $X \sim Unif(2, 5)$.

```
> media_u = c()
> set.seed(2534)
> for(i in 1:1000){
+   media_u[i] = mean(runif(n=100, min = 2, max = 5)) }
```

La Figura 1.2 muestra el resultado de la simulación donde puede observarse una distribución aproximadamente simétrica.

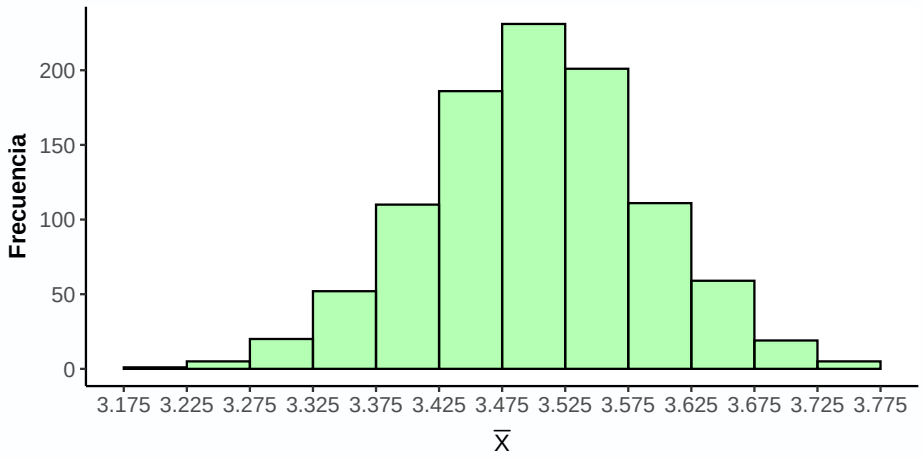


Figura 1.2: Distribución de la media para la variable X, con $X \sim unif(2, 5)$

Fuente: Elaboración propia

Seguidamente se comprueba que los valores de la simulación son aproximadamente iguales a los valores teóricos.

```
> # Valor teórico de la media (2 + 5)/2 = 3.5
> mean(media_u)

[1] 3.502093

> # Valor teórico de la variancia (5-2)^2/(12*100)=0.0075
> var(media_u)

[1] 0.007670457
```

Ejemplo 1.5 Realice una simulación de 1000 muestras de tamaño 100, donde $X_1, X_2, \ldots, X_{100}$ son variables aleatorias **iid** y que $X_i \sim Bin(n = 20,\ p = 0,4)$ para $1 \leq i \leq 100$. Indique los valores teóricos y los obtenidos de la simulación.

Solución

```
> media_bin = c()
> set.seed(1334)
> for(i in 1:1000){media_bin[i] = mean(rbinom(n=100, size = 20, prob = 0.4))}
```

Ahora se comprueban que los valores de la simulación son aproximadamente iguales a los valores teóricos.

```
> # Valor teórico de la media 20(0.4) = 8
> mean(media_bin)

[1] 8.0109

> # Valor teórico de la variancia 20(0.4)(0.6)/100)=0.048
> var(media_bin)

[1] 0.0477089
```

La Figura 1.3 muestra el resultado de la simulación.

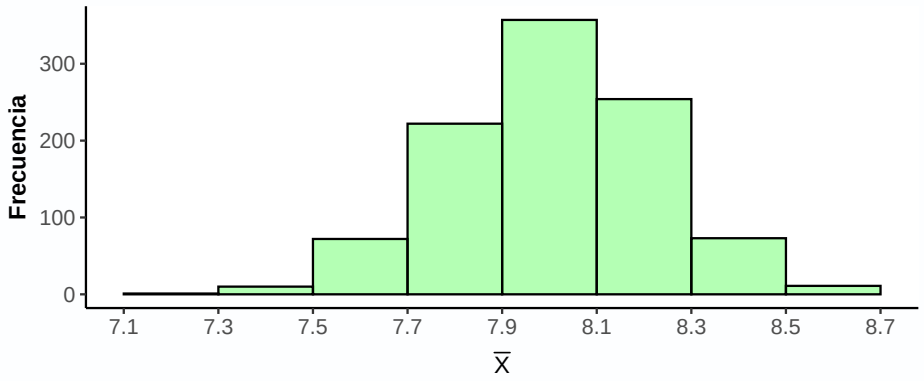


Figura 1.3: Distribución de la media para la variable X, con $X \sim Bin(n = 20,\ p = 0{,}4)$

Fuente: Elaboración propia

Aplicación

Gracias a la ayuda que proporciona el teorema del límite central, mediante una muestra del tamaño apropiado, puede describirse el comportamiento de ciertos estimadores, aprovechando el hecho de que la distribución es aproximadamente normal, lo que implica que pueden calcularse probabilidades. En el caso que no se cuente con disponibilidad de un software estadístico, pueden obtenerse las probabilidades utilizando la tabla de la distribución normal estándar, para ello, es necesario estandarizar la variable. Recuerde que si el estimador se distribuye normalmente, entonces

$$Z = \frac{\text{Estimador - Valor esperado}}{\text{Error estándar}}$$

es una variable aleatoria con distribución normal estándar.

Para el caso de la media, la variable normal estándar queda definida por:

$$Z = \frac{\bar{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} \tag{1.1}$$

A la desviación estándar de un estadístico usado como estimador de un parámetro se le conoce como error estándar del estimador, por ejemplo $\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$ sería el error estándar de la media.

Ahora bien, a que se llama una muestra "grande" o "adecuada". Según Wackerly, Muñoz, y Humberto (2010) algunas recomendaciones para considerar un tamaño de muestra adecuado son las siguientes:

- Si la población muestreada es normal, entonces la distribución muestral de $\bar{X}$ también será normal, sin importar el tamaño de muestra que se elija, es decir, si $X \sim N(\mu_X, \sigma_X^2)$, entonces $\bar{X} \sim N(\mu_X, \frac{\sigma_X^2}{n})$.
- Cuando la población muestreada es aproximadamente simétrica, la distribución muestral de $\bar{X}$ se vuelve aproximadamente normal para valores relativamente pequeños de n.
- Cuando la población muestreada está sesgada, el tamaño de muestra n deber ser más grande, por lo menos $n \geq 30$, antes de que la distribución muestral de $\bar{X}$ se vuelva aproximadamente normal. Por tanto, si $X \sim (\mu_X, \sigma_X^2)$, es decir, no se conoce la distribución de X, entonces (por el teorema del límite central) la distribución límite de $\frac{\bar{X} - \mu_X}{\sigma_X^2/n}$, cuando $n \to \infty$ es la distribución normal estándar.

De acuerdo con Wackerly et al. (2010), una vez que se tiene claro que la distribución muestral para la media es normal o aproximadamente normal, puede describirse el comportamiento de la media muestral $\bar{X}$ al calcular probabilidades para ciertos valores de $\bar{X}$ en el muestreo repetido. En caso de no tener a disposición un software que permita calcular probabilidades para los valores de $\bar{X}$ se recomienda:

- Hallar $\mu_{\bar{X}} = \mu_X$ y $\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$ (error estándar).
- Describa la probabilidad que desea obtener en términos de $\bar{x}$. Puede realizar un dibujo para ayudarse a localizar el área de la región bajo la curva normal.
- Realice el proceso de estandarización de la variable $\bar{X}$ utilizando la Ecuación 1.1.
- Utilice la tabla para la distribución normal estándar para calcular la probabilidad requerida.

Ejemplo 1.6 Una compañía estima que la media poblacional del precio de un determinado artículo es de $\mu_X = \$280$ con una desviación estándar poblacional de \$50. Si se decide seleccionar muestras aleatorias de 36 artículos, calcule

a. La probabilidad de que la media muestral sea a lo sumo de $260.

b. La probabilidad de que la media muestral se encuentre a no más de $10 de la media poblacional.

c. Los valores dentro de los cuales se encuentra el 50% central de la medias muestrales.

d. El monto promedio máximo del 35% de las medias muestrales de menor monto.

Solución

a) Debe hallarse $\mathbb{P}(\bar{X} \leq 260)$.

```
> sigma <- 50; n <- 36; mu <- 280; xbarra <- 260
> sigmaxbarra <- sigma/sqrt(n)
> p <- pnorm(260, 280, sigmaxbarra); p

[1] 0.008197536
```

La Figura 1.4 muestra la región bajo la curva que representa la probabilidad que la media muestral sea a lo sumo de $260.

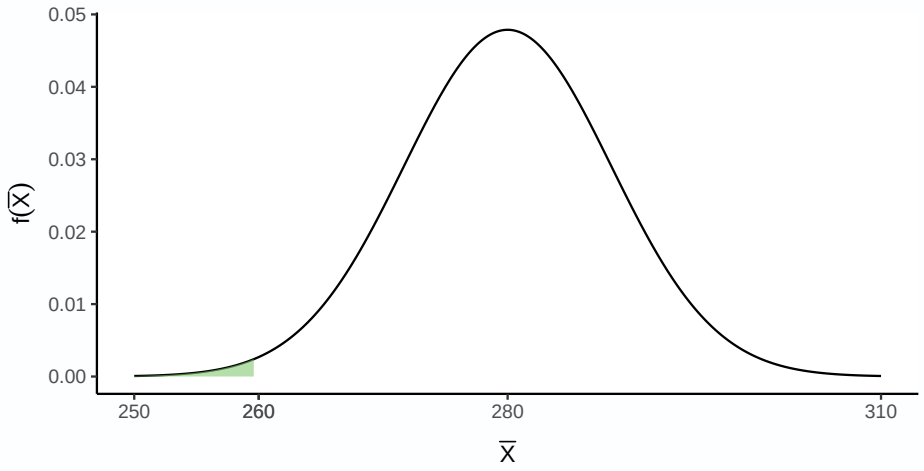


Figura 1.4: Probabilidad de que la media muestral sea a lo sumo de 260 para $n = 36$

Fuente: Elaboración propia

Nota: El código de R para la construcción de la Figura 1.4 puede verse en el Anexo 2.

b) Debe hallarse la probabilidad de que el promedio no se aleje en más de $10 de la media poblacional, es decir, $\mathbb{P}(270 \leq \bar{X} \leq 290)$.

```
> sigma <- 50; n <- 36; sigmaxbarra <- sigma/sqrt(n)
> p1 <- pnorm(270, 280, sigmaxbarra)
> p2 <- pnorm(290, 280, sigmaxbarra)
> p2 - p1

[1] 0.7698607
```

La Figura 1.5 muestra la región bajo la curva que representa la probabilidad de que $270 \leq \bar{X} \leq 290$. Un código de R que puede emplearse para la construcción de esta la Figura 1.5 se muestra en el Anexo 3

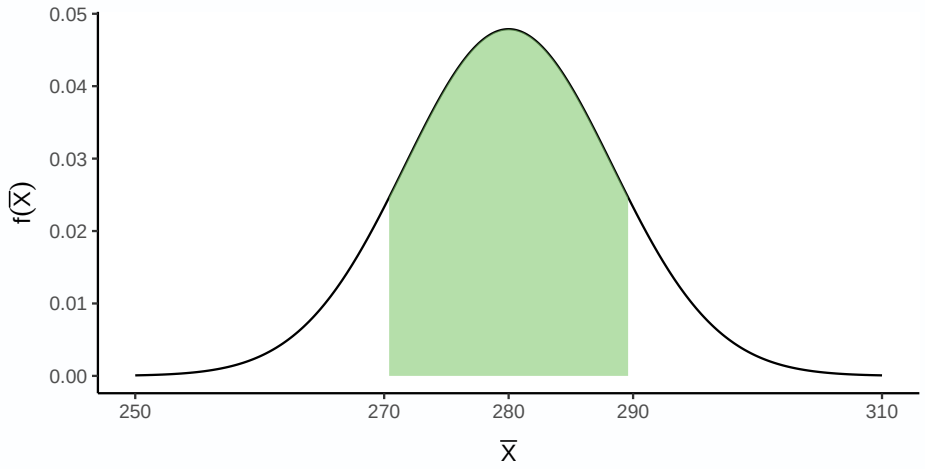


Figura 1.5: Probabilidad de que la media muestral esté entre 270 y 290 para $n = 36$

Fuente: Elaboración propia

c) Deben hallarse los límites de la región que determina el 50% de las medias muestrales, es decir, el cuantil 25 y el cuantil 75.

```
> sigma <- 50; n <- 36; sigmaxbarra <- sigma/sqrt(n)
> q1 <- qnorm(0.25, 280, sigmaxbarra)
> q2 <- qnorm(0.75, 280, sigmaxbarra)
> c(q1, q2)

[1] 274.3793 285.6207
```

La Figura 1.6 muestra la región bajo la curva que representa el 50% central de las medias muestrales.

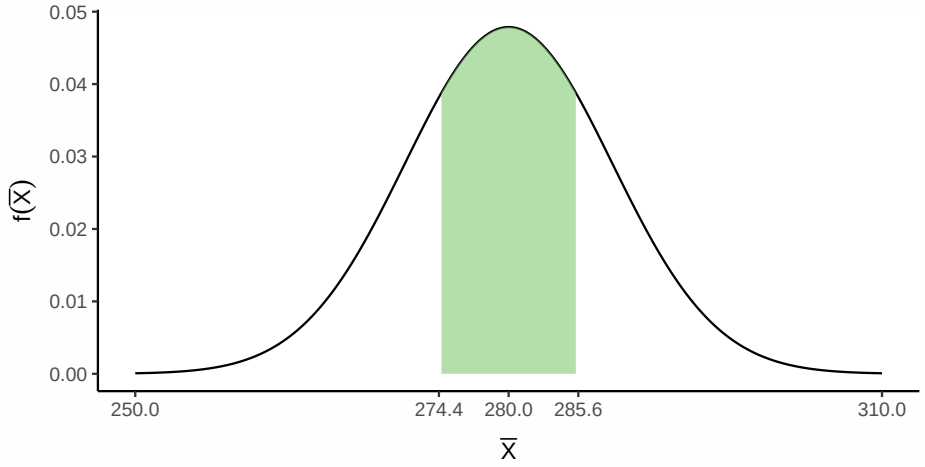


Figura 1.6: Valores que limitan el 50% central de las medias muestrales para $n = 36$

Fuente: Elaboración propia

Ejemplo 1.7 El tiempo que tarda un componente electrónico en fallar tiene una distribución normal, con $\mu_X = 110$ horas y una variancia poblacional de 441 horas. Se decide seleccionar muestras con reemplazo de tamaño 30, para determinar la vida útil del componente.

a. ¿Qué proporción de medias muestrales estarán entre 100 y 119 horas?

b. ¿Qué proporción de medias muestrales estarán Sobrepasando las 99 horas?

c. ¿Dentro de qué límites estará el 90% central de las medias muestrales?

d. ¿Cuánto es el porcentaje de componentes electrónicos con una duración media de a lo sumo 105 horas?

Ejemplo 1.8 En una población, la edad donde se inicia la vida laboral tiene una media de 25 años y desviación estándar de 5 años. Se elige aleatoriamente una muestra de 100 personas. Sea $\bar{X}$ la edad promedio de inicio a la vida laboral.

a) Calcule la media y la variancia para $\bar{X}$.

b) ¿Cuál es la probabilidad de que $\bar{X}$ esté comprendida entre 23 y 26 años?

c) ¿Cuál es la probabilidad de que $\bar{X}$ supere los 24 años?

d) ¿Cuál es la probabilidad de que $\bar{X}$ sea menor a los 25,5 años?

Solución

a) La media de $\bar{X}$ es $E[\bar{X}] = \mu_{\bar{X}} = 25$. La variancia de $\bar{X}$ es $\sigma^2_{\bar{X}} = \dfrac{\sigma^2_X}{n} = \dfrac{25}{100} = \dfrac{1}{4}$.

b) Debe hallarse $\mathbb{P}(23 < \bar{X} < 26)$, es decir, $F(26) - F(23)$.

```
> sigma <- 5; n <- 100; sigmaxbarra <- sigma/sqrt(n); mu = 25
> p1 <- pnorm(23, mean = mu, sd = sigmaxbarra)
> p2 <- pnorm(26, mean = mu, sd = sigmaxbarra)
> p2 - p1

[1] 0.9772182
```

La Figura 1.7 muestra la región bajo la curva que representa la probabilidad de que $23 < \bar{X} < 26$.

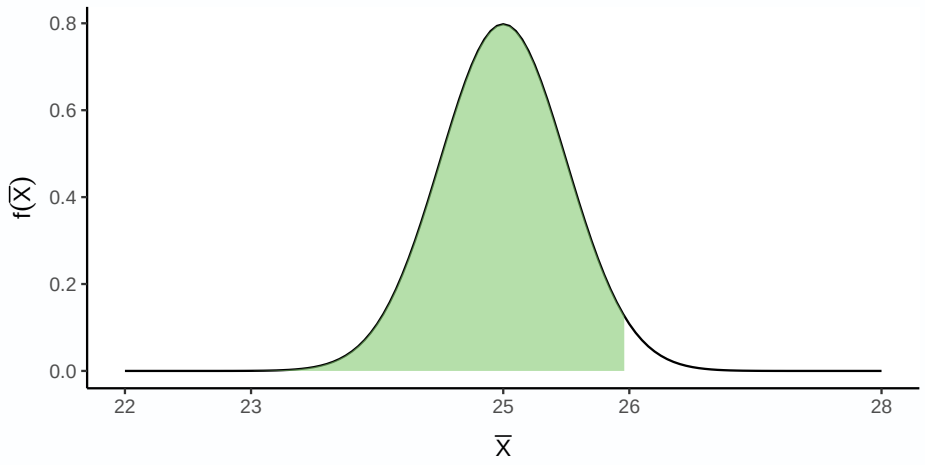


Figura 1.7: Probabilidad de que la media muestral esté entre 23 y 26 para $n = 100$

Fuente: Elaboración propia

c) Debe hallarse $\mathbb{P}(\bar{X} > 24)$, es decir, $1 - F(24)$.

```
> sigma <- 5; n <- 100; sigmaxbarra <- sigma/sqrt(n); mu = 25
> p1 <- 1 - pnorm(24, mean = mu, sd = sigmaxbarra); p1

[1] 0.9772499
```

La Figura 1.8 muestra la región bajo la curva que representa la probabilidad de que $\bar{X} > 24$.

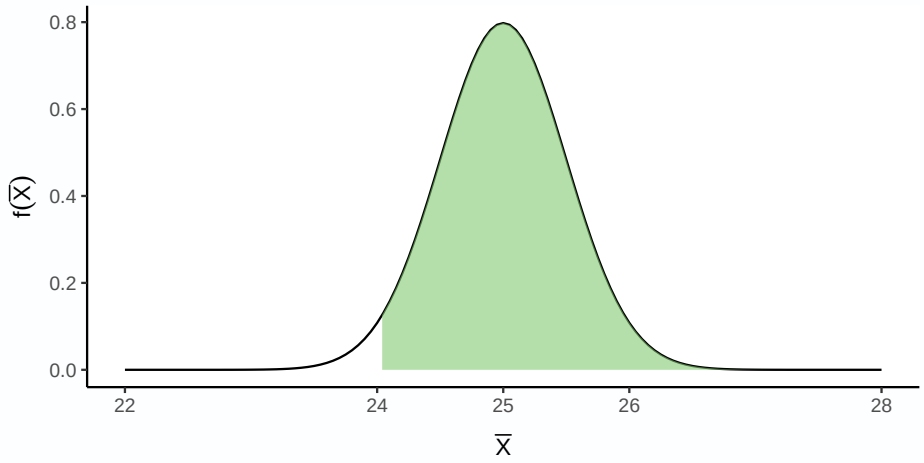


Figura 1.8: Probabilidad de que la media muestral sea superior a 24 para $n = 100$

Fuente: Elaboración propia

d) Debe hallarse $\mathbb{P}(\bar{X} < 25,5)$, es decir, $F(25,5)$.

```
> sigma <- 5; n <- 100; sigmaxbarra <- sigma/sqrt(n); mu = 25
> p1 <- pnorm(25.5, mean = mu, sd = sigmaxbarra); p1

[1] 0.8413447
```

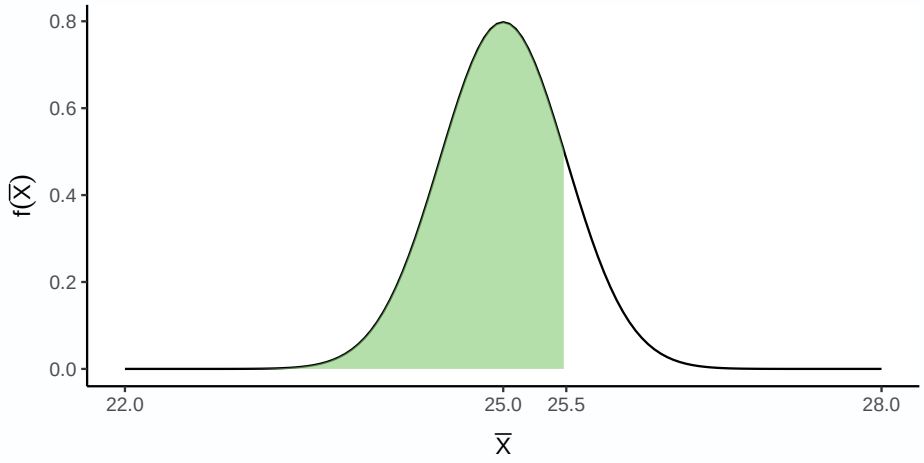


Figura 1.9: Probabilidad de que la media muestral sea inferior 25,5 para $n = 100$

Fuente: Elaboración propia

1.1.2. Distribución de muestreo para la proporción

En ocasiones es necesario resolver problemas donde la media no es el indicador más apropiado para analizar una población. Considere el caso en el que desea conocerse la opinión que tienen las personas sobre determinado tema, por ejemplo, intención de voto por una determinada candidatura. En este caso sería más recomendable utilizar la proporción P de personas de la población que deseen votar por dicha candidatura.

Recuerde que dada una población de tamaño N, si X representa la cantidad de elementos de la población que satisfacen una característica de interés, entonces la proporción de elementos que satisfacen la característica se denota por P y se define por

$$P = \frac{X}{N}$$

En el caso de una muestra de tamaño n, la proporción de elementos que satisfacen la característica de interés se denota por $\hat{p}$ y se define por

$$\hat{p} = \frac{X}{n}$$

Definición 1.9 Distribución muestral para la proporción

Una distribución de muestreo para la proporción muestral está dada por la distribución de probabilidad para los valores posibles del estadístico $\hat{p}$ que resulta de extraer repetidamente de la población muestras de tamaño n.

Ejemplo 1.9 Suponga que se cuenta con un grupo de 12 personas, donde 4 residen en zona urbana. Se van a seleccionar 5 personas al azar de ese grupo sin reemplazo. Construya la distribución de muestreo de la proporción para la cantidad de personas que residen en zona urbana.

Solución

La proporción de cada una de las muestras se detalla a continuación:

z. urbana	z. rural	$\hat{p}$	muestras	$f(\hat{p})$
0	5	0	${}_4C_0 \cdot {}_8C_5 = 56$	$\frac{56}{792}$
1	4	$\frac{1}{5}$	${}_4C_1 \cdot {}_8C_4 = 280$	$\frac{280}{792}$
2	3	$\frac{2}{5}$	${}_4C_2 \cdot {}_8C_3 = 336$	$\frac{336}{792}$
3	2	$\frac{3}{5}$	${}_4C_3 \cdot {}_8C_2 = 112$	$\frac{112}{792}$
4	1	$\frac{4}{5}$	${}_4C_4 \cdot {}_8C_1 = 8$	$\frac{8}{792}$

Si $x_1, x_2, \cdots, x_n$ es una muestra aleatoria (con reemplazo) de una población infinita con proporción P entonces, el valor esperado de $\hat{p}$ denotado por $E(\hat{p})$ o $\mu_{\hat{p}}$ y la variancia de $\hat{p}$, denotada por $Var(\hat{p})$ o $\sigma^2_{\hat{p}}$, se difinen por:

$$E(\hat{p}) = P$$

$$Var(\hat{p}) = \frac{P(1-P)}{n}$$

En su defecto, la desviación estándar de $\hat{p}$ se denota por $\sigma_{\hat{p}}$ y queda definida por:

$$\sigma_{\hat{p}} = \sqrt{\frac{P(1-P)}{n}}$$

Corolario 1.1 Sea $X \sim Bin(n,\ P)$ y $\hat{p} = \frac{X}{n}$, la proporción de éxitos de la muestra. Entonces, si n es suficientemente grande, la distribución de muestreo de $\hat{p}$ es aproximadamente normal con media P y desviación estándar $\sqrt{P(1-P)/n}$. De manera similar, la distribución de muestreo para X es aproximadamente normal con media nP y desviación estándar $\sqrt{nP(1-P)}$.

Por consiguiente, si se tiene una muestra n lo suficientemente grande, la distribución muestral de $\hat{p}$ puede aproximarse por medio de la distribución normal de manera similar al procedimiento usado para aproximar la distribución de probabilidad binomial.

Resumiendo se tiene que, $\hat{p} = \frac{X}{n}$ con $\mu_{\hat{p}} = P$ y $\sigma_{\hat{p}} = \sqrt{\frac{P(1-P)}{n}}$.

Una regla empírica para determinar si la aproximación es adecuada cuando ($n\hat{p} > 5$ y $n(1-\hat{p}) > 5$).

En la mayoría de los casos en los que se hace inferencia utilizando la proporción, las muestras son grandes y por lo tanto la distribución de muestreo de la proporción es aproximadamente normal, por lo que pueden calcularse probabilidades. En caso de no tener a disposición un software que realice los cálculos, puede utilizarse la tabla de la distribución normal estándar, pues $\hat{p}$ tiene una distribución normal con media $\mu_{\hat{p}} = P$ y desviación estándar $\sigma_{\hat{p}} = \sqrt{\dfrac{P(1-P)}{n}}$, entonces:

$$Z = \frac{\hat{p} - P}{\sqrt{\dfrac{P(1-P)}{n}}}, \text{ donde } Z \sim N(0,1)$$

También, puede definirse la distribución muestral de X como:

$$Z = \frac{X - nP}{\sqrt{nP(1-P)}}, \text{ donde } Z \sim N(0,1)$$

Ejemplo 1.10 La gerencia de una empresa fabricante de computadoras ha determinado que la proporción de las ventas del año anterior que incluyeron la entrega en un plazo de 10 días después de la compra fue de 0,25. Si se selecciona una muestra aleatoria de 310 ventas, calcule la probabilidad de que la proporción muestral de los pedidos entregados dentro de los 10 días siguientes sean de

a) entre 0,20 y 0,29.

b) a lo sumo 0,22.

c) al menos 0,28.

d) Dentro de qué valores se encuentra el 60% central de las proporciones muestrales.

Solución

a) Debe hallarse,

$$\mathbb{P}\left(0,20 < \hat{p} < 0,29 / \ \mu_{\hat{p}} = 0,25,\ \sigma_{\hat{p}} = \sqrt{\frac{0,25(1-0,25)}{310}}\right) = F(0,29) - F(0,20)$$

```
> P <- 0.25; n <- 310
> p1 <- pnorm(0.20, mean = P, sd = sqrt(P*(1 - P)/n))
> p2 <- pnorm(0.29, mean = P, sd = sqrt(P*(1 - P)/n))
> p2 - p1

[1] 0.9270496
```

La Figura 1.10 representa gráficamente la probabilidad determinada.

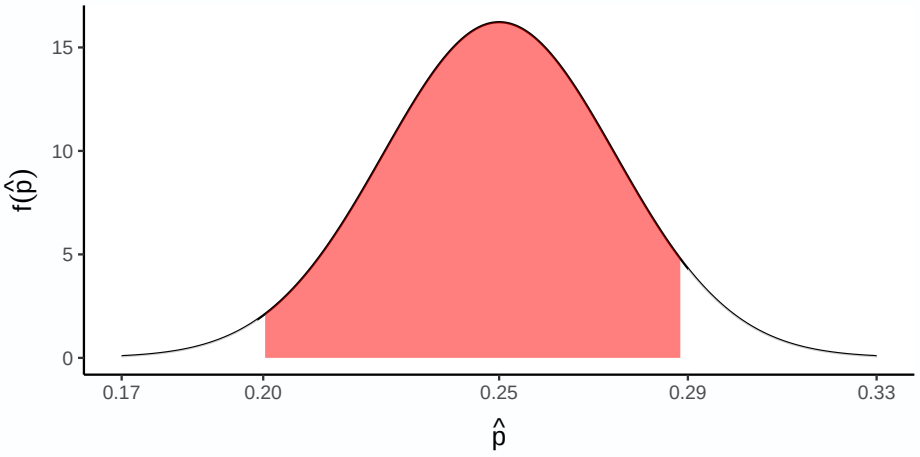


Figura 1.10: Probabilidad de que la proporción muestral esté entre 0,2 y 0,29 para $P = 0{,}25$ y $n = 310$

Fuente: Elaboración propia

b) Debe hallarse,

$$\mathbb{P}\left(\hat{p} \leq 0,22/\ \mu_{\hat{p}} = 0,25,\ \sigma_{\hat{p}} = \sqrt{\frac{0,25(1-0,25)}{310}}\right) = F(0,22)$$

La Figura 1.11 representa gráficamente la probabilidad requerida.

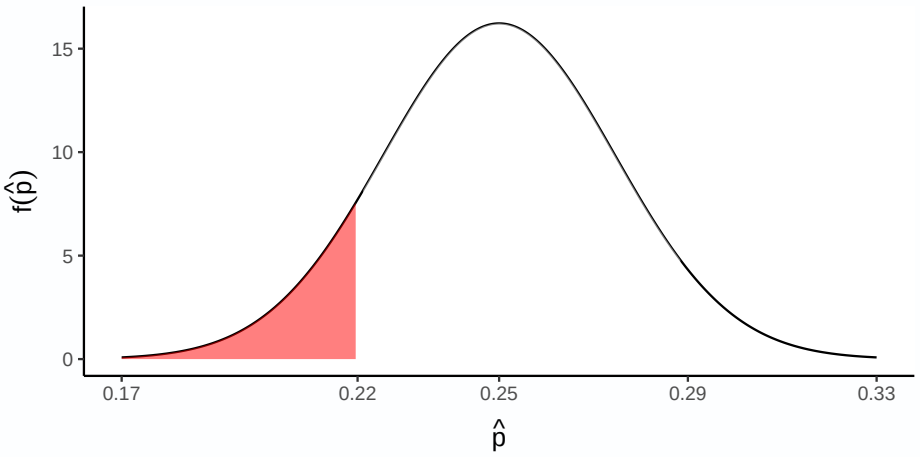


Figura 1.11: Probabilidad de que la proporción muestral sea a lo sumo 0,22 para $P = 0{,}25$ y $n = 310$

Fuente: Elaboración propia

```
> P <- 0.25; n <- 310
> p1 <- pnorm(0.22, mean = P, sd = sqrt(P*(1 - P)/n)); p1

[1] 0.1112635
```

c) Debe hallarse,

$$\mathbb{P}\left(\hat{p} \geq 0,28/\ \mu_{\hat{p}} = 0,25,\ \sigma_{\hat{p}} = \sqrt{\frac{0,25(1-0,25)}{310}}\right) = 1 - F(0,28)$$

```
> P <- 0.25; n <- 310
> p1 <- 1 - pnorm(0.28, mean = P, sd = sqrt(P*(1 - P)/n)); p1

[1] 0.1112635
```

La Figura 1.12 representa gráficamente la probabilidad determinada.

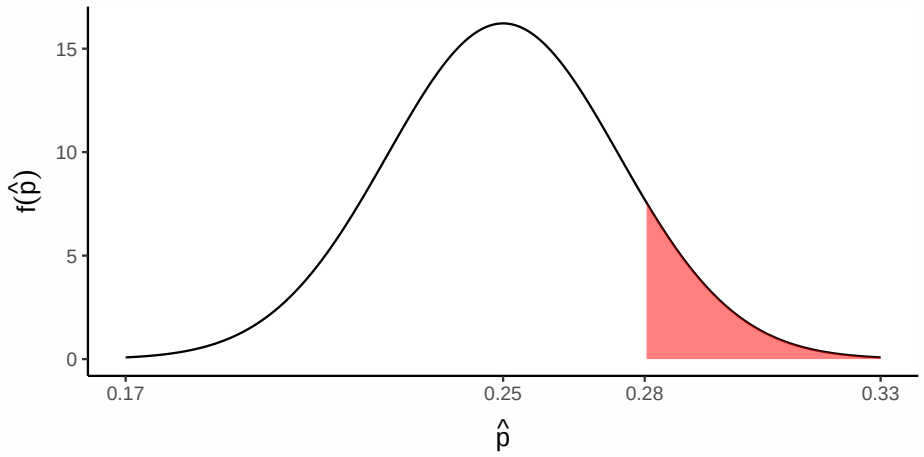


Figura 1.12: Probabilidad de que la proporción muestral sea al menos 0,28 para $P = 0,25$ y $n = 310$

Fuente: Elaboración propia

d) El 60% central de las proporciones muestrales se encuentra entre el cuantil 20 y el cuantil 80.

```
> P <- 0.25; n <- 310
> c20 <- qnorm(0.20, mean = P, sd = sqrt(P*(1 - P)/n))
> c80 <- qnorm(0.80, mean = P, sd = sqrt(P*(1 - P)/n))
> round(c(c20, c80), 4)

[1] 0.2293 0.2707
```

El 60% central de las proporciones muestrales se encuentra entre 22,93% y 27,07%

La Figura 1.13 muestra la región bajo la curva que representa el 60% central de las proporciones muestrales.

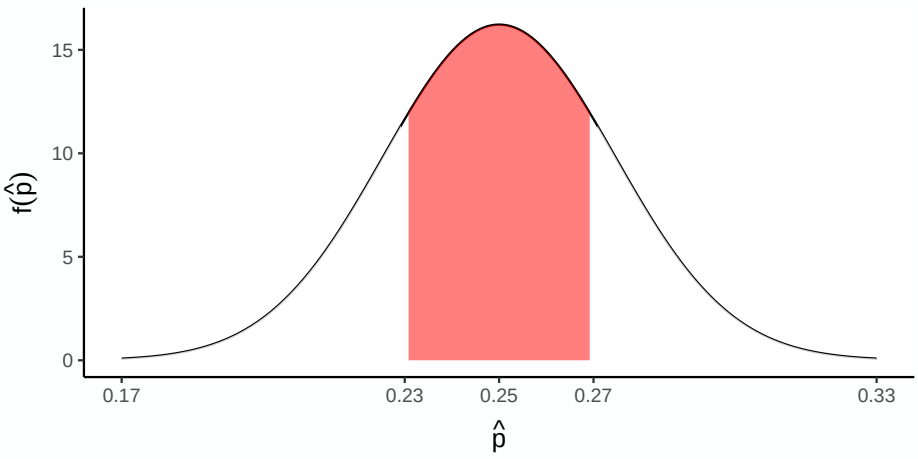


Figura 1.13: Región que representa el 60% de las proporciones muestrales para $P = 0{,}25$ y $n = 310$

Fuente: Elaboración propia

Es importante indicar que en ocasiones el verdadero valor de P no es conocido, en estas situaciones los cálculos se realizan sustituyendo el valor de P por alguna referencia de este, es decir, un valor de $\hat{p}$.

Ejemplo 1.11 Debido a la huelga de educadores efectuada en 1995 por el asunto del régimen de pensiones se realizó una encuesta a 500 personas jefes de hogar del Gran Área Metropolitana acerca de si la huelga era justa. Como resultado de la encuesta, se obtuvo que el 55% de las personas entrevistadas consideró que la huelga era justa.

a) ¿Cuál es la probabilidad de observar una proporción muestral de personas que apoyan la huelga de al menos el 60%?

b) ¿Cuál es la probabilidad de observar una proporción muestral de personas que apoyan la huelga de a lo sumo el 44%?

c) ¿Dentro de qué valores se encuentra el 40% central de las proporciones muestrales de personas que apoyan la huelga?

1.1.3. Factor de corrección para población finita

Cuando las muestras son obtenidas de poblaciones finitas relativamente pequeñas se acostumbraba introducir a la fórmula del error estándar un factor de corrección finita con el objetivo de minimizar el error. Como en la práctica se trabajan con poblaciones finitas muy grandes, esta técnica es poco utilizada, en general se estima que si el tamaño de la muestra es menor al 5% del tamaño de la población, la corrección es insignificante y se puede prescindir de ella.

En los casos en los que se utiliza el factor de corrección finita (f. c. f), este puede ser aplicado tanto a distribuciones muestrales para la media y la proporción, siempre y cuando el muestreo se realice sin reemplazo y está definido por $\frac{N-n}{N-1}$. Cuando n es grande, $\frac{N-n}{N-1} = 1$, donde:

- N: tamaño de la población.
- n: el tamaño de la muestra.

Ejemplo 1.12 La proporción de personas solteras de edades comprendidas entre 25 y 35 años en un pequeño pueblo de Costa Rica es de $\frac{2}{3}$. Suponga que se obtienen muestras de tamaño 36 de todas las personas del pueblo entre dichas edades.

a) Halle la media y la desviación estándar de la proporción de personas solteras entre 25 y 35 años.

b) Suponga que en el pueblo viven 226 personas entre 25 y 35 años y que el muestreo es sin reemplazo. ¿Cuál es la media y la desviación estándar de la proporción?

Capítulo 2

Estimación estadística

2.1. Estimación puntual

Definición 2.1 Estimador puntual

Según Freund, Miller, y Miller (2004), un estimador puntual es cualquier función $W(X_1, \cdots, X_n)$ de una muestra, donde $X_1, \cdots, X_n$ son variables aleatorias que se obtienen de la muestra. Esto es, cualquier estadístico es un estimador puntual.

Nota 1 *Tal y como se mencionó en el Capítulo 1, existe una diferencia entre un* ***estimador*** *y una* ***estimación****. Un estimador es una función de la muestra, mientras que una estimación es el valor de un estimador, es decir, el número.*

2.2. Propiedades de los estimadores

Los estimadores estadísticos cumplen con las siguientes propiedades: insesgados, eficientes, suficientes y consistentes.

2.2.1. Estimadores Insesgados

Definición 2.2 Estimador insesgado

Un estimador $\hat{\theta}$ es un estimador insesgado del parámetro θ si y solo si $E(\hat{\theta}) = \theta$, caso contrario el estimador se dice que es sesgado y su sesgo se denota por $B(\hat{\theta})$ y se define por:

$$B(\hat{\theta}) = E(\hat{\theta}) - \theta$$

Ejemplo 2.1 Sea X una variable aleatoria tal que $X \sim Bin(n, P)$, con n conocido y P desconocido. Muestre que la proporción muestral $\hat{p} = \dfrac{X}{n}$ es un estimador insesgado de P.

Solución

La proporción muestral $\hat{p}$ es un estimador insesgado de P si $E[\hat{p}] = P$.

Prueba

Si X $\sim Bin(n, P)$, entonces $E[X] = nP$. De esta forma se tiene que

$$E[\hat{p}] = E\left[\frac{X}{n}\right] = \frac{1}{n}E[X] = \frac{1}{n}nP = P$$

Efectivamente, la proporción muestral $\hat{p}$ es un estimador insesgado de P.

Ejemplo 2.2 Sea $X_1, \cdots, X_n$ una muestra aleatoria de una distribución $N(\mu_X, \sigma_X^2)$, entonces $\bar{X}$ es un estimador insesgado de μ_X.

Ejemplo 2.3 Sea $\hat{\mu}_X = \dfrac{1}{n+1}\displaystyle\sum_{i=1}^{n} X_i$. Muestre que $\hat{\mu}_X$ es un estimador sesgado de μ_X.

Solución

$\hat{\mu}_X$ es un estimador sesgado de μ_X si $E(\hat{\mu}_X) \neq \mu_X$.

$$E(\hat{\mu}_X) = E\left[\frac{1}{n+1}\sum_{i=1}^{n} x_i\right] = \frac{1}{n+1}E\left[\sum_{i=1}^{n} x_i\right] = \frac{1}{n+1}\sum_{i=1}^{n} E(x_i) = \frac{1}{n+1}(n\mu_X) = \frac{n\mu_X}{n+1} \neq \mu_X$$

Efectivamente, $\hat{\mu}_X$ es un estimador sesgado de μ_X.

Teorema 2.1 Si S^2 es la variancia de una muestra aleatoria de una población finita con variancia finita σ_X^2, entonces $E(S^2) = \sigma_X^2$.

Prueba

Sugerencia: Pruebe como lema el hecho que: $\displaystyle\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \mu_X)^2 - n(\mu_X - \bar{X})^2$

$$
\begin{aligned}
E[S^2] &= E\left[\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}\right] \\
&= E\left[\frac{\sum_{i=1}^{n}(X_i-\mu_X)^2-n(\mu_X-\bar{X})^2}{n-1}\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}E(X_i-\mu_X)^2-nE(\mu_X-\bar{X})^2\right] \\
&= \frac{1}{n-1}\left[n\sigma_X^2-n\frac{\sigma_X^2}{n}\right] \\
&= \sigma_X^2
\end{aligned}
$$

Estimadores Asintóticamente Insesgados

Definición 2.3 Un estimador $\hat{\theta}$ es asintóticamente insesgado si su posible sesgo tiende a cero al aumentar el tamaño muestral, es decir:

$$\lim_{n\to\infty} B(\hat{\theta}) = 0$$

Ejemplo 2.4 Sea $X_1, \cdots, X_n$ una muestra aleatoria de una pobalción normal con media μ_X y variancia σ_X^2). Compruebe que $\hat{\mu}_X = \frac{1}{n+1}\sum_{i=1}^{n} x_i$ es un estimador asintóticamente insesgado de μ_X.

Solución

Como se comprobó en el Ejemplo 2.3, $E(\hat{\mu}_X) = \frac{n\mu_X}{n+1}$. Así,

$$B(\hat{\mu}_X) = E(\hat{\mu}_X) - \mu_X = \frac{n\mu_X}{n+1} - \mu_X = \frac{n\mu_X - n\mu_X - \mu_X}{n+1} = -\frac{\mu_X}{n+1}$$

Por lo tanto,

$$\lim_{n\to\infty} B(\hat{\mu}_X) = \lim_{n\to\infty} -\frac{\mu_X}{n+1} = 0$$

Finalmente, $\hat{\mu}_X = \dfrac{1}{n+1}\displaystyle\sum_{i=1}^{n} x_i$ es un estimador asintóticamente insesgado de μ_X.

Error Cuadrático Medio de los Estimadores

Definición 2.4 El Error cuadrático medio, MSE por sus siglas en inglés (Mean Square Error), de un estimador $\hat{\theta}$ está definido por:

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = Var(\hat{\theta}) + [B(\hat{\theta})]^2,$$

donde $B(\hat{\theta})$ es el sesgo de $\hat{\theta}$, es decir, $B(\hat{\theta}) = E(\hat{\theta}) - \theta$.

Nota 2 *Recuerde que* $Var[\hat{\theta}] = E[(\hat{\theta} - E[\hat{\theta}])^2]$

Observación:

El MSE toma en cuenta tanto la variabilidad como el sesgo del estimador. En general, cuando se comparan dos estimadores $\hat{\theta}_1$ y $\hat{\theta}_2$ de θ, a menudo debe elegirse entre dos alternativas: poca variabilidad o poco sesgo.

Ejemplo 2.5 Sea X una variable aleatoria tal que $X \sim Bin(n, P)$, con n conocido y P desconocido. La proporción muestral $\hat{p}_1 = \dfrac{X}{n}$ es un estimador insesgado de P. Determine la $Var(P)$ y el $MSE(P)$.

Solución

$$E[\hat{p}_1] = P$$
$$Var[\hat{p}_1] = \frac{P(1-P)}{n}$$
$$MSE[\hat{p}_1] = \frac{P(1-P)}{n}$$

Ejemplo 2.6 Considere que $\hat{p}_2 = \dfrac{X+1}{n+2}$ es un estimador de P, determine el $MSE(\hat{p}_2)$.

Solución

El valor esperado de $\hat{p}_2$ está dado por:

$$E[\hat{p}_2] = \frac{1}{n+2}(E[X+1]) = \frac{1}{n+2}(E[X]+E[1]) = \frac{nP+1}{n+2}$$

Este es un caso de un estimador *sesgado* y el sesgo de $\hat{p}_2$ esta dado por:

$$B[\hat{p}_2] = \frac{nP+1}{n+2} - P = \frac{1-2P}{n+2}$$

La variancia de $\hat{p}_2$ es:

$$Var[\hat{p}_2] = Var\left[\frac{X+1}{n+2}\right] = \frac{1}{(n+2)^2}(Var[X]+Var[1]) = \frac{1}{(n+2)^2}nP(1-P) = \frac{nP(1-P)}{(n+2)^2}$$

El MSE de $\hat{p}_2$ corresponde a:

$$MSE[\hat{p}_2] = \frac{nP(1-P)}{(n+2)^2} + \left(\frac{1-2P}{n+2}\right)^2 = \frac{nP(1-P)+(1+2P)^2}{(n+2)^2}$$

2.2.2. Estimadores Eficientes

Definición 2.5 Si $\hat{\theta}_1$ y $\hat{\theta}_2$ son estimadores de θ y $Var(\hat{\theta}_1) \leq Var(\hat{\theta}_2)$ entonces, $\hat{\theta}_1$ es un estimador más eficiente de θ que $\hat{\theta}_2$.

Definición 2.6 Se dice que $\hat{\theta}$ es el mejor estimador insesgado de θ, si satisface las siguientes condiciones:

- $\hat{\theta}$ es insesgado.
- $Var(\hat{\theta}) \leq Var(\hat{\theta}_1)$, donde $\hat{\theta}_1$ es cualquier otro estimador insesgado de θ.

A este estimador se le acostumbra llamar *estimador insesgado de variancia mínima* (EIVM) o *mejor estimador insesgado*. La eficiencia de un estimador $\hat{\theta}$ se denota por $eff(\hat{\theta})$.

Se puede definir la eficiencia de un estimador, $\hat{\theta}$, como la inversa de su error cuadrático medio, es decir:

$$eff(\hat{\theta}) = \frac{1}{MSE(\hat{\theta})}$$

Un estimador $\hat{\theta}_1$ se dice que es más preciso o eficiente que un estimador $\hat{\theta}_2$ si, $eff(\hat{\theta}_1) \geq eff(\hat{\theta}_2)$, o bien si $MSE(\hat{\theta}_1) \leq MSE(\hat{\theta}_2)$.

Eficiencia relativa

En algunas ocasiones interesa saber si un estimador insesgado $\hat{\theta}_1$ es "mejor" que otro estimador insesgado $\hat{\theta}_2$. Para ello se utiliza el concepto de **eficiencia relativa** para comparar las variancias respectivas de ambos estimadores.

Definición 2.7 La eficiencia relativa de $\hat{\theta}_1$ con respecto a $\hat{\theta}_2$, se denota por $eff(\hat{\theta}_1, \hat{\theta}_2)$ y se define como:

$$eff(\hat{\theta}_1, \hat{\theta}_2) = \frac{eff(\hat{\theta}_1)}{eff(\hat{\theta}_2)} = \frac{MSE(\hat{\theta}_2)}{MSE(\hat{\theta}_1)}$$

.

En el caso de estimadores insesgados, la eficiencia relativa de $\hat{\theta}_1$ con respecto a $\hat{\theta}_2$ es:

$$eff(\hat{\theta}_1, \hat{\theta}_2) = \frac{Var(\hat{\theta}_2)}{Var(\hat{\theta}_1)}$$

Se dice que $\hat{\theta}_1$ es más eficiente que $\hat{\theta}_2$ si el cociente anterior es mayor que 1.

Ejemplo 2.7 Considere una muestra aleatoria Y_1, Y_2, Y_3 de una distribución exponencial con media θ de modo que $\hat{\theta}_1 = Y_1$, $\hat{\theta}_2 = \frac{1}{2}(Y_1 + Y_2)$ y $\hat{\theta}_3 = \bar{Y}$. Encuentre la eficiencia relativa de $\hat{\theta}_1$ respecto a $\hat{\theta}_3$ y de $\hat{\theta}_2$ respecto a $\hat{\theta}_3$.

Ejemplo 2.8 Suponga que se toma una muestra aleatoria de variables independientes de tamaño 5, de una variable $X \sim Exp\left(\lambda = \frac{1}{\theta}\right)$. Encuentre la eficiencia relativa de $\hat{\lambda}_1$ respecto a $\hat{\lambda}_2$, si los estimadores $\hat{\lambda}_1$ y $\hat{\lambda}_2$ se definen de la siguiente manera:

$$\hat{\lambda}_1 = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{5}$$

y

$$\hat{\lambda}_2 = \frac{2X_1 + X_2 + X_3 - X_4 + 3X_5}{6}$$

Solución

Recuerde que si $X \sim Exp\left(\lambda = \frac{1}{\theta}\right)$, entonces $E[X] = \frac{1}{\lambda}$ y $Var[X] = \frac{1}{\lambda^2}$

Ejemplo 2.9 Suponga que el verdadero largo de una pieza de metal para construcción, sigue una distribución normal, con promedio μ_X desconocido y desviación estándar conocida $\sigma_X = 0,5$ pulgadas. Si una muestra aleatoria de variables independientes de tamaño 3, de ese tipo de piezas, es tomada para estimar el valor de μ_X, ¿cuál de los estimadores $\hat{\mu}_{1_X}$ o $\hat{\mu}_{2_X}$ es mejor en términos de: sesgo, variancia y eficiencia relativa, si $\hat{\mu}_{1_X} = \frac{1}{3} \cdot (X_1 + X_2 + X_3)$ y $\hat{\mu}_{2_X} = \frac{1}{2} \cdot (X_1 + X_2)$?

Teorema 2.2 Desigualdad de Cramer-Rao

Si $\hat{\theta}$ es un estimador insesgado de θ y

$$Var(\hat{\theta}) \geq \frac{1}{n \cdot E\left[\left(\frac{\partial ln\left[f(X)\right]}{\partial \theta}\right)^2\right]}$$

Si se cumple que,

$$Var(\hat{\theta}) = \frac{1}{n \cdot E\left[\left(\frac{\partial ln\left[f(X)\right]}{\partial \theta}\right)^2\right]}$$

entonces, en algún sentido, $\hat{\theta}$ es el mejor estimador disponible y se le conoce como **estimador insesgado de variancia mínima de** θ.

Aquí la cantidad del denominador se refiere a la información sobre θ que es suministrada por la muestra. Así, la menor variancia es dada por el estimador que brinda mayor información.

Ejemplo 2.10 Muestre que $\bar{X}$ es un estimador insesgado de variancia mínima del promedio μ_X de una población normal.

Prueba

Debe mostrarse que $\frac{1}{n \cdot E\left[\left(\frac{\partial ln\left[f(X)\right]}{\partial \mu_X}\right)^2\right]} = \frac{\sigma_X^2}{n}$

En el caso de la distribución normal, se sabe que:

$$f(x) \quad = \quad \frac{1}{\sigma_X\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2}, \text{ para } x \in \mathbb{R}$$

$$
\begin{aligned}
ln\left[f(x)\right] &= ln\left[\frac{1}{\sigma_X\sqrt{2\pi}}\cdot e^{-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2}\right] \\
ln\left[f(x)\right] &= ln\left[\frac{1}{\sigma_X\sqrt{2\pi}}\right] + ln\left[e^{-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2}\right] \\
ln\left[f(x)\right] &= -ln\left(\sigma_X\sqrt{2\pi}\right) - \frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2 \\
\frac{\partial ln\left[f(x)\right]}{\partial \mu_X} &= \frac{1}{\sigma_X}\left(\frac{x-\mu_X}{\sigma_X}\right)
\end{aligned}
$$

Ahora,

$$
E\left[\left(\frac{\partial ln\left[f(X)\right]}{\partial \mu_X}\right)^2\right] = E\left[\frac{1}{\sigma_X}\left(\frac{x-\mu_X}{\sigma_X}\right)\right] = \frac{1}{\sigma_X^2}\cdot E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)^2\right] = \frac{1}{\sigma_X^2}\cdot 1 = \frac{1}{\sigma_X^2}
$$

Finalmente,

$$
\frac{1}{n\cdot E\left[\left(\frac{\partial ln\left[f(X)\right]}{\partial \mu_X}\right)^2\right]} = \frac{1}{n\cdot\frac{1}{\sigma_X^2}} = \frac{\sigma_X^2}{n}
$$

Ejemplo 2.11 Sea $X_1, \cdots, X_n$ una muestra aleatoria de una distribución de Poisson. Muestre que $\hat{\Theta} = \bar{X}$ es un estimador insesgado de variancia mínima de λ.

2.2.3. Estimadores Suficientes

Definición 2.8 Un estimador $\hat{\theta}$ de θ es suficiente, si utiliza toda la información sobre el parámetro contenida en la muestra para la estimación de θ, de modo que ningún otro estimador pueda proporcionar información adicional sobre el parámetro desconocido θ. Es decir, si todo el conocimiento acerca de θ que puede ser adquirido desde cada uno de los valores de la muestra, también puede ser adquirido solo por medio del valor de $\hat{\theta}$.

Por ejemplo, la media muestral es un estimador suficiente de la media poblacional.

2.2.4. Estimadores Consistentes

Definición 2.9 Un estimador $\hat{\theta}$ es un estimador consistente del parámetro θ, si su valor esperado tiende a θ cuando el tamaño de muestra aumenta. Esto es:

1. $E\left[\hat{\theta}\right] \to \theta$ cuando $n \to \infty$.
2. $Var(\hat{\theta}) \to 0$ cuando $n \to \infty$.

De forma equivalente, un estimador $\hat{\theta}$ es un estimador consistente del parámetro θ, si dado $\epsilon > 0$ se cumple que:

$$\lim_{n\to\infty} P(|\hat{\theta} - \theta| \geq \epsilon) = 0$$

o de manera equivalente:

$$\lim_{n\to\infty} P(|\hat{\theta} - \theta| < \epsilon) = 1$$

Teorema 2.3 Teorema de Chebyshev

Sean $k > 0$ y X una variable aleatoria con variancia finita σ_X. Entonces:

$$P(|X - \mu_X| < k\sigma_X) \geqslant 1 - \frac{1}{k^2}$$

O bien,

$$P(|X - \mu_X| \geq k\sigma_X) \leq \frac{1}{k^2}$$

Otra versión del teorema es:

Sean $\epsilon > 0$ y X una variable aleatoria con variancia finita σ_X. Entonces:

$$P(|X - E(X)| \geqslant \epsilon) \leqslant \frac{Var(X)}{\epsilon^2}$$

Sea $\epsilon = k\sigma_X$.

$$\begin{aligned}
\epsilon = k\sigma_X \quad &\Rightarrow \quad k = \frac{\epsilon}{\sigma_X} \\
&\Rightarrow \quad \frac{1}{k} = \frac{\sigma_X}{\epsilon} \\
&\Rightarrow \quad \frac{1}{k^2} = \frac{\sigma_X^2}{\epsilon^2} \\
&\Rightarrow \quad \frac{1}{k^2} = \frac{Var(X)}{\epsilon^2}
\end{aligned}$$

Ejemplo 2.12 Sean $X_1, X_2, \ldots, X_n$ una muestra aleatoria de tamaño n de una distribución con media μ_X y variancia σ_X^2. Muestre que $\bar{X}$ es un estimador consistente de μ_X.

Solución

Hay que probar que: $\lim_{n\to\infty} P(|\bar{X} - \mu_X| \geq \epsilon) = 0, \forall \epsilon > 0$.

$$\begin{aligned}
P(|\bar{X} - \mu_X| \geq \epsilon) \leq \frac{Var(\bar{X})}{\epsilon^2} \quad &\Rightarrow \quad P(|\bar{X} - \mu_X| \geq \epsilon) \leq \frac{\sigma_X^2}{n\epsilon^2} \\
&\Rightarrow \quad \lim_{n\to\infty} P(|\bar{X} - \mu_X| \geq \epsilon) \leq \lim_{n\to\infty} \frac{\sigma_X^2}{n\epsilon^2} \\
&\Rightarrow \quad \lim_{n\to\infty} P(|\bar{X} - \mu_X| \geq \epsilon) = 0
\end{aligned}$$

Por lo tanto, $\bar{X}$ es un estimador consistente de μ_X.

2.3. Métodos de estimación

2.3.1. El Método de Máxima Verosimilitud

Este método se debe al trabajo de R.A. Fisher, gran estadístico del siglo XX, quien, en sus escritos, Fisher (1934) y Fisher (1932) propuso un método general de estimación llamado **Método de Máxima Verosimilitud**. Él demostró las ventajas del método, mostrando que los estimadores máximo verosímiles son estimadores asintóticamente insesgados de variancia mínima.

Función de verosimilitud para distribuciones discretas

Definición 2.10 Sea $f(x|\theta)$ la función de probabilidad de masa para una distribución discreta con parámetro asociado θ. Suponga que $X_1, X_2, \cdots, X_n$ es una muestra aleatoria de esta distribución

y $x_1, x_2, \cdots, x_n$ son los valores observados en la muestra. Entonces la función de verosimilitud de θ está dada por:

$$\begin{aligned} L(\theta|x_1, x_2, \cdots, x_n) &= P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) \\ &= P(X_1 = x_1) \cdot P(X_2 = x_2) \cdots P(X_n = x_n) \\ &= \prod_{i=1}^{n} P(x_i|\theta). \end{aligned}$$

La función de verosimilitud es una función de θ y a veces se escribe $L(\theta) = L(\theta|x_1, x_2, \cdots, x_n)$.

El estimador máximo verosímil es el valor $\hat{\theta}$ tal que $L(\hat{\theta}) \geq L(\theta)$ para todo θ.

Nota 3

$$L(\theta) = L(\theta|x_1, x_2, \cdots, x_n) = f(x_1, x_2, \cdots, x_n|\theta)$$

Por tanto, el método de máxima verosimilitud consiste en maximizar la función de verosimilitud con respecto al parámetro θ dados los datos. El valor de θ que maximiza la función de verosimilitud se le llama estimador máximo verosímil de θ.

Ejemplo 2.13 Sean $X_1, \cdots X_n$ n variables aleatorias de Bernuolli independientes con parámetro P, $0 < P < 1$. Sea $X = \sum_{i=1}^{n} X_i$ el número de unos. Probar que el estimador máximo verosímil del parámetro P es $\hat{p} = \dfrac{X}{n}$.

Solución

$$\begin{aligned} L(p) &= P(X_1 = x_1, \cdots, X_n = x_n) \\ &= \prod_{i=1}^{n} P(X_i = x_i) \\ &= \prod_{i=1}^{n} P^{x_i} \prod_{i=1}^{n} (1-P)^{1-x_i} \\ &= P^{\sum_{i=1}^{n} x_i} \cdot (1-P)^{n - \sum_{i=1}^{n} x_i} \\ &= P^X (1-P)^{n-X} \end{aligned}$$

Igualando la derivada a cero y resolviendo para P se obtiene $P = \frac{X}{n}$, es decir, el estimador máximo verosímil está dado por $\hat{p} = \frac{X}{n}$.

Ejemplo 2.14 Sean $x_1, x_2, \cdots, x_n$ una muestra aleatoria de una distribución de Poisson con parámetro desconocido λ. Halle el estimador máximo verosímil para λ.

Función de verosimilitud para distribuciones continuas

Definición 2.11 Sea $f(x|\theta)$ una función de densidad de probabilidad para una variable aleatoria continua con parámetro asociado θ. Suponga que $X_1, X_2, \cdots, X_n$ es una muestra aleatoria de esta distribución y $x_1, x_2, \cdots, x_n$ son los valores observados en la muestra. Entonces la función de verosimilitud de θ está dada por:

$$L(\theta|x_1, x_2, \cdots, x_n) = f(x_1, x_2, \ldots, x_n|\theta) = f(x_1|\theta)f(x_2|\theta)\cdots f(x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \tag{2.1}$$

El estimador máximo verosímil es el valor $\hat{\theta}$ tal que $L(\hat{\theta}) \geq L(\theta)$ para todo θ.

Ejemplo 2.15 Si $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$ es una muestra aleatoria *iid* de una población exponencial, encuentre el estimador máximo verosímil para el parámetro λ.

Solución

$$\begin{aligned} L(\lambda|x_1, x_2, \cdots, x_n) &= f(x_1;\lambda)f(x_2;\lambda)\cdots f(x_n;\lambda) \\ &= \lambda e^{-\lambda x_1} \cdot \lambda e^{-\lambda x_2} \cdots \lambda e^{-\lambda x_n} \\ &= \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i} \end{aligned}$$

Aplicando el logaritmo natural se tiene que:

$$ln\left[L(\lambda)\right] = ln\left(\lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}\right)$$

Derivando a ambos lados se tiene que:

$$\frac{dln\left[L(\lambda)\right]}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i$$

Si $\frac{dln\left[L(\lambda)\right]}{d\lambda} = 0$ entonces:

$$\begin{aligned}
\frac{n}{\lambda} - \sum_{i=1}^{n} x_i &= 0 \\
\frac{n}{\lambda} &= \sum_{i=1}^{n} x_i \\
\lambda &= \frac{n}{\sum_{i=1}^{n} x_i} \\
\lambda &= \frac{1}{\frac{\sum_{i=1}^{n} x_i}{n}} \\
\lambda &= \frac{1}{\bar{x}}
\end{aligned}$$

Por lo tanto $\hat{\lambda}$ es el inverso del promedio.

Ejemplo 2.16 Si $x_1, x_2, \cdots, x_n$ son valores de una muestra aleatoria de tamaño n de una población uniforme con $a = 0$, encuentre el estimador máximo verosímil para el parámetro b.

Ejemplo 2.17 Si $X_1, X_2, \cdots, X_n$ constituyen una muestra aleatoria de tamaño n de una población normal con media μ_X y variancia σ_X^2, encuentre los estimadores máximo verosímiles para ambos parámetros.

2.3.2. Método de Momentos

Otra forma de encontrar estimadores para los parámetros de una distribución es el *método de momentos*. Sea $f(x)$ la función de densidad de probabilidad (o probabilidad de masa) de una variable aleatoria X. Para un entero positivo r, el r-ésimo (teórico) momento de X es:

$$E[X^r] = \int_{-\infty}^{+\infty} x^r f(x) dx \tag{2.2}$$

y el r-ésimo momento muestral es :

$$\frac{1}{n} \sum_{i=1}^{n} x_i^r \tag{2.3}$$

Sean $X_1, X_2, \cdots X_n$ una muestra aleatoria de una distribución con r parámetros desconocidos $\theta_1, \theta_2, \cdots \theta_r$. Sea $f(x; \theta_1, \theta_2, \cdots \theta_r)$ la función de densidad de probabilidad de esta distribución. El método de momentos iguala cada r teórico momento con su correspondiente r-ésimo momento muestral, para obtener un sistema de r ecuaciones con r incógnitas.

$$\begin{aligned}
\int_{-\infty}^{+\infty} x f(x; \theta_1, \theta_2, \cdots \theta_r) &= \frac{1}{n}\sum_{i=1}^{n} X_i \\
\int_{-\infty}^{+\infty} x^2 f(x; \theta_1, \theta_2, \cdots \theta_r) &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 \\
\int_{-\infty}^{+\infty} x^3 f(x; \theta_1, \theta_2, \cdots \theta_r) &= \frac{1}{n}\sum_{i=1}^{n} X_i^3 \\
&\vdots \\
\int_{-\infty}^{+\infty} x^r f(x; \theta_1, \theta_2, \cdots \theta_r) &= \frac{1}{n}\sum_{i=1}^{n} X_i^r
\end{aligned}$$

Las soluciones de este sistema de ecuaciones son los estimadores del método de momentos para $\theta_1, \theta_2, \cdots \theta_r$.

Nota 4 *En el caso de una variable aleatoria discreta, se deben reemplazar las integrales por sumatorias.*

Ejemplo 2.18 Suponga que $X_1, X_2, \cdots, X_n$ son muestras aleatorias de una distribución exponencial. Use el método de momentos para encontrar un estimador para λ.

Solución

El primer momento teórico es:

$$E[X] = \int_0^{\infty} \lambda x e^{-\lambda x} dx = \frac{1}{\lambda}$$

El primer momento muestral es el promedio:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Igualando ambos momentos se tiene:

$$\frac{1}{\lambda} = \bar{x}$$

$$\lambda = \frac{1}{\bar{X}}$$

Es decir,

$$\hat{\lambda} = \frac{1}{\bar{X}}$$

Ejemplo 2.19 Sean $X_1, X_2, \cdots, X_n$ variables aleatorias *iid* provenientes de una distribución uniforme de parámetros $a = 0$ y b. Use el método de momentos para encontrar un estimador para b.

Solución

El primer momento teórico es $E[X] = \frac{b}{2}$, mientras que el primer momento muestral es el promedio $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Igualando ambos momentos, se tiene que $\frac{b}{2} = \bar{X}$, es decir, $\hat{b} = 2\bar{X}$.
Note que este estimador puede arrojar resultados imposibles. Por ejemplo, si n= 4 con $x_1 = x_2 = x_3 = 1$, $x_4 = 9$, entonces $\bar{X}$=3 y b =6, lo cual es imposible si $x_4 = 9$.

Ejemplo 2.20 Dada una muestra aleatoria de tamaño n de una población uniforme con $b = 1$. Use el método de momentos para encontrar un estimador para a.

Ejemplo 2.21 Suponga que $x_1 = 1,3$, $x_2 = 1,8$, $x_3 = 2,1$, $x_4 = 2,25$ son los resultados de una muestra aleatoria cuya función de densidad de probabilidad está dada por:

$$f(x;\theta) = \frac{2(\theta - x)}{\theta}$$

para $0 < x < \theta$. Use el método de momentos para encontrar un estimador para θ.

Capítulo 3

Estimación por intervalos

Para estimar parámetros poblacionales, suele utilizarse dos tipos de estimaciones: la puntual y la de intervalo.

En los capítulos anteriores solo se han realizado estimaciones puntuales de parámetros como el promedio o la variancia. Pero no se han trabajado algunos otros aspectos como el posible tamaño del error en la estimación. Se podría complementar un estimador puntual $\hat{\theta}$ de θ con el tamaño de la muestra y el valor de $var(\theta)$ o con alguna otra información acerca de la distribución muestral de $\hat{\theta}$.

Con el fin de conocer acerca del tamaño del error, puede usarse la llamada **estimación por intervalo**. Una estimación por intervalo de θ es un intervalo de la forma $\hat{\theta_1} < \theta < \hat{\theta_2}$, donde $\hat{\theta_1}$ y $\hat{\theta_2}$ son valores tales que:

$$\mathcal{P}(\hat{\theta_1} < \theta < \hat{\theta_2}) = 1 - \alpha$$

Para alguna probabilidad específica $1 - \alpha$. El valor específico de $1 - \alpha$ se refiere a que el intervalo $\hat{\theta_1} < \theta < \hat{\theta_2}$ es un intervalo de confianza del $(1 - \alpha)$ 100% para θ.

También, $1 - \alpha$ se le llama grado de confianza del intervalo y los puntos del intervalo, $\hat{\theta_1}$ y $\hat{\theta_2}$ se conocen como límite inferior y superior del intervalo, respectivamente. Es decir, puede controlarse la precisión (amplitud) del intervalo, así como la confiabilidad (confianza), de que el verdadero valor del parámetro se encuentre contenido en dicho intervalo.

Frecuentemente el nivel de confianza se expresa como un porcentaje, de tal manera que típicamente se interpreta el intervalo de confianza como: *con una confianza del* $(1 - \alpha) \cdot 100\%$*,* θ *está contenido en el intervalo* $[L_i(X), L_s(X)]$. En este caso, confianza se refiere al proceso usado para construir el intervalo, no al intervalo en sí. Es decir, si se hubiera podido hacer un infinito número de muestras, $(1 - \alpha) \cdot 100\%$ de ellas deberían contener a θ.

3.1. Intervalo de confianza para la estimación de la media poblacional

3.1.1. Intervalo de confianza para la estimación de la media poblacional con variancia poblacional σ_X^2 conocida

Teorema 3.1 Sea $\bar{X}$ el promedio de una variable aleatoria X de tamaño n de una población normal con variancia conocida σ_X^2. Entonces hay una probabilidad $1-\alpha$ de que el error en la estimación del promedio poblacional sea menor que $z_{\frac{\alpha}{2}} \cdot \dfrac{\sigma_X}{\sqrt{n}}$, es decir,

$$P\left(\mid \bar{X} - \mu_X \mid < z_{\frac{\alpha}{2}} \cdot \frac{\sigma_X}{\sqrt{n}}\right) = 1-\alpha$$

Definición 3.1 El margen de error de un intervalo de confianza simétrico es la distancia medida desde el estimador hasta el final (principio) del intervalo. El intervalo de confianza puede expresarse de la siguiente forma: estimador $\pm$ margen de error, de modo que el margen de error E está dado por:

$$E = z_{\frac{\alpha}{2}} \cdot \frac{\sigma_X}{\sqrt{n}}$$

Si la población es finita de tamaño N, entonces el error en la estimación viene dado por:

$$E = z_{\frac{\alpha}{2}} \cdot \frac{\sigma_X}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

Ejemplo 3.1 Un equipo intenta determinar el tiempo promedio que tarda una muestra de 150 personas en trasladarse al trabajo. Por experiencia se considera que $\sigma_X = 6,2$ minutos para este tipo de datos. Determine, con una probabilidad del 99%, el error máximo en las estimaciones del tiempo promedio.

Solución

```
> n = 150; sigma = 6.2; alpha = 0.01; alpham = alpha/2
> zalpham = qnorm(1 - alpham)
> (E = zalpham*sigma/sqrt(n))

[1] 1.303957
```

Con una confianza del 99%, el error máximo que puede cometerse en la estimación del tiempo promedio real que tardan las personas en trasladarse al trabajo es de 1,3 minutos.

Para entender mejor el concepto de margen de error, puede utilizarse el proceso de se realizará una simulación con ayuda del software R.

Ejemplo 3.2 Suponga que desea construirse 100 intervalos de confianza, provenientes de variables aleatorias X, con $X \sim N(\mu_X = 25,\ \sigma_X^2 = 16)$ y $n = 30$.

Solución

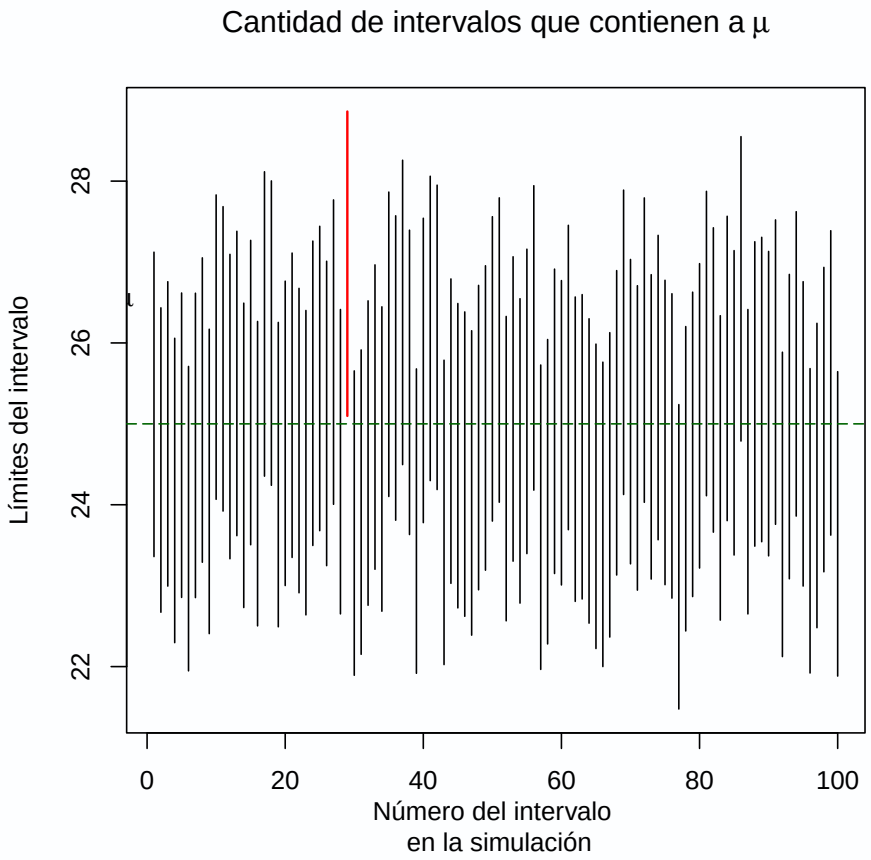


Figura 3.1: Intervalos de confianza para la variable $X \sim N(\mu_X = 25,\ \sigma_X = 4)$ con $n = 30$

Fuente: Elaboración propia

La Figura 3.1 muestra el resultado de la simulación, en la cual se identifica que de los 100 intervalos contruidos, existe uno que no contiene el valor promedio real, es decir, el 99% de los intervalos contienen el verdadero valor promedio, es decir 25.

Teorema 3.2 Sea $\bar{x}$ el promedio de una variable aleatoria X de tamaño n de una población normal con variancia conocida σ_X^2 entonces:

$$CI_{1-\alpha}(\mu_{\bar{X}}) = \left[\bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma_X}{\sqrt{n}},\ \ \bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma_X}{\sqrt{n}}\right] \tag{3.1}$$

es un intervalo del $(1-\alpha)100\%$ de confianza para el promedio de la población.

Para obtener un intervalo de confianza, con un nivel del $(1-\alpha)$, es necesario construir una región, tal que el área entre $-z_{\frac{\alpha}{2}}$ y $z_{\frac{\alpha}{2}}$ sea $(1-\alpha)$, en otras palabras:

$$P\left(-z_{\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} \leq z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

donde $z_{\frac{\alpha}{2}}$ puede definirse como $z_{\frac{\alpha}{2}} = 100\left(1 - \frac{\alpha}{2}\right)$ percentil de la distribución normal estándar y

$$\frac{\bar{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} = Z \sim N(0,1)$$

Ejemplo 3.3 Si una muestra aleatoria con $n = 20$ de una población normal con $\sigma_X^2 = 225$, tiene un promedio muestral de $\bar{x} = 64,3$. Construya e interprete un intervalo del 95% de confianza para el promedio poblacional $\mu_{\bar{X}}$.

Solución

```
> n = 20; sigma = 15; xbarra = 64.3; alpha = 0.05; alpham = alpha/2
> zalpham = qnorm(1 - alpham)
> round((E = zalpham*sigma/sqrt(n)), 2)

[1] 6.57

> c((Li = xbarra - E), (Ls = xbarra + E))

[1] 57.72608 70.87392
```

El cálculo también puede desarrollarse utilizando la función `zsum.test()` del paquete `BSDA` (Arnholt y Evans, 2017), cuyos argumentos principales son: `mean.x, sigma.x, n.x, conf.level` y `alternative`.

```
> library(BSDA)
> zsum.test(mean.x = 64.3, sigma.x = sqrt(225), n.x = 20, conf.level = 0.95,
+ alternative = "two.side")$conf.int

[1] 57.72608 70.87392
attr(,"conf.level")
[1] 0.95
```

Con una confianza del 95%, el intervalo [57,73, 70,87] contiene el promedio poblacional.

Ejemplo 3.4 Una muestra aleatoria de 40 componentes electrónicos reveló que la vida útil promedio es de 5,6 años. Además, se sabe que la desviación estándar poblacional es de 2,1 años.

a) Construya e interprete un intervalo de confianza del 98% para estimar la vida útil promedio real de los componentes.

b) Construya e interprete un intervalo de confianza del 95% para estimar la vida útil promedio real de los componentes.

c) Si la longitud del intervalo es de 1,2 años, determine el nivel de confianza utilizado en la contrucción de dicho intervalo.

Solución

a) Utilizando una confianza de 98%, el intervalo respectivo se obtiene de la siguiente manera:

```
> alpha = 0.02
> zsum.test(mean.x = 5.6, sigma.x = 2.1, n.x = 40, conf.level = 0.98,
+ alternative = "two.side")$conf.int

[1] 4.827561 6.372439
attr(,"conf.level")
[1] 0.98
```

Con una confianza del 98%, el intervalo [4,83, 6,37] contiene el tiempo de vida útil promedio real de los componentes electrónicos.

b) Si la confianza es del 95% se tiene que:

```
> alpha = 0.05
> zsum.test(mean.x = 5.6, sigma.x = 2.1, n.x = 40, conf.level = 0.95,
+ alternative = "two.side")$conf.int

[1] 4.949215 6.250785
attr(,"conf.level")
[1] 0.95
```

Con una confianza del 95%, el intervalo [4,95, 6,25] contiene el tiempo de vida útil promedio real de los componentes electrónicos.

c) Para obtener el nivel de confianza utilizado para construir un intervalo de longitud de 1,2 años se procede de la siguiente manera:

```
> n = 40; sigma = 2.1; xbarra = 5.6; alpha = 0.02;
> longitud <- 1.2
> E <- longitud/2
> zalpham = E*sqrt(n)/sigma
> alpham <- 1 - pnorm(zalpham)
> alpha <- 2*alpham
> confianza <- 1 - alpha; confianza

[1] 0.9292402
```

3.1.2. Intervalo de confianza para la estimación de la media poblacional con variancia poblacional σ_X^2 desconocida

Teorema 3.3 Si $X_i \sim N(\mu_X,\ \sigma_X^2)$, $i = 1, 2, \cdots, n$ con σ_X desconocida y $n < 30$, entonces un intervalo de confianza para $\mu_{\bar{X}}$ está dado por:

$$CI_{1-\alpha}(\mu_{\bar{X}}) = \left[\bar{x} - t_{\frac{\alpha}{2},n-1} \cdot \frac{s}{\sqrt{n}},\ \bar{x} + t_{\frac{\alpha}{2},n-1} \cdot \frac{s}{\sqrt{n}}\right] \tag{3.2}$$

Se dice que CI es un intervalo de confianza del $(1-\alpha)100\%$ para el promedio de la población.

Ejemplo 3.5 Una diseñadora industrial quiere determinar el promedio de tiempo que le lleva a una persona colaboradora ensamblar un juguete. La Tabla 3.1 expresa la cantidad de tiempo, en minutos, que tardó una muestra aleatoria de 36 personas en ensamblar un juguete en particular. Construya un intervalo de confianza del 98% para estimar el promedio poblacional. Asuma que el tiempo de ensamblaje tiene una distribución normal.

Tabla 3.1: Tiempo que les toma a quienes trabajan en la fábrica armar un juguete.

1	2	3	4	5	6	7	8	9	10	11	12
17	13	18	19	17	21	29	22	16	28	21	15
26	23	24	20	8	17	17	21	32	18	25	22
16	10	20	22	19	14	30	22	12	24	28	11

Solución

```
> datos<-c(17,13,18,19,17,21,29,22,16,28,21,15,26,23,
+          24,20,18,17,17,21,32,18,25,22,16,10,20,22,
+          19,14,30,22,12,24,28,11)
> n = length(datos); s = sd(datos); xbarra= mean(datos); alpha = 0.02;
> alpham = alpha/2
> talpham = qt(1 - alpham, n - 1)
> E = talpham*s/sqrt(n)
> c((Li = xbarra- E), (Ls = xbarra+ E))

[1] 18.01231 22.37658
```

Con una confianza del 98%, el tiempo promedio real que le lleva a una persona colaboradora ensamblar un juguete está entre 18,01 y 22,38 minutos.

Cuando se cuenta con los datos de la muestra, el intervalo de confianza puede construirse directamente en R con ayuda de la función `t.test()`.

```
> t.test(datos, conf.level = 0.98)$conf

[1] 18.01231 22.37658
attr(,"conf.level")
[1] 0.98
```

Ejemplo 3.6 Suponga que el nivel de colesterol total en la sangre se distribuye normalmente. Una persona quiere determinar el nivel de colesterol total promedio en una población costarricense. Para ello selecciona 12 personas y obtiene que el nivel de colesterol promedio es de $\bar{x} = 214,1$ minutos con $s = 49,6$. Construya e interprete un intervalo de confianza del 95% para estimar el promedio poblacional.

Solución

```
> n = 12; s = 49.6; xbarra= 214.1; alpha = 0.05; alpham = alpha/2
> talpham = qt(1 - alpham, n - 1)
> E = round(talpham*s/sqrt(n), 2)
> c((Li = xbarra- E), (Ls = xbarra + E))

[1] 182.59 245.61
```

En el caso de la variancia poblacional no conocida, también puede utilizarse la función `TTestA()` del paquete `DescTools` (Signorell et al., 2021), la cual incluye argumentos como `mx, sx, nx, conf.level` y `alternative`.

```
> library(DescTools)
> int <- TTestA(mx = 214.1, sx = 49.6, nx = 12, conf.level = 0.95)$conf.int
> int

[1] 182.5857 245.6143
attr(,"conf.level")
[1] 0.95
```

Con una confianza del 95%, el intervalo [182,59, 245,61] contiene el nivel de colesterol promedio real de la población.

Ejemplo 3.7 Suponga que la estatura del estudiantado de noveno año del Liceo de Heredia sigue una distribución normal. Para conocer la estatura promedio se selecciona una muestra de 25 estudiantes y se encontró que la media de la muestra es de 1,63 metros con una desviación estándar muestral de 5,84 centímetros. Construya un intervalo de confianza del 90% para la estatura media del estudiantado de noveno año de dicha institución.

Solución

```
> library(DescTools)
> int <- TTestA(mx = 1.63, sx = 0.0584, nx = 25, conf.level = 0.9)$conf.int
> int

[1] 1.610017 1.649983
attr(,"conf.level")
[1] 0.9
```

Con una confianza del 90%, el intervalo [1,61, 1,65] contiene la estatura promedio real del estudiantado de noveno año.

Ejemplo 3.8 La directora de una escuela desea determinar el tiempo promedio de traslado de sus estudiantes a la institución. Una muestra aleatoria de 20 estudiantes reveló un tiempo promedio de llegada de 34 minutos con desviación estándar de 6,9 minutos. Construya un intervalo de confianza del 98% para estimar el promedio poblacional, considerando que el tiempo de traslado tiene una distribución normal.

Solución

```
> library(DescTools)
> int <- TTestA(mx = 34, sx = 6.9, nx = 20, conf.level = 0.98)$conf.int
> int

[1] 30.08186 37.91814
attr(,"conf.level")
[1] 0.98
```

Con una confianza del 98%, el intervalo [30,08, 37,92] contiene el tiempo de traslado promedio real del estudiantado.

Ejemplo 3.9 Suponga que el ICE necesita desarrollar un proyecto hidroeléctrico en la zona sur del país, pero necesita contar con personal de amplia experiencia en varias áreas. El número de personas capacitadas para realizar el trabajo es de 820 y se quiere saber cuánto tiempo llevan trabajando en el puesto que actualmente ocupan, para ello se toma una muestra de 50 personas y se obtiene que en promedio tienen 6,7 años de experiencia en sus puestos con una desviación estándar muestral de 0,9 años. Se sabe que los años de experiencia se distribuyen normalmente. Calcule una estimación de intervalo de los años de experiencia en el puesto de las personas colaboradoras capacitadas, de modo que se pueda tener una confianza del 95% en la media de la población.

Solución

```
> N=820; n = 50; s = 0.9; xbarra = 6.7; alpha = 0.05; alpham = alpha/2
> talpham = qt(1 - alpham, n - 1)
> E = round(talpham*(s/sqrt(n))*sqrt((N - n)/(N - 1)), 2)
> c((Li = xbarra - E), (Ls = xbarra + E))

[1] 6.45 6.95
```

Con una confianza del 95%, el intervalo [6,45, 6,95] contiene los años de experiencia promedio real de las personas colaboradoras.

3.2. Intervalos de confianza para una proporción

Dado que el estimador máximo verosímil de P es $\hat{p}$, el cual tiene la ventaja de ser asintótico cuando $n \to \infty$, es decir,

$$\hat{p} \sim N\left(P,\ \frac{P(1-P)}{n}\right)$$

3.2.1. Intervalo de confianza de Wald para una proporción

Sabiendo que el error estándar de $\hat{p}$ viene dado por

$$\hat{\sigma}_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

cuando la población es infinita, entonces un intervalo de confianza de $100(1-\alpha)\%$ para la proporción poblacional, dadas estas condiciones, viene dado por:

$$CI_{1-\alpha}(P) = \left[\hat{p} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] \tag{3.3}$$

En caso de contar con poblaciones finitas de tamaño N, el intervalo de confianza de $100(1-\alpha)\%$ viene dado por:

$$CI_{1-\alpha}(P) = \left[\hat{p} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \cdot \sqrt{\frac{N-n}{N-1}},\ \hat{p} + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \cdot \sqrt{\frac{N-n}{N-1}}\right] \tag{3.4}$$

Ejemplo 3.10 Suponga que se requiere seleccionar una muestra de tamaño 210, para estimar la proporción de estudiantes que están satisfechos con su carrera. Sea X el número de estudiantes satisfechos con la carrera que eligieron. Si en la muestra $x = 130$ estudiantes están satisfechos con la carrera elegida, construya e interprete un intervalo de confianza del 95% para la proporción de estudiantes satisfechos con su carrera.

Solución

Una forma para estimar el intervalo de Wald es la que se muestra a continuación.

```
> x <- 130; n <- 210; p1 <- x / n
> alpha <- 0.05; alpham <- alpha/2
> E <- qnorm(1 - alpham)*sqrt(p1*(1 - p1)/n)
> c((Li = round(p1 - E, 4)), (Ls = p1 + E))

[1] 0.553400 0.684728
```

El intervalo de Wald puede estimarse de manera directa utilizando la función `wald.ci()` del paquete `fastR2` (Pruim, 2018).

```
> library(mosaic)
> library(fastR2)
> wald.ci(x = 130, n = 210)

[1] 0.5533672 0.6847280
attr(,"conf.level")
[1] 0.95
```

O bien, puede utilizarse la función `ciAAllx()` del paquete `proportion` (Subbiah y Rajeswaran, 2017), cuyos argumentos son `x, n, alp, h`. El intervalo de Wald puede observarse en la fila 1 de la salida de R.

```
> library(proportion)
> ciAAllx(x = 130, n = 210, alp = 0.05, h = 0)[1,]

    method   x LowerLimit UpperLimit LowerAbb UpperAbb ZWI
1 Adj-Wald 130  0.5533672   0.684728       NO       NO  NO
```

Con una confianza del 95%, la proporción real de estudiantes satisfechos con su carrera se encuentra entre 55,34% y 68,47%.

No obstante, debido a su baja precisión su uso no es recomendado a menos que se cuente con muestras lo suficientemente grandes y que $n\hat{p} > 5$ y $n(1-\hat{p}) > 5$. Aunque el intervalo de Wald es poco preciso, su uso es frecuente y su precisión mejora al realizar la conocida corrección de Yates.

3.2.2. Corrección de Yates

Considerando la Expresión 3.3 propuesta por Wald para un tamaño de muestra n, el intervalo de confianza de $100(1-\alpha)\%$ para la proporción muestral $\hat{p}$, luego de aplicar la corrección de Yates, viene dado por:

$$\hat{p} \pm \left(z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}} + \frac{1}{2n} \right)$$

Ejemplo 3.11 Considere los datos del Ejemplo 3.10 y construya e interprete un intervalo de confianza del 95% para la proporción real de estudiantes satisfechos con la carrera elegida utilizando la corrección de Yates.

Solución

```
> x = 130; n = 210; p1 <- x/n
> alpha <- 0.05; alpham <- alpha/2
> E <- qnorm(1 - alpham)*sqrt((p1*(1 - p1)/n))
> c((Li = p1 - E - 1/(2*n)), (Ls = p1 + E + 1/(2*n)))

[1] 0.5509863 0.6871090
```

Con una confianza del 95%, la proporción real de estudiantes satisfechos con su carrera se encuentra entre 55,1% y 68,71%.

La corrección de Yates para el intervalo de Wald puede obtenerse directamente con ayuda de la función `ciCAllx()`, utilizando el argumento `c = 1/(2*n)`. Esta corrección puede observarse en la fila 1 de la salida de R.

```
> # library(proportion) es requerida
> ciCAllx(x = 130, n = 210, alp = 0.05, c = 1/(2*210))[1,]

  method   x LowerLimit UpperLimit LowerAbb UpperAbb ZWI
1   Wald 130  0.5509863   0.687109       NO       NO  NO
```

3.2.3. Intervalos de confianza de Wilson para una proporción

Sea X una variable aleatoria binomial tal que $X \sim Bin(n, P)$ y que $\hat{p} = X/n$ es un estimador insesgado de P. Además, para valores "grandes" de n la variable z se comporta, aproximadamente, como una normal estándar, cuando:

$$z = \frac{\hat{p} - P}{\sqrt{\frac{P(1-P)}{n}}} \tag{3.5}$$

La Ecuación 3.5 indica que:

$$P\left(-z_{\frac{\alpha}{2}} \leq \frac{\hat{p}-P}{\sqrt{\frac{P(1-P)}{n}}} \leq z_{\frac{\alpha}{2}}\right) \approx 1-\alpha \tag{3.6}$$

Considerando una de las igualdades en la Ecuación 3.6 se tiene que:

$$-z_{\frac{\alpha}{2}} = \frac{\hat{p}-P}{\sqrt{\frac{P(1-P)}{n}}} \tag{3.7}$$

Al despejar P de la Ecuación 3.7 se obtiene:

$$\left(\frac{n}{z^2_{\frac{\alpha}{2}}}+1\right)P^2 - \left(\frac{2n\hat{p}}{z_{\frac{\alpha}{2}}}\right)P + \frac{n\hat{p}^2}{z_{\frac{\alpha}{2}}} = 0$$

Resolviendo para P, se obtiene un intervalo de confianza del 95%, donde:

$$li = \frac{\hat{p}+z^2_{\frac{\alpha}{2}}/(2n) - z_{\frac{\alpha}{2}}\sqrt{\hat{p}(1-\hat{p})/n + z^2_{\frac{\alpha}{2}}/(4n^2)}}{1+z^2_{\frac{\alpha}{2}}/n} \tag{3.8}$$

$$ls = \frac{\hat{p}+z^2_{\frac{\alpha}{2}}/(2n) + z_{\frac{\alpha}{2}}\sqrt{\hat{p}(1-\hat{p})/n + z^2_{\frac{\alpha}{2}}/(4n^2)}}{1+z^2_{\frac{\alpha}{2}}/n} \tag{3.9}$$

Teorema 3.4 Considere una muestra aleatoria de tamaño n de una población donde el parámetro P, $0 < P < 1$, indica la verdadera proporción de éxitos de una cierta característica binaria. Sea X la cantidad de éxitos en la muestra n y $\hat{p} = X/n$ la proporción muestral. Entonces un intervalo de confianza aproximado de $(1-\alpha)100\%$ para P es $(li,\ ls)$, con:

$$li = \frac{\hat{p}+z^2_{\frac{\alpha}{2}}/(2n) - z_{\frac{\alpha}{2}}\sqrt{\hat{p}(1-\hat{p})/n + z^2_{\frac{\alpha}{2}}/(4n^2)}}{1+z^2_{\frac{\alpha}{2}}/n}$$

$$ls = \frac{\hat{p}+z^2_{\frac{\alpha}{2}}/(2n) + z_{\frac{\alpha}{2}}\sqrt{\hat{p}(1-\hat{p})/n + z^2_{\alpha/2}/(4n^2)}}{1+z^2_{\frac{\alpha}{2}}/n}$$

Este intervalo de confianza es llamado intervalo de confianza de puntaje del $(1-\alpha)100\%$ para la proporción P.

Ejemplo 3.12 Considere los datos del Ejemplo 3.10 y construya e interprete un intervalo de confianza del 95% para la proporción real de estudiantes satisfechos con la carrera elegida utilizando el método de Wilson.

Solución

```
> x <- 130; n <- 210; p <- x/n
> alpha <- 0.05; alpham <- alpha/2; zalpham <- qnorm(1 - alpham)
> num_i <- p + zalpham^2/(2*n) - zalpham*sqrt(p*(1 - p)/n + zalpham^2/(4*n^2))
> num_s <- p + zalpham^2/(2*n) + zalpham*sqrt(p*(1 - p)/n + zalpham^2/(4*n^2))
> den <- 1 + zalpham^2/n
> int <- c(li <- num_i/den, ls <- num_s/den); int

[1] 0.5517861 0.6820320
```

Con una confianza del 95%, la proporción real de estudiantes satisfechos con su carrera se encuentra entre 55,18% y 68,2%.

El intervalo de confianza de Wilson puede obtenerse en R con ayuda de la función `prop.test()` indicando el argumento `correct = F`.

```
> x <- 130; n <- 210; p <- x/n
> prop.test(x = 130, n = 210, conf.level = 0.95, correct = F,
+           alternative = "two.sided")$conf.int

[1] 0.5517861 0.6820320
attr(,"conf.level")
[1] 0.95
```

O bien, puede utilizarse la función `ciAAllx()` indicando el argumento `h = 0`. El intervalo de Wilson puede observarse en la fila 4 de la salida de R que indica el método `Adj-Score`.

```
> ciAAllx(x = 130, n = 210, alp = 0.05, h = 0)[4,]

     method   x LowerLimit UpperLimit LowerAbb UpperAbb ZWI
4 Adj-Score 130  0.5517861   0.682032       NO       NO  NO
```

Ejemplo 3.13 En una encuesta realizada a 846 personas sobre la situación económica del país, la proporción de personas que están de acuerdo con las medidas adoptadas por el gobierno es alrededor del 45%. Sea X la cantidad de personas que simpatizan con las medidas del gobierno

en una muestra de tamaño n, construya e interprete un intervalo de confianza del 95% para la proporción real de personas que están de acuerdo con las medidas económicas adoptadas por el gobierno.

Solución

```
> n <- 846; p1 <- 0.45; alpha <- 0.05; alpham <- alpha/2; x <- 846*0.45
> int <- prop.test(x, n, conf.level=1 - alpha, correct=F)$conf.int
> int

[1] 0.4167775 0.4836745
attr(,"conf.level")
[1] 0.95
```

Con una confianza del 95%, la proporción real de personas que están de acuerdo con las medidas económicas adoptadas por el gobierno se encuentra entre 41,68% y 48,37%.

El intervalo de Wilson también mejora su precisión al utilizar la correción de Yates. Para realizar esta corrección puede utilizarse la función `ciCAllx()` indicando el argumento `c = 1/(2*n)`. El intervalo ajustado puede observarse en la fila 3 de la salida de R.

Ejemplo 3.14 Considere el Ejemplo 3.12, construya e interprete un intervalo de confianza del 95% para la proporción real de estudiantes satisfechos con la carrera elegida utilizando la corrección de Yates.

Solución

```
> ciCAllx(x = 130, n = 210, alp = 0.05, c = 1/(2*210))[3,]

  method   x LowerLimit UpperLimit LowerAbb UpperAbb ZWI
3  Score 130   0.549372  0.6842925       NO       NO  NO
```

También puede utilizarse la función `prop.test()` indicando el argumento `correct = T`.

```
> x <- 130; n <- 210; p <- x/n; alpha = 0.05
> intervalo <- prop.test(x=130, n=210, conf.level=0.95, correct=T)$conf
> intervalo

[1] 0.5493720 0.6842925
attr(,"conf.level")
[1] 0.95
```

Con una confianza del 95%, la proporción real de estudiantes satisfechos con su carrera se encuentra entre 54,94% y 68,43%.

3.2.4. Intervalo de confianza de Agresti-Coull para una proporción

La fórmula para calcular intervalos de confianza de Wilson puede resultar tediosa, a no ser que se cuente con un software que facilite la estimación. Agresti y Coull (1998) proponen una forma simple de aproximar un intervalo de confianza para P, de la siguiente manera:

Si X expresa el número de éxitos en una muestra de tamaño n, y sean $\widetilde{X} = X + 2$, $\widetilde{n} = n + 4$ y $\widetilde{p} = \dfrac{\widetilde{X}}{\widetilde{n}}$. Entonces una aproximación para un intervalo de confianza del $(1-\alpha)100\%$ para P está dado por:

$$CI_{1-\alpha}(P) = \left[\widetilde{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\widetilde{p}(1-\widetilde{p})}{\widetilde{n}}},\ \widetilde{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\widetilde{p}(1-\widetilde{p})}{\widetilde{n}}}\right] \tag{3.10}$$

Ejemplo 3.15 Considere los datos del Ejemplo 3.10 y construya e interprete un intervalo de confianza del 95% de Agresti-Coull, para la proporción de estudiantes satisfechos con su carrera.

Solución

Para este caso $\widetilde{p} = \dfrac{132}{214} = 0{,}6168$ donde el intervalo de confianza solicitado es:

Solución

```
> x <- 130; n <- 210; p1 <- (x + 2) / (n + 4)
> alpha <- 0.05; alpham <- alpha/2
> E <- qnorm(1 - alpham)*sqrt(p1*(1 - p1)/(n + 4))
> c((Li = p1 - E), (Ls = p1 + E))

[1] 0.5516864 0.6819585
```

Con una confianza del 95%, la proporción real de estudiantes satisfechos con su carrera se encuentra entre 55,17% y 68,2%.

3.3. Intervalos de confianza para la varianza de una población normal

Considere una variable aleatoria X de una población normal tal que $X \sim N(\mu_X, \sigma_X^2)$, de la cual se toma una muestra aleatoria de tamaño n. El intervalo de confianza para σ_X^2 está basado en el hecho de que:

$$\frac{(n-1)s^2}{\sigma_X^2} \sim \chi^2_{n-1}$$

Un intervalo de confianza de $(1-\alpha)100\%$ está dado por:

$$CI_{1-\alpha}(\sigma_X^2) = \left[\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2};\, n-1}},\ \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2};\, n-1}}\right] \tag{3.11}$$

Es importante resaltar que el supuesto de normalidad de la muestra debe ser validado antes de construir el intervalo.

Ejemplo 3.16 Suponga que para 16 personas el consumo de gasolina del motor de sus vehículos muestran una desviación estándar de 2,2 galones. Suponiendo que el consumo de gasolina se comporta normalmente, construya e interprete un intervalo de confianza del 99% para σ_X^2, que estime la verdadera variabilidad del consumo de gasolina de los motores.

Solución

```
> n = 16; s = 2.2
> alpha <- 0.01; alpham <- alpha/2
> Li <- ((n-1)*s^2)/qchisq(p = 1 - alpham, df = n - 1)
> Ls <- ((n-1)*s^2)/qchisq(p = alpham, df = n - 1)
> c(Li, Ls)

[1]  2.213326 15.779468
```

Con una confianza del 99%, la variabilidad real en el consumo de combustible del motor está entre 2,21 y 15,78 galones.

Ejemplo 3.17 Suponga que para un grupo de 30 personas seleccionadas aleatoriamente se obtiene la desviación estándar $s = 4,52$ respecto a la nota promendio alcanzada en un curso de Estadística. Construya e interprete un intervalo de confianza del 96% para la variancia poblacional.

Solución

```
> n = 30; s = 4.52
> alpha <- 0.04; alpham <- alpha/2
> sup <- qchisq(alpham, df = n - 1, lower.tail = T)
> inf <- qchisq(p = 1 - alpham, df = n - 1, lower.tail = T)
> c(Li <- (n - 1)*s^2/inf, Ls <- (n - 1)*s^2/sup)

[1] 12.68896 38.04181
```

Con una confianza del 96%, la variabilidad real de la nota promedio en el curso de Estadística está entre 12,69 y 38,04.

3.4. Intervalos de confianza para el cociente de varianzas de dos poblaciones normales

Se va a considerar la contrucción de intervalos de confianza para σ_X^2/σ_Y^2, donde hay dos poblaciones normales e independientes, $N(\mu_X, \sigma_X^2)$ y $N(\mu_Y, \sigma_Y^2)$, de las cuales se toman muestras de tamaño n_X y n_Y, respectivamente. Por lo general, se busca verificar si el 1 se encuentra en el intervalo, lo que indicaría que las variancias son iguales.

Para construir el intervalo de confianza para σ_X^2/σ_Y^2, se usa un teorema que dice que: si se tienen dos muestras aleatorias $X_1, X_2, \ldots, X_{n_X}$ y $Y_1, Y_2, \ldots, Y_{n_Y}$, que son tomadas de poblaciones normales independientes, es decir, $X \sim N(\mu_X, \sigma_X^2)$ y $Y \sim N(\mu_Y, \sigma_Y^2)$, entonces la variable aleatoria:

$$\frac{\dfrac{s_Y^2}{\sigma_Y^2}}{\dfrac{s_X^2}{\sigma_X^2}} \sim F_{n_Y-1,\ n_X-1}$$

Por lo que el intervalo de confianza del $(1-\alpha)100\%$ para $\dfrac{\sigma_X^2}{\sigma_Y^2}$ está dado por:

$$CI_{1-\alpha}\left(\frac{\sigma_X^2}{\sigma_Y^2}\right) = \left[f_{\frac{\alpha}{2};\ n_Y-1,\ n_X-1} \cdot \frac{s_X^2}{s_Y^2}, f_{1-\frac{\alpha}{2};\ n_Y-1,\ n_X-1} \cdot \frac{s_X^2}{s_Y^2}\right] \tag{3.12}$$

Donde $f_{\frac{\alpha}{2};\ n_Y-1,\ n_X-1}$ representa el $100 \cdot \frac{\alpha}{2}$ percentil de la distribución F con $n_Y - 1$ y $n_X - 1$ grados de libertad y $f_{1-\frac{\alpha}{2};\ n_Y-1,\ n_X-1}$ representa el $100\left(1-\frac{\alpha}{2}\right)$ percentil de dicha distribución.

Ejemplo 3.18 Se realiza un estudio para determinar si los colegios urbanos tenían mejor rendimiento en las pruebas de bachillerato que los colegios rurales. Se decidió comparar el rendimiento en matemáticas para una muestra de 11 colegios urbanos y 5 rurales, suponiendo que las muestras son independientes y siguen una distribución normal. La desviación estándar para los colegios urbanos fue de 18 puntos y para los rurales de 11 puntos. Elabore un intervalo de confianza del 95% para la razón de las variancias de las dos poblaciones.

Solución

```
> n_x <- 11; n_y <- 5; s_x <- 18; s_y <- 11; alpha <- 0.05; alpham <- alpha/2
> i <- qf(p = 0.025, df1 = n_x - 1, df2 = n_y - 1, lower.tail = T)
> s <- qf(p = 0.975, df1 = n_x - 1, df2 = n_y - 1, lower.tail = T)
> c(Li <- i*(18^2/11^2), Ls <- s*(18^2/11^2))

[1]  0.5992572 23.6811358
```

Con una confianza del 95%, la razón entre las varianzas de las dos poblaciones se encuentra entre 0,6 y 23,68. Como el intervalo incluye al uno, entonces con una confianza del 95%, puede concluirse que las varianzas de las poblaciones son iguales.

O bien,

```
> n_x <- 5; n_y <- 11; s_x <- 11; s_y <- 18; alpha <- 0.05; alpham <- alpha/2
> i <- qf(p = 0.025, df1 = n_x - 1, df2 = n_y - 1, lower.tail = T)
> s <- qf(p = 0.975, df1 = n_x - 1, df2 = n_y - 1, lower.tail = T)
> c(Li <- i*(s_x^2/s_y^2), Ls <- s*(s_x^2/s_y^2))

[1] 0.0422277 1.6687325
```

Con una confianza del 95%, la razón entre las varianzas de las dos poblaciones se encuentra entre 0,04 y 1,67. Como el intervalo incluye al uno, entonces con una confianza del 95%, puede concluirse que las varianzas de las poblaciones son iguales.

Cuando se cuenta con los datos de la muestra, el intervalo de confianza para la razón de varianzas puede obtenerse directamente en R con ayuda de la función `var.test()`.

Ejemplo 3.19 Se realizó un estudio para determinar si existen diferencias en el rendimiento en el curso de Estadística en las universidades públicas y en las universidades privadas. A continuación, la Tabla 3.2 presenta las notas promedio en el curso de Estadística pertenecientes a estudiantes procedentes de ambos tipos de universidades. Elabore un intervalo de confianza del 97% para la razón de las variancias de las dos poblaciones.

Tabla 3.2: Notas del curso de Estadística según el tipo de universidad

	1	2	3	4	5	6	7	8	9	10
Público	67	60	70	76	40	72	52	51	73	44
Privado	93	82	75	74	45	59	59	94	90	46

Solución

```
> pub <- c(67, 60, 70, 76, 40, 72, 52, 51, 73, 44)
> pri <- c(93, 82, 75, 74, 45, 59, 59, 94, 90, 46)
> n_x <- length(pub); n_y <- length(pri); s_x <- sd(pub); s_y <- sd(pri)
> alpha <- 0.03; alpham <- alpha/2
> i <- qf(p = 0.015, df1 = n_x - 1, df2 = n_y - 1, lower.tail = T)
> s <- qf(p = 0.985, df1 = n_x - 1, df2 = n_y - 1, lower.tail = T)
> c(Li <- i*(s_x^2/s_y^2), Ls <- s*(s_x^2/s_y^2))

[1] 0.1032232 2.3122565
```

Al tener los datos de las variables en estudio, el contraste puede realizarse de forma directa utilizando la función `var.test()`.

```
> var.test(pub, pri, conf.level = 0.97)$conf.int

[1] 0.1032232 2.3122565
attr(,"conf.level")
[1] 0.97
```

Con una confianza del 97%, la razón entre las varianzas de las dos poblaciones se encuentra entre 0,1 y 2,31. Como el intervalo incluye al uno, entonces con una confianza del 97%, puede concluirse que las varianzas de las poblaciones son iguales.

Ejemplo 3.20 Construya un intervalo de confianza del 95% para el cociente de variancias, considerando los datos proporcionados en la Tabla 3.2 del Ejemplo 3.19.

3.5. Intervalos de confianza para la diferencia de medias de dos poblaciones

Teorema 3.5 Sean X y Y variables aleatorias tal que $X \sim N(\mu_X,\ \sigma_X^2)$ y $Y \sim N(\mu_Y,\ \sigma_Y^2)$ entonces:

$$X - Y \sim N(\mu_X - \mu_Y,\ \sigma_X^2 + \sigma_Y^2)$$

Para muestras de tamaño n_X y n_Y se tiene que:

$$\bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y,\ \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}\right)$$

3.5.1. Las muestras son independientes con variancias poblacionales conocidas e iguales

Teorema 3.6 Sean X y Y variables aleatorias tal que $X \sim N(\mu_X,\ \sigma_X^2)$ y $Y \sim N(\mu_Y,\ \sigma_Y^2)$, con $\sigma_X = \sigma_Y = \sigma$ conocida, entonces para muestras de tamaño $n_X,\ n_Y$ se tiene que:

$$\bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y,\ \sigma^2\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)\right)$$

De esta manera, el error dado en la estimación de las diferencias está dado por

$$E = z_{\frac{\alpha}{2}} \cdot \sigma\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}$$

Por lo tanto, el intervalo de confianza, dadas estas condiciones, estaría definido por

$$CI_{1-\alpha}(\mu_X - \mu_Y | \sigma_X = \sigma_Y = \sigma \text{ pero conocidas}) = [(\bar{x} - \bar{y}) - E,\ (\bar{x} - \bar{y}) + E] \tag{3.13}$$

Ejemplo 3.21 Suponga que se quiere comparar el rendimiento promedio entre dos grupos de un curso particular. Suponga que el rendimiento en el curso se distribuye normalmente. Para cada grupo se tiene que $\sigma_X = \sigma_Y = 3$, donde $n_X = 15$, $\sum_{i=1}^{n_X} x_i = 120$, $n_Y = 22$, $\sum_{i=1}^{n_Y} y_i = 184$, respectivamente. Construya e interprete un intervalo de confianza del 95%, para la diferencia de las medias poblacionales.

Solución

```
> nx = 15; ny = 22; sigma = 3; xbarra = 8; ybarra = 8.40
> alpha = 0.05; alpham = alpha/2
> zalpham = qnorm(1 - alpham)
> diferencia = (xbarra - ybarra)
> E = round(zalpham*sigma*sqrt(1/nx + 1/ny), 2)
> c((Li = diferencia - E), (Ls = diferencia + E))

[1] -2.37  1.57
```

Este intevalo puede obtenerse de forma directa utilizando la función `zsum.test()` del paquete `BSDA` (Arnholt y Evans, 2017).

```
> zsum.test(mean.x = 8, sigma.x = 3, n.x = 15, mean.y = 8.40, sigma.y = 3,
+   n.y = 22, alternative = "two.sided", conf.level = 0.95)$conf.int

[1] -2.368853  1.568853
attr(,"conf.level")
[1] 0.95
```

Con una confianza del 95%, el intervalo $[-2{,}37,\ 1{,}57]$ contiene la diferencia promedio real en la nota del curso entre los dos grupos. Como el intervalo incluye al cero, entonces, con una confianza del 95%, no puede concluirse la existencia de una diferencia en el rendimiento promedio de ambos grupos.

3.5.2. Muestras independientes con variancias desconocidas, pero que se asumen iguales

Es común que las variancias poblacionales no se conocen, en cuyo caso se tendría que aplicar el siguiente teorema:

Teorema 3.7 Sean $X_i \sim N(\mu_X,\ \sigma_Y^2),\ i = 1, 2, \cdots, n_X$ y $Y_i \sim N(\mu_Y,\ \sigma_Y^2),\ j = 1, 2, \cdots, n_Y$, dos muestras aleatorias e independientes con medias y variancias muestrales $\bar{X},\ S_X^2,\ \bar{Y},\ S_Y^2$, respectivamente. Entonces:

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{S_p\sqrt{\dfrac{1}{n_X} + \dfrac{1}{n_Y}}}$$

tiene una distribución t de student con $n_X + n_Y - 2$ grados de libertad.

Por lo tanto, una aproximación para un intervalo de confianza del $(1-\alpha)100\%$ para $\mu_X - \mu_Y$ está dada por:

$$CI_{1-\alpha}(\mu_X - \mu_Y) = (\bar{X} - \bar{Y}) \pm t_{(\frac{\alpha}{2},\, n_Y+n_Y-2)} S_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}} \tag{3.14}$$

donde:

$$S_p = \sqrt{\frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2}}$$

Ejemplo 3.22 Suponga que el nivel de colesterol total en la sangre se distribuye normalmente. En una investigación se desea determinar nivel promedio colesterol por región. Para ello se consideran 12 personas que residen en zona urbana y obtiene que el nivel promedio de colesterol, para esta muestra, es de $\bar{x} = 215,9$ mg/dl con $s_X = 49,8$. Luego considera 16 personas residentes en zona rural y se obtiene que el nivel de colesterol promedio, para esta muestra, es de $\bar{y} = 211,5$ mg/dl con $s_Y = 49,1$. Construya e interprete un intervalo de confianza del 95% para estimar la diferencia poblacional.

Solución

```
> nx = 12; ny = 16; sx = 49.8; sy = 49.1; xbarra = 215.9
> ybarra = 211.5; alpha = 0.05; alpham = alpha/2
> talpham = qt(1 - alpham, nx + ny - 2)
> diferencia = (xbarra - ybarra)
> sp = sqrt(((nx - 1)*sx^2 + (ny - 1)*sy^2)/(nx + ny - 2))
> E = round(talpham*sp*sqrt(1/nx + 1/ny), 2)
> c((Li = diferencia - E), (Ls = diferencia + E))

[1] -34.38  43.18
```

El intervalo de confianza puede obtenerse de manera directa utilizando la función `TTestA()` del paquete `DescTools` (Signorell et al., 2021).

```
> library(DescTools)
> TTestA(mx = 215.9, sx = 49.8, nx = 12, my = 211.5, sy = 49.1, ny = 16,
+        conf.level = 0.95, alternative = "two.sided", var.equal = T)$conf.int

[1] -34.37536  43.17536
attr(,"conf.level")
[1] 0.95
```

Con una confianza del 95%, el intervalo $[-34{,}38,\ 43{,}18]$ contiene la diferencia promedio real en el nivel de colesterol total entre las personas residentes en ambas zonas. Como el intervalo incluye al cero, entonces, con una confianza del 95%, puede concluirse que no existen diferencias en el nivel de colesterol total entre las personas residentes en ambas zonas.

Ejemplo 3.23 Suponga que el salario es una variable aleatoria que se distribuye normalmente. Una persona desea determinar el salario promedio en dos tipos de universidades. Para ello considera que 30 docentes de universidades públicas tienen un salario promedio de 750 000 con $s = 6360,4$. Luego, una muestra de 26 docentes de universidades privadas indican un salario promedio de 711 000 con $s = 6275,5$. Construya e interprete un intervalo de confianza del 96% para estimar la diferencia poblacional.

3.5.3. Muestras pareadas o dependientes

En muchas situaciones se tienen muestras aleatorias que no son independientes, pues las observaciones tienen relación de manera natural o por diseño. A este tipo de muestras también se les conoce como muestras pareadas.

En estos casos, en los cuales se diseña una prueba de diferencia por pares (X_i, Y_i), para todo $i = 1, 2, \ldots, n$, se tiene como supuesto que la distribución de esas diferencias es normal, es decir, si $D = (X_1 - Y_1, X_2 - Y_2, \ldots)$ denota las diferencias de la población y $d = (x_1 - y_1, x_2 - y_2, \ldots, x_n - y_n)$ denota las diferencias muestrales, entonces $D \sim N(\mu_D = \mu_X - \mu_Y, \sigma_D^2)$ y

$$T = \frac{\bar{D} - \mu_D}{\frac{s_D}{\sqrt{n}}}$$

tiene una distribución t de student con $n - 1$ grados de libertad.

Un intervalo de confianza para μ_D, cuando σ_D es desconocida, está dado por:

$$CI_{1-\alpha}(\mu_X - \mu_Y = \mu_D) = \left[\bar{d} - t_{\frac{\alpha}{2},\, n-1}\frac{s_D}{\sqrt{n}}, \bar{d} + t_{\frac{\alpha}{2},\, n-1}\frac{s_D}{\sqrt{n}}\right] \tag{3.15}$$

Este intervalo puede construirse directamente en R utilizando la función `t.test()` indicanco el argumento `paired = T`.

Ejemplo 3.24 Se realiza un experimento para medir el gasto calórico en personas entrenadas, después de una sesión de pesas en circuito y con descanso. Suponga que las diferencias entre ambas poblaciones, se distribuyen aproximadamente normal. Los datos se resumen en la Tabla 3.3. Construya e interprete un intervalo de confianza del 95% para la diferencia de las medias poblacionales.

Tabla 3.3: Experimento de gasto calórico con personas entrenadas según tipo de ejercicio.

	1	2	3	4	5	6
Circuito	106	98	123	97	88	95
Descanso	102	94	118	91	83	90

Solución

```
> circuito <- c(106,98,123,97,88,95)
> descanso <- c(102,94,118,91,83,90)
> t.test(x = circuito, y = descanso, paired = T, conf.level = 0.98)$conf.int

[1] 3.799229 5.867437
attr(,"conf.level")
[1] 0.98
```

Con una confianza del 98%, el intervalo [3,8, 5,87] contiene la diferencia promedio real en el gasto calórico promedio en ambas rutinas. Como el intervalo no incluye al cero, entonces, con una confianza del 98%, puede concluirse que existen diferencias en el gasto calórico promedio en ambas rutinas. Es decir, el gasto calórico es mayor en la rutina de circuito.

Ejemplo 3.25 Un docente realiza una prueba diagnóstica a sus 10 estudiantes del curso de inferencia estadística al inicio del ciclo. Posteriormente, el último día de clases aplica otra prueba equivalente a la primera (no es la misma, pero si muy similar) con el objetivo de verificar si el estudiantado mejoró sus conocimientos de estadística descriptiva, luego de llevar el curso de inferencia. Suponga que la nota en las pruebas son aproximadamente normales. Los datos se resumen en la Tabla 3.4. Construya e interprete un intervalo de confianza del 98% para la diferencia de las medias poblacionales.

Tabla 3.4: Notas en las pruebas de diagnóstico iniciales y finales en el curso de inferencia estadística

	1	2	3	4	5	6	7	8	9	10
P. inicial	40	73	56	87	88	23	63	77	80	85
P. final	45	68	70	93	79	35	59	72	74	91

Solución

```
> inicial <- c(40,73,56,87,88,23,63,77,80,85)
> final <- c(45,68,70,93,79,35,59,72,74,91)
> t.test(x = inicial, y = final, paired = T, conf.level = 0.98)$conf.int

[1] -8.687349  5.887349
attr(,"conf.level")
[1] 0.98
```

3.6. Intervalo de confianza para la diferencia de dos proporciones

En muchas ocasiones es necesario estimar la diferencia entre dos parámetros binomiales P_X y P_Y, basados en muestras aleatorias independientes de tamaños n_X y n_Y. Si el número de éxitos en las muestras son x e y respectivamente, entonces, las proporciones muestrales están dadas por $\hat{p}_X$ y $\hat{p}_Y$ y es posible verificar que:

a) $E(\hat{p}_X - \hat{p}_Y) = P_X - P_Y$

b) $Var(\hat{p}_X - \hat{p}_Y) = \dfrac{P_X(1-P_X)}{n_X} + \dfrac{P_Y(1-P_Y)}{n_Y}$

De esta manera, el siguiente teorema garantiza un procedimiento para aproximar intervalos de confianza para la diferencia de proporciones.

Teorema 3.8 Sean $X \sim Bin(n_X,\ P_X)$ y $Y \sim Bin(n_Y,\ P_Y)$, con $\hat{p}_X = \dfrac{x}{n_X}$ y $\hat{p}_Y = \dfrac{y}{n_Y}$. Para n_X y n_Y suficientemente grandes, se tiene que un intervalo de confianza aproximado del $(1-\alpha)100\%$ para $P_X - P_Y$ está dado por:

$$CI_{1-\alpha}(P_X - P_Y) = \hat{p}_X - \hat{p}_Y \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n_X} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{n_Y}} \tag{3.16}$$

Ejemplo 3.26 Una muestra de 631 estudiantes residentes en zona urbana reveló que 475 están de acuerdo con el pago electrónico en el transporte público del país. Por su parte, en una muestra de 677 estudiantes residentes de zona rural, 424 indicaron estar de acuerdo con esta iniciativa. Construya e interprete un intervalo de confianza del 95%, para la diferencia de las proporciones de estudiantes que están de acuerdo con la medida en ambas poblaciones.

Solución

```
> x <- 475; y <- 424; nx <- 631; ny <- 677
> px <- x/nx; py <- y/ny;
> alpha <- 0.05; alpham <- alpha/2
> E <- qnorm(1 - alpham)*sqrt((px*(1-px)/nx)+(py*(1-py)/ny))
> diferencia <- px - py
> c((Li = diferencia - E), (Ls = diferencia + E))

[1] 0.07687197 0.17608985
```

Con una confianza del 95%, la diferencia real de las proporciones de estudiantes que están de acuerdo con el pago electrónico en ambas poblaciones está entre 7,69% y 17,61%. Como el intervalo no incluye al cero, entonces, con una confianza del 95%, puede concluirse que existe diferencia en la proporción real de estudiantes que están de acuerdo con la implementación del pago electrónico en ambas poblaciones. Es decir, la población rural tiene mayor aceptación por la implementación de dicho servicio.

El intervalo de confianza para la diferencia de proporciones puede obtenerse en R por medio de la función `prop.test()`

```
> x <- 475; y <- 424; nx <- 631; ny <- 677
> intervalo <- prop.test(c(x, y), c(nx, ny), correct = F)$conf;
> intervalo

[1] 0.07687197 0.17608985
attr(,"conf.level")
[1] 0.95
```

Por otra parte, Agresti y Caffo (2000) proponen la siguiente aproximación para intervalos de confianza para la diferencia de proporciones:

$$CI_{1-\alpha}(P_X - P_Y) = \widetilde{p}_X - \widetilde{p}_Y \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\widetilde{p}_X(1-\widetilde{p}_X)}{\widetilde{n}_X} + \frac{\widetilde{p}_Y(1-\widetilde{p}_Y)}{\widetilde{n}_Y}} \tag{3.17}$$

donde $\widetilde{p}_i = \dfrac{X_i + 1}{n_i + 2}$ para $i = 1, 2$.

Ejemplo 3.27 Utilice la aproximación de Agresti y Caffo y construya e interprete un intervalo de confianza del 95%, para la diferencia de las proporciones de estudiantes que están de acuerdo con el pago electrónico en ambas poblaciones según los datos del Ejemplo 3.26.

Solución

```
> x <- 475 + 1; y <- 424 + 1; nx <- 631 + 2; ny <- 677 + 2
> px <- x/nx; py <- y/ny;
> alpha <- 0.05; alpham <- alpha/2
> E <- qnorm(1 - alpham)*sqrt((px*(1-px)/nx)+(py*(1-py)/ny))
> diferencia <- px - py
> c((Li = diferencia - E), (Ls = diferencia + E))

[1] 0.07649083 0.17561767
```

O bien,

```
> x <- 475 + 1; y <- 424 + 1; nx <- 631 + 2; ny <- 677 + 2
> intervalo <- prop.test(c(x, y), c(nx, ny), correct = F)$conf
> intervalo

[1] 0.07649083 0.17561767
attr(,"conf.level")
[1] 0.95
```

Con una confianza del 95%, la diferencia real de las proporciones de estudiantes que están de acuerdo con el pago electrónico en ambas poblaciones está entre 7,65% y 17,56%. Como el intervalo no incluye al cero, entonces, con una confianza del 95%, puede concluirse que existe diferencia en la proporción real de estudiantes que están de acuerdo con la implementación del pago electrónico en ambas poblaciones. Es decir, la población rural tiene mayor aceptación por la implementación de dicho servicio.

Capítulo 4

Contraste de hipótesis con base en una muestra

4.1. Generalidades para una prueba de hipótesis

Los problemas de estimación estadística tienen la finalidad de aproximar un parámetro poblacional a través de los datos revelados por una muestra. Aunque no se tenga idea del valor real del parámetro, usualmente se trata de indagar lo más que se pueda. Es por ello que, cuando se logra tener alguna idea acerca de este valor, se plantea una proposición que permita contrastarlo con ayuda de la información muestral que se posee. En este procedimiento se plantean dos ideas o proposiciones, la que propone la persona que realiza la investigación y la negación de la proposición planteada. Estas proposiciones que se establecen respecto a uno o varios parámetros poblacionales se denominan hipótesis. Ambas hipótesis plantean afirmaciones relacionadas con el parámetro en estudio y lo que se busca es verificar cuál de ellas presenta la afirmación correcta, para ello se recurre a los datos revelados por una muestra.

De esta manera la hipótesis que plantea una igualdad y en consideración es sometida a prueba, recibe el nombre de **hipótesis nula** y se denota por H_0, mientras que la que plantea una negación de la hipótesis nula se denomina **hipótesis alternativa**, la cual se denota por H_1. La hipótesis alternativa constituye un planteamiento complementario a lo planteado en alternativa H_0 y en consecuencia, representa la afirmación que se considera aceptable en caso de que la información muestral revele que la hipótesis nula no sea considerada como verdadera. Dado un parámetro θ, existen tres formas distintas de plantear un contraste de hipótesis:

Tabla 4.1: Contraste de hipótesis para un parámetro θ

Hipótesis Nula	Hipótesis Alternativa
$H_0 : \theta = \theta_0$	$H_1 : \theta \neq \theta_0$
$H_0 : \theta \geq \theta_0$	$H_1 : \theta < \theta_0$
$H_0 : \theta \leq \theta_0$	$H_1 : \theta > \theta_0$

Ejemplo 4.1 Autoridades de educación han indicado que la nota promedio en el examen de bachillerato en matemática podría ser menor a 60. Si esta afirmación puede sustentarse estadísticamente, se pondrá en marcha un plan de ajuste que permita aumentar dicho valor medio. Plantee las hipótesis del problema en términos matemáticos así como del contexto del problema.

Solución

Dado que la afirmación de que la nota promedio en el examen de bachillerato en matemática podría ser menor a 60, esta se considera la hipótesis alternativa, de esta forma, las hipótesis del problema se plantean como sigue:

$H_0 : \mu \geq 60$: La nota promedio real en el examen de bachillerato en matemática es al menos 60.

$H_1 : \mu < 60$: La nota promedio real en el examen de bachillerato en matemática es significativamente menor a 60.

Definición 4.1 Región de rechazo

Dado un contraste de hipótesis, la región de rechazo es aquella que establece el conjunto de valores para los cuales la hipótesis nula debe ser rechazada.

Definición 4.2 Valores críticos

Dado un contraste de hipótesis, los valores críticos son aquellos que delimitan la región que se establece para considerar el rechazo de la hipótesis nula.

Definición 4.3 Estadístico de prueba

Dado un contraste de hipótesis, el estadístico de prueba es un valor sobre el cual, la decisión de rechazar o no la hipótesis nula se fundamenta. El estadístico de prueba se obtiene a partir de los datos de la muestra.

4.1.1. Errores en el proceso de contraste

Cuando se realiza un procedimiento de prueba de hipótesis puede ocurrir que la hipótesis nula sea rechazada cuando en realidad es verdadera, o bien, se acepte cuando debió rechazarse. Ambas situaciones indican la presencia de un error. Estos errores en el contraste de hipótesis se denominan error tipo I y error tipo II.

Definición 4.4 Error tipo I

Dado un contraste de hipótesis, se comete error tipo I cuando se rechaza la hipótesis nula dado que ella es verdadera. La probabilidad de cometer el error tipo I se llama nivel de significancia de la prueba y se denota con la letra α, donde:

$$\alpha = \mathbb{P}(\text{rechazar } H_0 | H_0 \text{ es verdadera})$$

Ejemplo 4.2 Considere la hipótesis nula de que la nota promedio del examen de bachillerato en matemática es igual a 60 y asuma que la nota sigue una distribución normal con $\sigma = 6$. La hipótesis alternativa afirma que la nota del examen es significativamente diferente a 60.

a) Determine el error tipo I para una muestra de $n = 9$, si la hipótesis nula es rechazada cuando la nota promedio es menor a 56 o mayor a 64.

b) Encuentre la probabilidad de cometer el error tipo I para una muestra de $n = 36$, si la hipótesis nula es rechazada cuando la nota promedio es menor a 58 o mayor a 62.

Solución

Parte a)

$$\begin{aligned} \alpha &= \mathbb{P}(\text{rechazar } H_0 | H_0 \text{ es verdadera}) \\ &= \mathbb{P}(\bar{x} < 56 \ \vee \ \bar{x} > 64 \mid \mu_0 = 60) \end{aligned}$$

```
> p1 <- pnorm(q = 56, mean = 60, sd = 6/sqrt(9))
> p2 <- 1 - pnorm(q = 64, mean = 60, sd = 6/sqrt(9))
> alpha = p1 + p2; alpha

[1] 0.04550026
```

Definición 4.5 Error tipo II

Dado un contraste de hipótesis, se comete error tipo II cuando no se rechaza la hipótesis nula dado que ella es falsa. La probabilidad de cometer el error tipo II se denota con la letra β, donde:

$$\beta = \mathbb{P}\ (\text{no rechazar } H_0 | H_0 \text{ es falsa})$$

Ejemplo 4.3 Considere los datos del Ejemplo 4.2 parte a) y determine el error tipo II cuando la verdadera nota promedio es 55.

Solución

De acuerdo con los datos del Ejemplo 4.2, $\beta = \mathbb{P}(56 < \bar{x} < 64 \mid \mu = 55)$. Es decir:

```
> p1 <- pnorm(q = 56, mean = 55, sd = 6/sqrt(9))
> p2 <- pnorm(q = 64, mean = 55, sd = 6/sqrt(9))
> beta1 <- p2 - p1; beta1

[1] 0.3085341
```

Ejemplo 4.4 Las personas asistentes a una oficina bancaria afirman que el tiempo promedio de atención al cliente es de al menos 25 minutos con desviación estándar de 2,8 minutos. Para analizar dicha afirmación se recolectó información de una muestra de 30 personas, la cual se presenta en la Tabla 4.2.

Tabla 4.2: Tiempo de atención al cliente

	1	2	3	4	5	6	7	8	9	10
1	21	21	25	23	20	25	22	21	25	22
2	22	22	21	24	20	24	24	20	24	21
3	23	24	22	24	21	23	21	19	23	20

a) Plantee las hipótesis del problema en términos estadísticos y del contexto.

b) Interprete en términos del problema los errores tipo I y tipo II.

c) ¿Cuál es la probabilidad de cometer el error tipo II, si la hipótesis nula se rechaza si el tiempo de atención promedio muestral es menor a 23 cuando el verdadero valor es de 22 minutos?

Solución

Parte a)

$H_0 : \mu \geq 25$: El tiempo promedio real de atención al cliente es de al menos 25 minutos.

$H_1 : \mu < 25$: El tiempo promedio real de atención al cliente es significativamente menor a 25 minutos.

Parte b)

Error tipo I: Decir que el tiempo promedio real de atención al cliente es significativamente menor a 25 minutos cuando en realidad no lo es.

Error tipo II: Decir que el tiempo promedio real de atención al cliente es de al menos 25 minutos cuando en realidad es significativamente menor.

Parte b)

$$\begin{aligned}\beta &= \mathbb{P}(\text{error tipo II}) \\ &= \mathbb{P}(\text{no rechazar } H_0 | H_0 \text{ es falsa}) \\ &= \mathbb{P}(\bar{x} \geq 23 \mid \mu = 22)\end{aligned}$$

```
> tiempo <- c(21,22,23,21,22,24,25,21,22,23,24,24,20,20,21,
+             25,24,23,22,24,21,21,20,19,25,24,23,22,21,20)
> mu <- 22; n <- length(tiempo); sigma <- 2.8
> beta <- 1 - pnorm(23, mean = mu, sd = sigma/sqrt(n))
> beta

[1] 0.02522363
```

4.1.2. Potencia de Prueba

Definición 4.6 Se llama potencia de prueba de un contraste de hipótesis a la probabilidad de rechazar la hipótesis nula cuando en verdad es falsa.

Es decir, dada una hipótesis alternativa compuesta por $H_1 : \theta \in \Theta_1$, la función potencia de prueba, denotada por $Potencia(\theta)$, está dada por:

$$Potencia(\theta) = \mathbb{P}(\text{rechazar } H_0 | H_0 \text{ es falsa}) = 1 - \beta(\theta)$$

Donde $\beta(\theta)$ es la probabilidad de cometer error tipo II, dado θ.

En otras palabras, la potencia de prueba se puede interpretar como la probabilidad de que el contraste detecte una diferencia, que en realidad existe. Note que $Potencia(\theta)$ es una función del parámetro θ, el cual tiene un valor en el subespacio paramétrico Θ_1 de la hipótesis alternativa ($\theta \in \Theta_1$). Cada hipótesis alternativa simple debería tener una potencia para el valor de θ.

Ejemplo 4.5 Considere los datos del Ejemplo 4.4 y determine la potencia de la prueba.

```
> potencia <- 1 - beta; potencia

[1] 0.9747764
```

Cuando la hipótesis nula es simple, $\theta = \theta_0$, la potencia de prueba en θ_0 es equivalente al nivel de significancia, es decir, Potencia$(\theta_0) = \alpha$.

Ejemplo 4.6 Considere los datos del Ejemplo 4.2 parte a) y verifique que la potencia de la prueba es igual al nivel de significancia cuando la verdadera nota promedio es 60.

Solución

```
> significancia <- alpha
> p1 <- pnorm(q = 64, mean = 60, sd = 6/sqrt(9))
> p2 <- pnorm(q = 56, mean = 60, sd = 6/sqrt(9))
> beta1 <- p1 - p2
> potencia1 <- 1 - beta1; potencia1

[1] 0.04550026
```

4.1.3. Valor p

Muchas personas se oponen a establecer *a priori* el nivel de significancia cuando prueban una hipótesis. En su lugar, prefieren tomar sus decisiones al rechazar o no rechazar la hipótesis nula, basados en el *valor p*.

Definición 4.7 Valor p

El *valor p* (*p value*) puede considerarse como la probabilidad, bajo un modelo estadístico especificado, de que un estadístico que resume alguna característica de los datos (por ejemplo, la diferencia de las medias al comparar dos grupos) sea igual o más extrema que su valor observado en la muestra (Wasserstein y Lazar, 2016).

Si se consideran valores de un estadístico de prueba s observados a partir de una muestra aleatoria, el cálculo del respectivo *valor p* para el contraste de hipótesis, se resume en la Tabla 4.3.

Tabla 4.3: Cálculo del *valor p* para distribuciones continuas

Hipótesis Alternativa	Valor p
$H_1 : \theta \neq \theta_0$	$2\mathbb{P}(S \geq \lvert s_{obs}\rvert \mid H_0)$
$H_1 : \theta > \theta_0$	$\mathbb{P}(S \geq \lvert s_{obs}\rvert \mid H_0)$
$H_1 : \theta < \theta_0$	$\mathbb{P}(S \geq \lvert s_{obs}\rvert \mid H_0)$

Es importante notar que el *valor p* no se ajusta *a priori*, pero si es determinado después de que la muestra ha sido tomada. Un valor bajo del *valor p* indica que las diferencias observadas tan

grandes o más grandes que las encontradas en la muestra son raras, por lo que no ocurren con mucha probabilidad. Un *valor p*, pequeño, permite apoyar H_1, pero se debe tener presente el nivel de significancia α, pues H_0 se rechaza si el *valor* $p < \alpha$.

4.2. Contraste de hipótesis para la media poblacional de una distribución normal

Según se indica en la Tabla 4.1, una hipótesis relacionada con la media poblacional μ puede plantearse de tres formas distintas, considerando a μ_0 para denotar el valor nulo de la media se tiene que:

Tabla 4.4: Contraste de hipótesis sobre la media de una población

H. Nula	H. Alternativa	Tipo de prueba
$H_0 : \mu = \mu_0$	$H_1 : \mu \neq \mu_0$	De dos colas
$H_0 : \mu \leq \mu_0$	$H_1 : \mu > \mu_0$	De cola derecha
$H_0 : \mu \geq \mu_0$	$H_1 : \mu < \mu_0$	De cola izquierda

4.2.1. Contraste de hipótesis para la media poblacional de una distribución normal con variancia poblacional conocida

Si una variable aleatoria X se distribuye normalmente con media μ y desviación estándar σ conocida, entonces para una muestra de tamaño n la región de rechazo para cada contraste de hipótesis para la media poblacional μ señalada en la Tabla 4.4 se establece de la siguiente manera:

Tabla 4.5: Región de rechazo para el contraste de hipótesis sobre la media de una población con σ conocida

H. Nula	H. Alternativa	Región de rechazo
$H_0 : \mu = \mu_0$	$H_1 : \mu \neq \mu_0$	$\bar{x} < \mu_0 - z_{\frac{\alpha}{2}} \cdot \dfrac{\sigma}{\sqrt{n}}$ ó $\bar{x} > \mu_0 + z_{\frac{\alpha}{2}} \cdot \dfrac{\sigma}{\sqrt{n}}$
$H_0 : \mu \leq \mu_0$	$H_1 : \mu > \mu_0$	$\bar{x} > \mu_0 + z_{\alpha} \cdot \dfrac{\sigma}{\sqrt{n}}$
$H_0 : \mu \geq \mu_0$	$H_1 : \mu < \mu_0$	$\bar{x} < \mu_0 - z_{\alpha} \cdot \dfrac{\sigma}{\sqrt{n}}$

Ejemplo 4.7 Considere la información del Ejemplo 4.2 que plantea la hipótesis nula de que la nota promedio del examen de bachillerato en matemática es igual a 60 y asuma que la nota sigue una distribución normal con $\sigma = 6$. Plantee la zona de rechazo para el contraste considerando una muestra de tamaño 49 y un nivel de significancia de 5% y represéntela gráficamente.

Solución

En este caso la hipótesis alternativa se plantea por: $H_1 : \mu \neq 60$. A su vez, la hipótesis nula se rechaza si $\bar{x} < \mu_0 - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ ó $\bar{x} > \mu_0 + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$.

```
> alpha <- 0.05; alpham <- alpha/2; mu <- 60; sigma <- 6; n <- 49
> zalpham <- qnorm(1 - alpham)
> lirechazo <- mu - zalpham*sigma/sqrt(n);
> lsrechazo <- mu + zalpham*sigma/sqrt(n);
> c(lirechazo, lsrechazo)

[1] 58.32003 61.67997
```

La hipótesis nula se rechaza si $\bar{x} < 58{,}32$ ó $\bar{x} > 61{,}68$.

Gráficamente, la región anterior se representa de la siguiente manera:

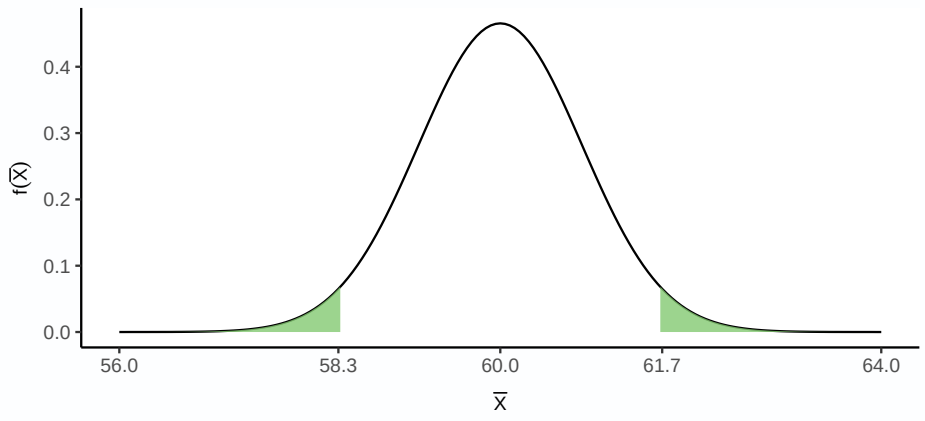


Figura 4.1: Región de rechazo para $H_1 : \mu \neq 60$, $\sigma = 6$ y $n = 49$

Fuente: Elaboración propia

Por otro lado,

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

es una variable aleatoria tal que $Z \sim N(0,\ 1)$. El estadístico de prueba para llevar a cabo el contraste de hipótesis para la media de una población que se distribuye normalmente con variancia poblacional conocida se denota por z_c y se define por

$$z_c = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

De esta manera, la región de rechazo para la hipótesis nula según lo planteado en la Tabla 4.4 queda definida de la siguiente manera:

Tabla 4.6: Región de rechazo para el contraste de hipótesis sobre la media de una población con σ conocida

H. Nula	**H. Alternativa**	**Región de rechazo**
$H_0 : \mu = \mu_0$	$H_1 : \mu \neq \mu_0$	$z_c < -z_{\frac{\alpha}{2}}$ ó $z_c > z_{\frac{\alpha}{2}}$
$H_0 : \mu \leq \mu_0$	$H_1 : \mu > \mu_0$	$z_c > z_\alpha$
$H_0 : \mu \geq \mu_0$	$H_1 : \mu < \mu_0$	$z_c < -z_\alpha$

Ejemplo 4.8 Considere los datos del Ejemplo 4.4.

a) Platee las hipótesis del problema en términos estadísticos y en términos del problema.

b) Contraste la hipótesis que el tiempo promedio de atención al cliente es de al menos 25 minutos considerando el estadístico de prueba.

c) Contraste la hipótesis que el tiempo promedio de atención al cliente es de al menos 25 minutos haciendo uso del valor p.

Solución

a) Las hipótesis del problema son:

$H_0 : \mu \geq 25$: El tiempo promedio real de atención al cliente es de al menos 25 minutos.

$H_1 : \mu < 25$: El tiempo promedio real de atención al cliente es significativamente menor a 25 minutos.

b) Considerando el estadístico de prueba, la hipótesis nula se rechaza si $z < -z_\alpha$.

```
> tiempo <- c(21,22,23,21,22,24,25,21,22,23,24,24,20,20,21,
+             25,24,23,22,24,21,21,20,19,25,24,23,22,21,20)
> alpha <- 0.05; sigma <- 2.8; n <- length(tiempo); mu <- 25
> xbarra <- mean(tiempo)
> z <- (xbarra - mu)/(sigma/sqrt(n))
> zalpha <- qnorm(1 - alpha)
> c(l1 <- -zalpha, z)

[1] -1.644854 -5.412021
```

Gráficamente,

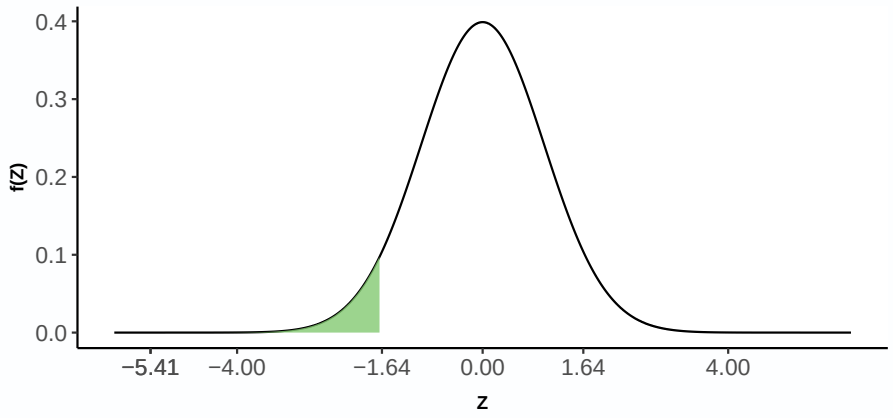


Figura 4.2: Región de rechazo utilizando z para $H_1 : \mu < 25$ y $n = 30$

Fuente: Elaboración propia

Como $-5{,}41 < -1{,}64$, es decir, $-5{,}41$ está en zona de rechazo. Existen evidencias, al nivel del 5%, de que el tiempo promedio real de atención al cliente es significativamente menor a 25 minutos.

Cuando se cuenta con los datos de la muestra, este tipo de contraste de hipótesis puede realizarse en R con ayuda de la función `ZTest()` del paquete `DescTools`

```
> library(DescTools)
> alpha <- 0.05; sigma <- 2.8; n <- 30; mu <- 25
> zalpha <- qnorm(1 - alpha)
> z <- ZTest(tiempo, mu = mu, alternative = "less", sd_pop = sigma,
+       conf.level = 1 - alpha)$statistic
> c(zalpha, z)

                  z
 1.644854 -5.412021
```

c) Si se considera el valor p, el procedimiento sería de la siguiente forma:

```
> p <- 1 - pnorm(abs(z)); p

           z
3.115877e-08
```

Utilizando la función `ZTest()` el procedimiento sería el siguiente:

```
> p <- ZTest(tiempo, mu = mu, alternative = "less", sd_pop = sigma,
+       conf.level = 1 - alpha)$p.value; p

[1] 3.115877e-08
```

Como $0 < 0{,}05$, puede concluirse que existen evidencias, al nivel del 5%, de que el tiempo promedio real de atención al cliente es significativamente menor a 25 minutos.

Ejemplo 4.9 Considere los datos del Ejemplo 4.4 y determine la potencia de la prueba considerando que el verdadero valor promedio es de 23 minutos.

Solución

Considérese inicialmente que la región de rechazo para este contraste está dada para valores de $\bar{x}$ tales que $\bar{x} < \mu - z_\alpha \cdot \dfrac{\sigma}{\sqrt{n}}$.

```
> alpha <- 0.05; sigma <- 2.8; n <- 30; mu <- 25
> zalpha <- qnorm(1 - alpha)
> rechazo <- mu - zalpha*sigma/sqrt(n); rechazo

[1] 24.15914
```

La hipótesis nula se rechaza si $\bar{x} < 24{,}16$, por lo tanto, la potencia de prueba está dada por $\beta = \mathbb{P}(\bar{x} < 24{,}16 \mid \mu = 23)$.

```
> potencia <- pnorm(24.15914, mean = 23, sd = 2.8/sqrt(30)); potencia

[1] 0.9883187
```

Ejemplo 4.10 En una compañía de aviación indican que el peso promedio de un determinado artículo es de 60 kilogramos con desviación estándar de 4,8 kilogramos. Contraste la afirmación anterior considerando que el peso se distribuye de manera normal y una muestra de 30 artículos reveló la información de la Tabla 4.7. Utilice un nivel de significancia de 5%.

Tabla 4.7: Peso de un artículo determinado

	1	2	3	4	5	6	7	8	9	10
1	59	54	72	65	60	60	60	61	61	63
2	64	60	56	73	65	66	66	66	69	69
3	70	64	60	56	56	56	58	59	59	59

Solución

$H_0 : \mu = 60$: El peso promedio real del artículo determinado es igual a 60 kilogramos.

$H_1 : \mu \neq 60$: El peso promedio real del artículo determinado es significativamente diferente a 60 kilogramos.

Considerando el método del valor p, el contraste se resuelve de la siguiente manera:

```
> alpha <- 0.05; sigma <- 4.8; n <- 30; mu = 60
> peso <- c(59,64,70,54,60,64,72,56,60,65,73,56,60,65,56,
+           60,66,56,60,66,58,61,66,59,61,69,59,63,69,59)
> p <- ZTest(peso, mu = mu, alternative = "two.sided", sd_pop = sigma,
+        conf.level = 1 - alpha)$p.value; p

[1] 0.01205962
```

Como 0,0121 < 0,05, existen evidencias, al nivel del 5%, de que el peso del artículo determinado es significativamente diferente a 60 kilogramos.

Considerando que la región de rechazo está definida por $\bar{x} < \mu_0 - z_{\frac{\alpha}{2}} \cdot \dfrac{\sigma}{\sqrt{n}}$ o $\bar{x} > \mu_0 + z_{\frac{\alpha}{2}} \cdot \dfrac{\sigma}{\sqrt{n}}$, el contraste puede resolverse de la siguiente manera:

```
> alpham <- alpha/2; xbarra <- mean(peso)
> zalpham <- qnorm(1 - alpham)
> lirechazo <- mu - zalpham*sigma/sqrt(n)
> lsrechazo <- mu + zalpham*sigma/sqrt(n)
> c(lirechazo, lsrechazo, xbarra)

[1] 58.28237 61.71763 62.20000
```

Gráficamente, la región de rechazo se muestra en la Figura 4.3

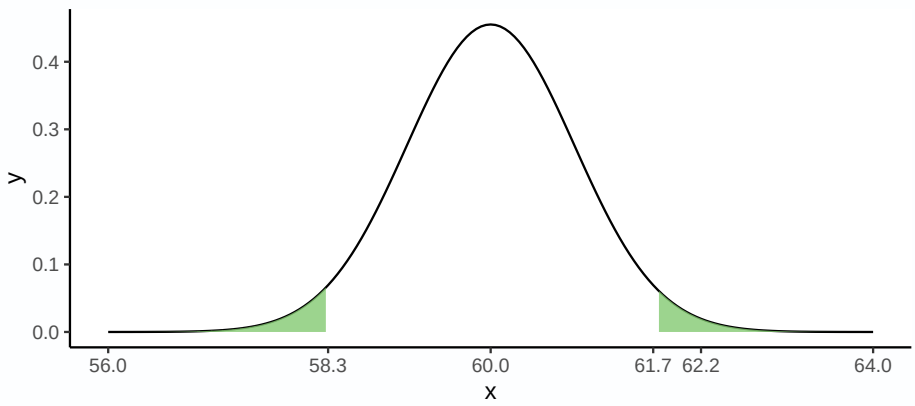


Figura 4.3: Región de rechazo para $H_1 \neq 60$, $\sigma = 4,8$ y $n = 30$

Fuente: Elaboración propia

Como 62,2 > 61,72, es decir, 62,2 está en zona de rechazo. Existen evidencias, al nivel del 5%, de que el peso promedio real de un determinado artículo es significativamente diferente a 60 kilogramos.

O bien, la solución podría plantearse considerando el estadístico de prueba, sabiendo que la hipótesis nula se rechaza si $z < -z_{\frac{\alpha}{2}}$ o $z > z_{\frac{\alpha}{2}}$. El estadístico de prueba está dado por:

```
> alpha <- 0.05; alpham <- alpha/2; sigma <- 4.8; n <- 30; mu = 60
> xbarra <- mean(peso)
> z <- (xbarra - mu)/(sigma/sqrt(n))
> zalpham <- qnorm(1 - alpham)
> c(l1 <- -zalpham, l2 <- zalpham, z)

[1] -1.959964  1.959964  2.510395
```

Como 2,51 > 1,96, es decir, 2,51 está en zona de rechazo. Existen evidencias, al nivel del 5%, de que el peso promedio real de un determinado artículo es significativamente diferente a 60 kilogramos.

Ejemplo 4.11 En una empresa de alta tecnología se afirmó que el tiempo promedio que las personas colaboradoras demoraban en tomar el café de la tarde era a lo sumo 32,6 minutos con una desviación estándar poblacional de 6,1 minutos. Si una muestra aleatoria de 36 personas reveló la información presentada en la Tabla 4.8, puede decirse con un nivel de significancia del 2% que las personas colaboradoras de la empresa tardan más de 32,6 minutos en su café de la tarde.

Tabla 4.8: Tiempo de demora en tomar el café

	1	2	3	4	5	6	7	8	9
1	31	32	33	31	32	34	35	31	32
2	34	34	30	30	31	35	34	33	29
3	28	29	30	29	35	34	33	32	31
4	29	30	33	33	30	33	28	30	32

Solución

$H_0 : \mu \leq 32,6$: El tiempo promedio real que tardan las personas colaboradoras de la empresa en tomar el café de la tarde es a lo sumo 32,6 minutos.

$H_1 : \mu > 32,6$: El tiempo promedio real que tardan las personas colaboradoras de la empresa en tomar el café de la tarde es significativamente mayor a 32,6 minutos.

```
> tiempo <- c(31,32,33,31,32,34,35,31,32,33,34,34,30,30,31,
+             35,34,33,29,28,28,29,30,29,35,34,33,32,31,30,
+             29,30,33,33,30,32)
> alpha <- 0.02; sigma <- 6.1; n <- 36; mu = 32.6
> p <- ZTest(tiempo, mu = mu, alternative = "two.sided", sd_pop = sigma,
+        conf.level = 1 - alpha)$p.value; p

[1] 0.3586017
```

Como 0,3586 > 0,02, no existen evidencias, al nivel del 2%, para decir que el tiempo promedio real que tardan las personas colaboradoras de la empresa en tomar el café de la tarde es significativamente mayor a 32,6 minutos.

Considerando el estadístico de prueba, la hipótesis nula se rechaza si $z > z_\alpha$. El estadístico de prueba está dado por:

```
> z <- ZTest(tiempo, mu = mu, alternative = "two.sided", sd_pop = sigma,
+         conf.level = 1 - alpha)$statistic
> zalpha <- qnorm(1 - alpha)
> c(l2 <- zalpha, z)

                    z
 2.0537489 -0.9180328
```

Como $-0{,}92 < 2{,}05$, es decir, $-0{,}92$ está en zona de no rechazo. No existen evidencias, al nivel del 2%, para decir que el tiempo promedio real que tardan las personas colaboradoras de la empresa en tomar el café de la tarde es significativamente mayor a 32,6 minutos.

4.2.2. Contraste de hipótesis para la media poblacional de una distribución normal con variancia poblacional desconocida

Si una variable aleatoria X se distribuye normalmente con media μ y desviación estándar σ desconocida, entonces para una muestra de tamaño n la regón de rechazo para cada contraste de hipótesis para la media poblacional μ señalada en la Tabla 4.4 se establece de la siguiente manera:

Tabla 4.9: Región de rechazo para el contraste de hipótesis sobre la media de una población con σ desconocida

Nula	Alternativa	Región de rechazo
$H_0 : \mu = \mu_0$	$H_1 : \mu \neq \mu_0$	$\bar{x} < \mu_0 - t_{\frac{\alpha}{2},\, n-1} \cdot \dfrac{s}{\sqrt{n}}$ ó $\bar{x} > \mu_0 + t_{\frac{\alpha}{2},\, n-1} \cdot \dfrac{s}{\sqrt{n}}$
$H_0 : \mu \leq \mu_0$	$H_1 : \mu > \mu_0$	$\bar{x} > \mu_0 + t_{\alpha,\, n-1} \cdot \dfrac{s}{\sqrt{n}}$
$H_0 : \mu \geq \mu_0$	$H_1 : \mu < \mu_0$	$\bar{x} > \mu_0 - t_{\alpha,\, n-1} \cdot \dfrac{s}{\sqrt{n}}$

Por otro lado, el estadístico de prueba para llevar a cabo el contraste de hipótesis para la media de una población que se distribuye normalmente con variancia poblacional desconocida está dado por t, donde

$$T = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Considerando que $T \sim T_{n-1}$, la región de rechazo para la hipótesis nula de la Tabla 4.4 queda definida de la siguiente manera:

Tabla 4.10: Región de rechazo para el contraste de hipótesis sobre la media de una población con σ conocida

Nula	Alternativa	Región de rechazo
$H_0 : \mu = \mu_0$	$H_1 : \mu \neq \mu_0$	$t < -t_{\frac{\alpha}{2}, n-1}$ ó $t > t_{\frac{\alpha}{2}, n-1}$
$H_0 : \mu \leq \mu_0$	$H_1 : \mu > \mu_0$	$t > t_{\alpha, n-1}$
$H_0 : \mu \geq \mu_0$	$H_1 : \mu < \mu_0$	$t < -t_{\alpha, n-1}$

Ejemplo 4.12 Una muestra aleatoria de tamaño 25 es tomada de una población normal, tal que $X \sim N(\mu_X,\ \sigma_X)$, donde $\bar{x} = 4$ y $s = 2,89$.

a) Pruebe la hipótesis nula de que $H_0 : \mu = 2,5$ versus, $H_1 : \mu \neq 2,5$, con un nivel de significancia de 0,05.

b) Calcule la potencia de prueba para $\mu_1 = 4$, asumiendo que $\sigma = 2,5$.

Solución

a) Considerando el estadístico de prueba, la hipótesis nula se rechaza si $t < -t_{\frac{\alpha}{2},\ n-1}$ o $t > t_{\frac{\alpha}{2},\ n-1}$. El estadístico de prueba está dado por:

```
> alpha <- 0.05; alpham <- alpha /2; n <- 25; mu = 2.5; xbarra <- 4; s <- 2.89;
> t <- (xbarra - mu)/(s/sqrt(n))
> talpham <- qt(1 - alpham, n - 1)
> c(l1 <- -talpham, l2 <- talpham, t)

[1] -2.063899  2.063899  2.595156
```

El valor de t también puede obtenerse utilizando la función `TTestA()`.

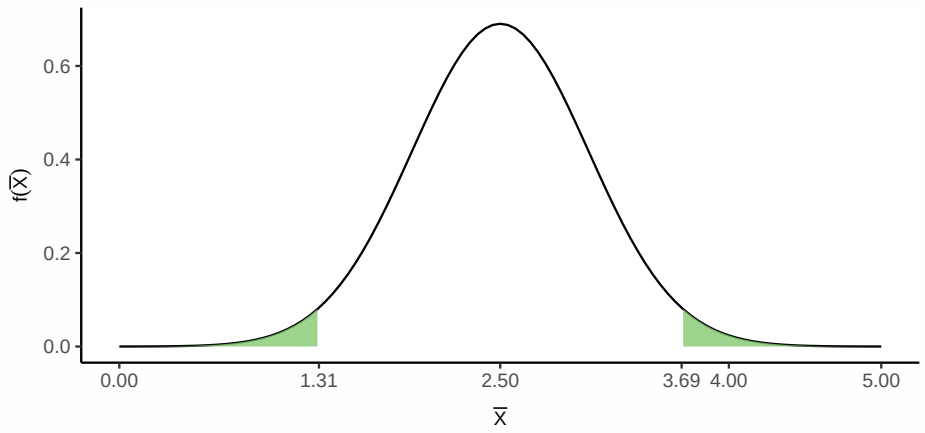
```
> TTestA(mx = 4, sx = 2.89, nx = 25, alternative = "two.sided",
+        mu = 2.5, var.equal = T)$statistic

       t
2.595156
```

Como 2,6 $\geq$ 2,06, es decir, 2,6 está en zona de rechazo. Por lo tanto, existen evidencias, al nivel del 5%, para decir que el promedio real es significativamente distinto de 2,5.

Por otro lado, si se considera la región de rechazo, la hipótesis nula se rechaza si $\bar{x} < \mu_0 - t_{\frac{\alpha}{2},n-1} \cdot \frac{s}{\sqrt{n}}$ o $\bar{x} > \mu_0 + t_{\frac{\alpha}{2},n-1} \cdot \frac{s}{\sqrt{n}}$.

```
> alpha <- 0.05; n <- 25; mu = 2.5; xbarra <- 4; s <- 2.89; alpham <- alpha/2
> lirechazo <- mu - talpham*s/sqrt(n)
> lsrechazo <- mu + talpham*s/sqrt(n)
> c(lirechazo, lsrechazo, xbarra)

[1] 1.307067 3.692933 4.000000
```

Gráficamente,



Se tiene que $4 > 3{,}69$, es decir, 4 está en zona de rechazo. Por lo tanto, existen evidencias, al nivel del 5%, para decir que el promedio real es significativamente distinto de 2,5.

b) El valor de la potencia de prueba se tiene a continuación:

```
> sigma <- 2.5
> lirechazo <- mu - talpham*s/sqrt(n)
> lsrechazo <- mu + talpham*s/sqrt(n)
> p1 <- pnorm(lirechazo, mean = 4, sd = sigma/sqrt(n))
> p2 <- 1 - pnorm(lsrechazo, mean = 4, sd = sigma/sqrt(n))
> potencia <- p1 + p2; potencia

[1] 0.7304364
```

Si se consideran valores del estadístico t observado a partir de los datos de una muestra aleatoria, el cálculo del respectivo valor p para el contraste de hipótesis, se resume en la Tabla 4.11.

Tabla 4.11: Cálculo del valor p para el contraste de hipótesis para la media de una población con σ desconocida

H. Nula	H. Alternativa	Valor p
$H_0 : \mu = \mu_0$	$H_1 : \mu \neq \mu_0$	$2 \cdot \mathbb{P}(T \geq \lvert t_{obs} \rvert)$
$H_0 : \mu \leq \mu_0$	$H_1 : \mu > \mu_0$	$\mathbb{P}(T \geq \lvert t_{obs} \rvert)$
$H_0 : \mu \geq \mu_0$	$H_1 : \mu < \mu_0$	$\mathbb{P}(T \geq \lvert t_{obs} \rvert)$

Ejemplo 4.13 Considere los datos del Ejemplo 4.12 y contraste la hipótesis nula de que $H_0 : \mu = 2,5$ versus, $H_1 : \mu \neq 2,5$, con un nivel de significancia de 0,05 haciendo uso del valor p.

Solución

```
> alpha <- 0.05; alpham <- alpha /2; n <- 25; mu = 2.5; xbarra <- 4; s <- 2.89
> t <- (xbarra - mu)/(s/sqrt(n))
> p <- 2*(1 - pt(abs(t), n - 1)); p

[1] 0.01587728
```

El valor p anterior también puede obtenerse utilizando la función `TTestA()`.

```
> library(DescTools)
> TTestA(mx = 4, sx = 2.89, nx = 25, alternative = "two.sided",
+        mu = 2.5, var.equal = T)$p.value

[1] 0.01587728
```

Como 0,0159 < 0,05 se rechaza H_0. Por lo tanto, existen evidencias, al nivel del 5%, para decir que el promedio real es significativamente distinto de 2,5.

Si se cuenta con los datos de la muestra, este tipo de contraste puede resolverse en R utilizando la función `t.test()`.

Ejemplo 4.14 Suponga que las notas del segundo examen parcial del Curso Probabilidad y Estadística realizado durante el II Ciclo, siguen una distribución normal. Para el grupo 04 las notas son 88, 33, 50, 48, 70, 75, 79, 60, 73, 74, 43, 58, 90 y 28.

a) Se desea probar que el rendimiento poblacional promedio de la prueba fue de 65, si se considera el supuesto de que las calificaciones se distribuyen normalmente y se utiliza un nivel de significancia del 4%, indique si los resultados muestrales contradicen la afirmación anterior.

b) Calcule e interprete la potencia de prueba si se considera que $\mu_1 = 80$, asumiendo que $\sigma = 20$.

Solución

a) En primer lugar se plantean las hipótesis del problema.

$H_0 : \mu = 65$: El rendimiento promedio real en la prueba del curso es de 65.

$H_1 : \mu \neq 65$: El rendimiento promedio real en la prueba del curso es significativamente distinto de 65.

En segundo lugar, se selecciona el método que se empleará en la solución del problema. Considerando el *valor p*, el contraste de hipótesis puede resolverse en R de la siguiente manera:

```
> rendimiento <- c(88, 33, 50, 48, 70, 75, 79, 60, 73, 74, 43, 58, 90, 28)
> alpha <- 0.04; n <- length(rendimiento); mu = 65
> p <- t.test(rendimiento, mu = mu, altenartive = "two.sided",
+         conf.level = 1 - alpha)$p.value; p

[1] 0.5845331
```

Como 0,5845 > 0,04, no existen evidencias, al nivel del 4%, para decir que el rendimiento promedio real en la prueba del curso es significativamente distinto de 65. Es decir, los datos muestrales no contradicen la afirmación de que $\mu = 65$.

Considerando el estadístico de prueba, la hipótesis nula se rechaza si $t < -t_{\frac{\alpha}{2},\, n-1}$ o $t > t_{\frac{\alpha}{2},\ n-1}$. El estadístico de prueba está dado por:

```
> talpham <- qt(1 - alpham, n - 1)
> t <- t.test(rendimiento, mu = mu, altenartive = "two.sided",
+         conf.level = 1 - alpha)$statistic
> c(l1 <- -talpham, l2 <- talpham, t)

                                t
-2.1603687  2.1603687 -0.5607058
```

Como $-2{,}16 \leq -0{,}56 \leq 2{,}16$, es decir, $-0{,}56$ está en zona de no rechazo. Por lo tanto, no existen evidencias, al nivel del 4%, para decir que el rendimiento promedio real en la prueba del curso es significativamente distinto de 65. Es decir, no se contradice la afirmación de que $\mu = 65$.

b) Considerando que la región de rechazo está definida por $\bar{x} < \mu_0 - t_{\frac{\alpha}{2},n-1} \cdot \frac{s}{\sqrt{n}}$ o $\bar{x} > \mu_0 - t_{\frac{\alpha}{2},n-1} \cdot \frac{s}{\sqrt{n}}$, entonces la potencia de la prueba (probabilidad de rechazar H_0 siendo falsa) viene dada por:

$$potencia = \mathbb{P}(\bar{x} < \mu_0 - t_{\frac{\alpha}{2},n-1} \cdot \frac{s}{\sqrt{n}} \mid \mu = 80) + \mathbb{P}(\bar{x} > \mu_0 - t_{\frac{\alpha}{2},n-1} \cdot \frac{s}{\sqrt{n}} \mid \mu = 80)$$

```
> alpha <- 0.04; alpham <- alpha/2; mu <- 65; sigma <- 20
> n <- 14; talpham <- qt(1 - alpham, n - 1)
> lirechazo <- mu - talpham*s/sqrt(n)
> lsrechazo <- mu + talpham*s/sqrt(n)
> p1 <- pnorm(lirechazo, mean = 80, sd = sigma/sqrt(n))
> p2 <- 1 - pnorm(lsrechazo, mean = 80, sd = sigma/sqrt(n))
> potencia <- p1 + p2; potencia

[1] 0.9942236
```

Ejemplo 4.15 Se realiza una prueba de control de calidad para determinar el peso neto de cierto tipo de confites, para ello se tomó una muestra de 22 confites, la cual dio como resultado una media de 76 g y una desviación estándar de 12,5 g. Estos resultados preocupan a la gerencia pues se supone que el peso neto de los confites es a lo sumo 70 g, si se utiliza un nivel de significancia de 0,025, el valor arrojado por la muestra es probatoria de que el peso neto promedio de los confites es mayor a 70 g. Suponga que el peso de los confites sigue una distribución aproximadamente normal.

Ejemplo 4.16 Un estudio realizado por el Ministerio de Salud en conjunto con el Ministerio de Educación Pública afirma que la niñez de primaria en el circuito 01 de la provincia de Heredia tienen un sobrepeso promedio aproximado de al menos 5 kilos, esta afirmación se contrapone a una muestra de 18 infantes de las escuelas del circuito que fueron pesados y en promedio su sobrepeso fue de 4,2 kilos con una desviación estándar de 1,35 kilos. Suponga que el peso de la niñez sigue una distribución aproximadamente normal.

a) Indique si hay alguna razón para dudar de la validez de que la niñez tiene en promedio un sobre peso de 5 kilos utilizando un nivel de significancia del 1%.

b) Indique si hay alguna razón para dudar de la validez de que la niñez tiene en promedio un sobre peso de 5 kilos utilizando un nivel de significancia del $2,5\%$.

4.3. Contraste de hipótesis para la proporción de éxitos en un experimento Binomial (aproximación Normal)

Si $X = \sum_{i=1}^{n} Y_i$, donde $Y_i \sim Bernoulli(P)$ entonces $X \sim Bin(n, P)$. El cálculo numérico necesario para el método exacto requiere el uso de una computadora sobre todo si la muestra es grande. Afortunadamente, es posible encontrar una aproximación a la distribución exacta cuando las muestras son grandes, gracias a las propiedades de los estimadores máximo verosímiles, pues

$$\hat{p} = \frac{X}{n} \sim N\left(P, \sqrt{\frac{P(1-P)}{n}}\right)$$

cuando $n \to \infty$. Cuando nP y $n(1-P)$ son ambos mayores a 5 (o 10 en algunos casos)

$$\hat{p} \sim N\left(P, \sqrt{\frac{P(1-P)}{n}}\right)$$

da una aproximación razonable a la distribución muestral de $\hat{p}$. Al estandarizar el estadístico de prueba, bajo el supuesto de que $H_0 : P = P_0$ es cierta, entonces:

$$Z = \frac{\hat{p} - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} \sim N(0,1)$$

Cuando $|\hat{p} - P_0| > \frac{1}{2n}$, se suele aplicar la corrección por continuidad.

Una hipótesis relacionada con la proporción poblacional P puede plantearse de tres formas distintas, considerando a P_0 para denotar el valor nulo de la proporción se tiene que:

Tabla 4.12: Contraste de hipótesis sobre la proporción de una población

H. Nula	**H. Alternativa**	**Tipo de prueba**
$H_0 : P = P_0$	$H_1 : P \neq P_0$	De dos colas
$H_0 : P \leq P_0$	$H_1 : P > P_0$	De cola derecha
$H_0 : P \geq P_0$	$H_1 : P < P_0$	De cola izquierda

Si una variable aleatoria X se distribuye binomialmente con parámetro P desconocido, entonces, para una muestra de tamaño n la región de rechazo para cada contraste de hipótesis para la proporción poblacional P se establece de la siguiente manera:

Tabla 4.13: Región de rechazo para el contraste de hipótesis sobre la proporción de una población

H. Nula	**H. Alternativa**	**Región de rechazo**
$H_0 : P = P_0$	$H_1 : P \neq P_0$	$\hat{p} < P_0 - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{P_0(1-P_0)}{n}}$ ó $\hat{p} > P_0 + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{P_0(1-P_0)}{n}}$
$H_0 : P \leq P_0$	$H_1 : P > P_0$	$\hat{p} > P_0 + z_{\alpha} \cdot \sqrt{\frac{P_0(1-P_0)}{n}}$
$H_0 : P \geq P_0$	$H_1 : P < P_0$	$\hat{p} < P_0 - z_{\alpha} \cdot \sqrt{\frac{P_0(1-P_0)}{n}}$

Ejemplo 4.17 Un reciente estudio sobre empleo señala que alrededor del 20% de todas las personas graduadas en carreras de ciencias sociales, encuentran trabajo en su especialidad en menos de dos años de haberse egresado de bachillerato. Una muestra aleatoria de 500 profesionales encontró que 90 de ellos encontraron trabajo en su campo de especialización en menos de dos años, a partir de su graduación de bachillerato. Con un nivel de significancia del 4%, investigue si se puede decir que la afirmación del estudio es correcta.

Solución

Las hipótesis del problema se plantean de la siguiente manera:

$H_0 : P = 0,20$: La proporción real de personas graduadas en carreras de ciencias sociales que encuentran trabajo en su especialidad en menos de dos años de haberse egresado de bachillerato es de 20%.

$H_1 : P \neq 0,20$: La proporción real de personas graduadas en carreras de ciencias sociales que encuentran trabajo en su especialidad en menos de dos años de haberse egresado de bachillerato es significativamente distinta de 20%.

La hipótesis nula se rechaza si

$$\hat{p} < P_0 - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{P_0(1-P_0)}{n}} \text{ ó } \hat{p} > P_0 + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{P_0(1-P_0)}{n}}$$

```
> alpha <- 0.04; alpham <- alpha/2; p0 <- 0.20; n <- 500; x <- 90
> zalpham <- qnorm(1 - alpham); pest <- x / n
> lirechazo <- p0 - zalpham*sqrt(0.20*0.8/n);
> lsrechazo <- p0 + zalpham*sqrt(0.20*0.8/n);
> c(lirechazo, lsrechazo, pest )

[1] 0.1632614 0.2367386 0.1800000
```

La región de rechazo se muestra gráficamente como sigue:

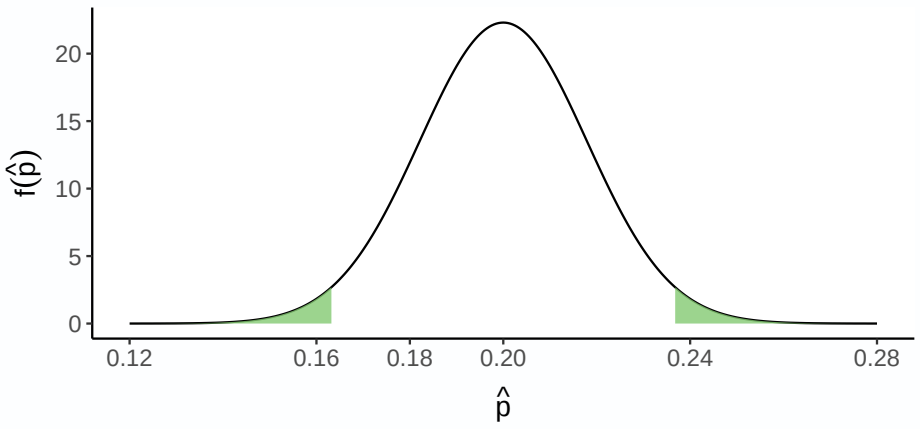


Figura 4.4: Región de rechazo para $H_1 : P \neq 0,20$ y $n = 500$

Fuente: Elaboración propia

Como 0,1633 $<$ 0,18 $<$ 0,2367, la hipótesis nula no se rechaza. Es decir, no existen evidencias, al nivel del 4%, para decir que la proporción real de personas graduadas en carreras de ciencias sociales que encuentran trabajo en su especialidad en menos de dos años de haberse egresado de bachillerato es significativamente distinta de 20%.

Por otro lado, el estadístico de prueba para llevar a cabo el contraste de hipótesis para la proporción de una población que se distribuye binomialmente está dado por z, tal que

$$Z = \frac{\hat{p} - P_0}{\sqrt{\frac{P_0(1 - P_0)}{n}}}$$

Considerando que $Z \sim N(0,\ 1)$, la región de rechazo para la hipótesis nula de la Tabla 4.12 queda definida de la siguiente manera:

Tabla 4.14: Región de rechazo para el contraste de hipótesis sobre la proporción de una población

H. Nula	H. Alternativa	Región de rechazo
$H_0 : P = P_0$	$H_1 : P \neq P_0$	$z < -z_{\frac{\alpha}{2}}$ ó $z > z_{\frac{\alpha}{2}}$
$H_0 : P \leq P_0$	$H_1 : P > P_0$	$z > z_\alpha$
$H_0 : P \geq P_0$	$H_1 : P < P_0$	$z < -z_\alpha$

Ejemplo 4.18 Considere la información del Ejemplo 4.17 y realice el contraste de hipótesis utilizando el estadístico de prueba.

Solución

Las hipótesis del problema se plantean de la siguiente manera:

$H_0 : P = 0,20$: La proporción real de personas graduadas en carreras de ciencias sociales que encuentran trabajo en su especialidad en menos de dos años de haberse egresado de bachillerato es de 20%.

$H_1 : P \neq 0,20$: La proporción real de personas graduadas en carreras de ciencias sociales que encuentran trabajo en su especialidad en menos de dos años de haberse egresado de bachillerato es significativamente distinta de 20%.

Considerando el estadístico de prueba, la hipótesis nula se rechaza si

$$z < -z_{\frac{\alpha}{2}} \text{ ó } z > z_{\frac{\alpha}{2}}$$

```
> alpha <- 0.04; alpham <- alpha/2; p0 <- 0.20; n <- 500; x <- 90
> zalpham <- qnorm(1 - alpham); pest <- x / n
> z <- (pest - p0) / sqrt(p0*(1 - p0)/n)
> c(-zalpham, zalpham, z)

[1] -2.053749  2.053749 -1.118034
```

Gráficamente,

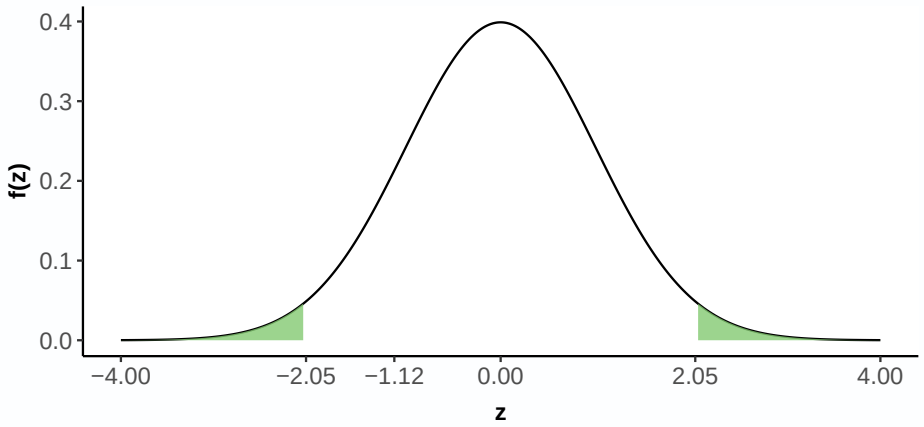


Figura 4.5: Región de rechazo utilizando z para $H_1 : P \neq 0,20$ y $n = 500$

Fuente: Elaboración propia

Como puede observarse en la Figura 4.5, $-2{,}05 < -1{,}12 < 2{,}05$, es decir, $-1{,}12$ está en zona de no rechazo. Por lo tanto, no existen evidencias, al nivel del 4%, para decir que la proporción real de personas graduadas en carreras de ciencias sociales que encuentran trabajo en su especialidad en menos de dos años de haberse egresado de bachillerato es significativamente distinta de 20%.

Si se consideran valores del estadístico z observado a partir de los datos de una muestra aleatoria, el cálculo del respectivo valor p para el contraste de hipótesis se resume en la Tabla 4.15.

Tabla 4.15: Cálculo del valor p para el contraste de hipótesis sobre la proporción de una población

H. Nula	**H. Alternativa**	**Valor p**
$H_0 : P = P_0$	$H_1 : P \neq P_0$	$2 \cdot \mathbb{P}(Z \geq \lvert z_{obs} \rvert)$
$H_0 : P \leq P_0$	$H_1 : P > P_0$	$\mathbb{P}(Z \geq \lvert z_{obs} \rvert)$
$H_0 : P \geq P_0$	$H_1 : P < P_0$	$\mathbb{P}(Z \geq \lvert z_{obs} \rvert)$

Ejemplo 4.19 Considere la información del Ejemplo 4.17 y realice el contraste de hipótesis utilizando el valor p.

Solución

Considerando el valor p, la hipótesis nula se rechaza si $p < \alpha$, donde $p = 2 \cdot \mathbb{P}(Z \geq \lvert z_{obs} \rvert)$.

```
> alpha <- 0.04; alpham <- alpha/2; p0 <- 0.20; n <- 500; x <- 90
> zalpham <- qnorm(1 - alpham); pest <- x / n
> z <- (pest - p0) / sqrt(p0*(1 - p0)/n); z

[1] -1.118034

> p <- 2*(1 - pnorm(abs(z))); p

[1] 0.2635525
```

Como 0,2636 > 0,04, se tiene que la hipótesis nula no se rechaza.

4.4. Contraste Chi-cuadrado

Recordando que si X es el número de éxitos en una muestra grande de tamaño n, la distribución de muestreo de la variable

$$Z = \frac{x - nP_0}{\sqrt{nP_0(1-P_0)}} = \frac{\hat{p} - P_0}{\sqrt{\dfrac{P_0(1-P_0)}{n}}}$$

es aproximadamente normal estándar. Sin embargo, el cuadrado de una variable aleatoria normal estándar es una variable aleatoria con distribución Chi-cuadrado (ji-cuadrado) con un grado de libertad. Así la expresión

$$\chi^2 = \frac{(x - nP_0)^2}{nP_0(1-P_0)}$$

es un valor de una variable aleatoria que tiene aproximadamente la distribución Chi-cuadrado con 1 grado de libertad. De esta manera, la hipótesis nula se rechaza si:

$$\chi^2 > \chi^2_{\alpha,\,1}$$

donde, $\chi^2_{\alpha,\,1}$ representa el $100(1-\alpha)$ percentil de la distribución Chi-cuadrado con un grado de libertad. Sin embargo, cuando $|\hat{p} - P_0| > \dfrac{1}{2n}$ puede aplicarse la corrección de Yates o corrección por continuidad. En este caso, el estadístico de prueba se expresa por:

$$\chi^2 = \frac{\left(|\hat{p} - P_0| - \frac{1}{2n}\right)^2}{\frac{P_0(1-P_0)}{n}}$$

Ejemplo 4.20 Considere la información del Ejemplo 4.17 y realice el contraste de hipótesis utilizando el estadístico chi-cuadrado.

```
> alpha <- 0.04; p0 <- 0.20; n <- 500; x <- 90
> chialpha <- qchisq(1 - alpha, 1); pest <- x / n
> chi2 <- (abs(pest - p0) - 1/(2*n))^2 / (p0*(1 - p0)/n)
> c(chialpha, chi2)

[1] 4.217885 1.128125
```

Se puede observar que 1,13 < 4,22, es decir, la hipótesis nula no se rechaza.

El contraste de hipótesis para la proporción poblacional utilizando el estadístico chi-cuadrado puede realizarse en R con ayuda de la función `prop.test()`. Es importante mencionar que la función `prop.test()` considera una corrección de Yates (corrección por continuidad) al obtener el estadístico chi-cuadrado, por lo que dicho valor no corresponde necesariamente al valor de z^2.

Ejemplo 4.21 Considere la información del Ejemplo 4.17 y realice el contraste de hipótesis utilizando la función `prop.test()`.

Solución

```
> chi2 <- prop.test(x = 90, n = 500, conf.level = 0.96, p = 0.20)$stat
> chialpha <- qchisq(0.96, 1)
> c(chialpha, chi2)

          X-squared
 4.217885  1.128125
```

Si se consideran valores del estadístico χ^2 observado a partir de los datos de una muestra aleatoria, el cálculo del respectivo valor p para el contraste de hipótesis se resume en la Tabla 4.16.

Tabla 4.16: Cálculo del valor p para el contraste de hipótesis (estadístico χ^2) sobre la proporción de una población

H. Nula	**H. Alternativa**	**Valor** p
$H_0 : P = P_0$	$H_1 : P \neq P_0$	$\mathbb{P}(\chi^2 \geq \chi^2_{obs})$
$H_0 : P \leq P_0$	$H_1 : P > P_0$	$\mathbb{P}(\chi^2 \geq \chi^2_{obs})$
$H_0 : P \geq P_0$	$H_1 : P < P_0$	$\mathbb{P}(\chi^2 \geq \chi^2_{obs})$

Ejemplo 4.22 Considere la información del Ejemplo 4.17 y realice el contraste de hipótesis utilizando el valor p.

Solución

Considerando el valor p, el contraste se realiza de la siguiente manera:

```
> valorp <- 1 - pchisq(chi2, 1); valorp

X-squared
0.2881756

> # o bien,
> prop.test(x = 90, n = 500, conf.level = 0.96,
+           alternative = "two.sided", p = 0.20)$p.value

[1] 0.2881756
```

Se puede observar que 0,2882 > 0,04, es decir, la hipótesis nula no se rechaza.

Ejemplo 4.23 Muchas personas están recurriendo a productos genéricos como una alternativa para reducir los costos de los medicamentos. Un artículo publicado en una revista particular da el resultado de un estudio en el que participaron 102 personas, de las cuales solo 47 conocían el nombre genérico de un determinado medicamento. Tomando un nivel de significancia del 1%, proporciona esto una fuerte evidencia para concluir que menos de la mitad de las personas participantes conocen el nombre genérico de la metadina.

Ejemplo 4.24 En una fábrica de suministros de computadora se está en el proceso de decidir si produce o no una versión de teclado inalámbrico. El departamento de investigación de mercados de la compañía utilizó un sondeo telefónico a 2500 casas y encontró que en 140 de ellas comprarían el nuevo teclado. Un estudio más extenso hecho un año antes mostró que el $4,5\%$ de los usuarios compraría ese tipo de teclado. Con un nivel de significancia del 2%, ¿debe la compañía concluir que hay un incremento en el interés, por parte de los usuarios de adquirir el teclado inalámbrico?

4.5. Determinación del tamaño de muestra para contraste de hipótesis

Cuando se realiza una prueba de hipótesis con un nivel de significancia α (error tipo I) y además se desea reducir el β (error tipo II) a un valor determinado, la única forma de reducir el β sin alterar el α es aumentando el tamaño de la muestra. Para determinar el tamaño de muestra apropiado puede utilizarse las siguientes fórmulas.

4.5.1. Determinación del tamaño de muestra para el contraste de la media con β y α fijos

Para estimar el tamaño mínimo de muestra necesario para llevar a cabo una prueba acerca de la media poblacional μ_0, con un nivel de significancia α, una probabilidad de cometer error tipo II β y un valor alternativo μ_1 se utiliza la expresión:

$$n = \begin{cases} \left[\dfrac{\sigma_X(z_\alpha + z_\beta)}{\mu_0 - \mu_1}\right]^2 & \text{si la prueba es de una cola} \\ \\ \left[\dfrac{\sigma_X(z_{\frac{\alpha}{2}} + z_\beta)}{\mu_0 - \mu_1}\right]^2 & \text{si la prueba es de dos colas} \end{cases}$$

Ejemplo 4.25 Considere la afirmación de que la velocidad promedio en un punto específico de la autopista General Cañas es de 100 km/h. ¿Cuán grande debe ser la muestra aleatoria que se debe tomar para probar la afirmación de que la velocidad promedio en un punto específico de la autopista General Cañas es de 100 km/h contra la hipótesis de que la velocidad promedio es significativamente diferente a 100 km/h? Considere $\sigma = 16$ km/h, la probabilidad de cometer error tipo I de 0,01 y que la probabilidad de cometer un error tipo II debe ser 0,20 para un $\mu = 93$ km/h.

Solución

```
> alpha <- 0.05; beta <- 0.20; sigma <- 16; mu0 <- 100; mu1 <- 93
> alpham <- alpha / 2
> zalpham <- qnorm(1 - alpham); zbeta <- qnorm(1 - beta)
> n <- (sigma*(zalpham + zbeta)/(mu0 - mu1))^2; n

[1] 41.00639
```

De acuerdo con la información proporcionada, es necesario un tamaño de muestra aproximado de 42 vehículos.

Ejemplo 4.26 Considérese el Ejemplo 4.25. ¿Cuán grande debe ser la muestra aleatoria que se debe tomar para realizar el contraste anterior si la probabilidad de cometer un error tipo II debe ser 0,05?

Solución

```
> alpha <- 0.05; beta <- 0.05; sigma <- 16; mu0 <- 100; mu1 <- 93
> alpham <- alpha / 2
> zalpham <- qnorm(1 - alpham); zbeta <- qnorm(1 - beta)
> n <- (sigma*(zalpham + zbeta)/(mu0 - mu1))^2; n

[1] 67.89073
```

Se requiere un tamaño mínimo de muestra aproximado de 68 vehículos cuando la probabilidad de cometer error tipo II debe ser de 0,05.

Ejemplo 4.27 Considérese la necesidad de probar una hipótesis nula $\mu = 40$ contra la hipótesis alternativa de que $\mu < 40$ con base en una muestra aleatoria grande de una población con $\sigma = 4$. Si la probabilidad de un error tipo I es de 0,05 y la probabilidad de un error tipo II es de 0,12 para $\mu = 38$, determine el tamaño de muestra requerido.

4.5.2. Determinación del tamaño de muestra para la proporción con β y α fijos

Para estimar el tamaño mínimo de muestra necesario para llevar a cabo una prueba acerca de la proporción poblacional P_0, con un nivel de significancia α, una probabilidad de cometer error tipo II β y un valor alternativo P_1 se utiliza la expresión:

$$n = \begin{cases} \left[\dfrac{z_\alpha \sqrt{P_0(1-P_0)} + z_\beta \sqrt{P_1(1-P_1)}}{P_0 - P_1} \right]^2 & \text{si la prueba es de una cola} \\ \left[\dfrac{z_{\frac{\alpha}{2}} \sqrt{P_0(1-P_0)} + z_\beta \sqrt{P_1(1-P_1)}}{P_0 - P_1} \right]^2 & \text{si la prueba es de dos colas} \end{cases}$$

Capítulo 5

Contraste de hipótesis con base en dos muestras

5.1. Contraste de hipótesis para la diferencia de medias de dos poblaciones

Sean dos variables aleatorias X y Y, cuyas medias están dadas por μ_X y μ_Y, respectivamente, una hipótesis relacionada con la diferencia de las medias poblacionales $\mu_X - \mu_Y$ puede plantearse de tres formas distintas, considerando a $(\mu_X - \mu_Y)_0 = \delta_0$ para denotar el valor nulo de la diferencia, se tiene que:

Tabla 5.1: Contraste de hipótesis para la diferencia de dos medias poblacionales

H. Nula	H. Alternativa	Tipo de prueba
$H_0 : \mu_X - \mu_Y = \delta_0$	$H_1 : \mu_X - \mu_Y \neq \delta_0$	De dos colas
$H_0 : \mu_X - \mu_Y \leq \delta_0$	$H_1 : \mu_X - \mu_Y > \delta_0$	De cola derecha
$H_0 : \mu_X - \mu_Y \geq \delta_0$	$H_1 : \mu_X - \mu_Y < \delta_0$	De cola izquierda

5.1.1. Diferencia de medias de dos poblaciones independientes que se distribuyen normalmente y las variancias poblacionales son conocidas

Sean X y Y dos variables aleatorias que se distribuyen normalmente con medias μ_X y μ_Y y desviación estándar σ_X y σ_Y, respectivamente, entonces para muestras de tamaño n_X y n_Y, la región de rechazo para cada contraste de hipótesis para la diferencia de las medias poblacionales $\mu_X - \mu_Y$ señaladas en la Tabla 5.1 se establece de la siguiente manera:

Cuando se muestrean dos poblaciones que se distribuyen de manera normal con variancias conocidas, el contraste de hipótesis para la diferencia de las medias puede expresarse como:

$$H_0 : \mu_X - \mu_Y = \delta_0$$

Considerando el supuesto de que H_0 es cierta, la variable

$$Z = \frac{\bar{X} - \bar{Y} - \delta_0}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \sim N(0,1)$$

El estadístico considerado es:

$$Z = \frac{\bar{X} - \bar{Y} - \delta_0}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}}$$

Ejemplo 5.1 Una persona investigadora desea conocer si las personas que ingresan a la carrera A obtienen mejores puntajes en la prueba de aptitud académica (PAA), utilizada para ingresar a la UNA, que las personas que ingresan a la carrera B. Para ello, considera dos muestras aleatorias, la primera constituida por el puntaje de 64 de personas de la carrera A, que tradicionalmente sigue una distribución normal con $\sigma = 100$. La segunda muestra consiste en 144 puntajes de personas de la carrera B, los cuales siguen una distribución aproximadamente normal con $\sigma = 108$.

a) Realice el contraste de hipótesis para la diferencia poblacional de las notas entre ambas carreras, con un nivel de significancia del 10%, si se sabe que la diferencia muestral fue de 30 puntos en favor de las personas de la carrera A.

b) Encuentre la potencia de prueba para este contraste, si se desea verificar si existe diferencia entre los puntajes de ambas poblaciones, asumiendo que $H_1 : \mu_X - \mu_Y = 40$.

Solución a)

$H_0 : \mu_X \leq \mu_Y$: El puntaje promedio real en la PAA obtenido en la carrera A es menor o igual al que se obtiene en la carrera B.

$H_1 : \mu_X > \mu_Y$: El puntaje promedio real en la PAA es significativamente distinto en ambas carreras.

Además, $H_0 : \mu_X \leq \mu_Y$ es equivalente a $H_0 : \mu_X - \mu_Y \leq 0$

```
> alpha <- 0.10
> n1 <- 64; sigma1 <- 100
> n2 <- 144; sigma2 <- 108
> difmu <- 0
> difxbarra <- 30
> num <- (difxbarra) - (difmu)
> den <- sqrt(sigma1^2/n1 + sigma2^2/n2)
> z <- num/den
> zalpha <- qnorm(1 - alpha)
> c(l2 <- zalpha, z)

[1] 1.281552 1.947682
```

El estadístico 1,95 > 1,28, es decir, 1,95 está en zona de rechazo. Por lo tanto, existen evidencias, al nivel del 10%, para decir que el puntaje promedio real en la PAA es significativamente mayor en las personas de la carrera A.

Considerando el valor p se tiene:

```
> z

[1] 1.947682

> p <- 1 - pnorm(abs(z)); p

[1] 0.02572649
```

El 0,0257 < 0,1, es decir, se rechaza H_0. Por lo tanto, existen evidencias, al nivel del 10%, para decir que el puntaje promedio real en la PAA es significativamente mayor en las personas de la carrera A.

Ejemplo 5.2 El análisis de una muestra aleatoria formada por 30 estudiantes residentes en zona rural reveló un rendimiento promedio en un curso determinado de 79,8. Una segunda muestra aleatoria de 36 estudiantes residentes en zona urbana mostró un rendimiento promedio en el curso de 84,7. Suponga que las dos distribuciones del rendimiento son normales con $\sigma_{rural} = 8,0$ y $\sigma_{urbana} = 9,0$, respectivamente. Utilizando un nivel de significancia del 1%, puede decirse que el rendimiento en las dos poblaciones es diferente.

Solución

Las hipótesis del problema son las siguientes:

$H_0 : \mu_X = \mu_Y$: El rendimiento promedio real en el curso es el mismo según la zona de residencia.

$H_1 : \mu_X \neq \mu_Y$: El rendimiento promedio real en el curso es significativamente distinto según la zona de residencia.

Además, $H_0 : \mu_X = \mu_Y$ es equivalente a $H_0 : \mu_X - \mu_Y = 0$

```
> alpha <- 0.01; alpham <- alpha/2
> n1 <- 30; xbarra1 <- 79.8; sigma1 <- 8
> n2 <- 36; xbarra2 <- 84.7; sigma2 <- 9
> difmu <- 0
> num <- (xbarra1 - xbarra2) - (difmu)
> den <- sqrt(sigma1^2/n1 + sigma2^2/n2)
> z <- num/den
> zalpham <- qnorm(1 - alpham)
> c(l1 <- -zalpham, l2 <- zalpham, z)

[1] -2.575829  2.575829 -2.340420
```

El estadístico $-2{,}34$ está contenido en el intervalo $[-2{,}58,\ 2{,}58]$, es decir, $-2{,}34$ está en zona de no rechazo. Por lo tanto, no existen evidencias, al nivel del 1%, para decir que el rendimiento promedio real en el curso es significativamente distinto en ambas zonas.

Cuando se tiene información de los datos, el contraste puede llevarse a cabo en R con la función `ZTest()`.

5.1.2. Diferencia de medias de dos poblaciones independientes que se distribuyen normalmente y las variancias poblacionales son desconocidas

Considerando variancias iguales

Si las muestras provienen de poblaciones que se distribuyen normalmente, entonces se tiene que:

$$T_{\sigma_X=\sigma_Y} = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{S_p\sqrt{\dfrac{1}{n_X} + \dfrac{1}{n_Y}}} \sim t_{n_x+n_y-2}$$

donde:

$$S_p^2 = \frac{(n_X - 1)S_1^2 + (n_Y - 1)S_2^2}{n_X + n_Y - 2}$$

Considerando variancias diferentes

Si las muestras provienen de poblaciones que se distribuyen normalmente, entonces se tiene que:

$$T_{\sigma_X \neq \sigma_Y} = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}} \sim t_\nu$$

donde:

$$\nu = \frac{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right)^2}{\frac{(s_X^2/n_X)^2}{n_X - 1} + \frac{(s_Y^2/n_Y)^2}{n_Y - 1}}$$

Ejemplo 5.3 Se realizó un estudio para determinar si en las universidades públicas se tiene mejor rendimiento que en las universidades privadas en el curso de Estadística. A continuación, la Tabla 5.2 presenta las notas promedio en el curso de Estadística pertenecientes a estudiantes procedentes de ambos tipos de universidades.

Tabla 5.2: Notas obtenidas en el curso de Estadística según el tipo de universidad

	1	2	3	4	5	6	7	8	9	10
Público	93	82	75	74	45	59	59	94	90	46
Privado	67	60	70	76	40	72	52	51	73	44

Se desea conocer, con un nivel de significancia del 3%, si una persona que pertenece a la universidad pública, tiene un rendimiento promedio mayor al de una persona procedente de la universidad privada en el curso de Estadística.

Solución

$H_0 : \mu_X \leq \mu_Y$: El rendimiento promedio real en el curso de Estadística en la universidad pública es menor o igual al rendimiento promedio en la universidad privada.

$H_1 : \mu_X > \mu_Y$: El rendimiento promedio real en el curso de Estadística es significativamente mayor en las universidades públicas.

Primero se evalúa si las variancias son iguales.

```
> pri <- c(67, 60, 70, 76, 40, 72, 52, 51, 73, 44)
> pub <- c(93, 82, 75, 74, 45, 59, 59, 94, 90, 46)
> var.test(pub, pri, conf.level = 0.97)$conf.int

[1] 0.432478 9.687743
attr(,"conf.level")
[1] 0.97
```

Como las variancias son iguales puede realizarse el contraste utilizando la función `t.test()` indicando el argumento `var.equal = T`.

```
> t.test(x=pub, y=pri, alternative = "greater", var.equal = T,
+        conf.level = 0.97)$p.value

[1] 0.06771821
```

Como 0,07 > 0,03 no se rechaza la hipótesis nula. Por lo tanto, no existen evidencias, al nivel del 3%, para decir que el rendimiento promedio real en el curso es significativamente mayor en las universidades públicas.

Considerando la estadística de prueba se tiene que:

```
> alpha <- 0.03; nx = 10; ny = 10
> talpha <- qt(1 - alpha, nx + ny - 2)
> t <- t.test(x=pub, y=pri, alternative = "greater", var.equal = T,
+        conf.level = 0.97)$statistic
> c(talpha, t)

                t
2.007067 1.563112
```

El resultado indica que 1,56 < 2,01, es decir, 1,56 está en zona de no rechazo. Por lo tanto, no existen evidencias, al nivel del 3%, para decir que el rendimiento promedio real en el curso es significativamente mayor en las universidades públicas.

Ejemplo 5.4 Suponga que el nivel de colesterol total en la sangre se distribuye normalmente. En una investigación se desea verificar si el nivel promedio de colesterol total presenta diferencias por zona de residencia asumiendo que las variancias poblacionales son iguales. Para ello se consideran 12 personas que residen en zona urbana y obtiene que el nivel promedio de colesterol es de $\bar{x} = 215,9$

mg/dl con $s_X = 49,8$. Luego se consideran 16 personas residentes en zona rural y se obtiene que el nivel de colesterol promedio, para esta muestra, es de $\bar{y} = 211,5$ mg/dl con $s_Y = 49,1$. Utilice un nivel de significancia del 5%.

Solución

$H_0 : \mu_X = \mu_Y$: No existen diferencias en el nivel promedio de colesterol total por zona de residencia.

$H_1 : \mu_X \neq \mu_Y$: Existen diferencias significativas en el nivel promedio de colesterol total por zona de residencia.

En el caso de variancias poblacionales no conocidas pero que se consideran iguales, puede utilizarse la función `TTestA()` del paquete `DescTools` (Signorell et al., 2021).

```
> library(DescTools)
> TTestA(mx = 214.1, sx = 49.6, nx = 12, my = 211.5, sy = 49.1, ny = 16,
+ conf.level = 0.95, alternative = "two.sided", var.equal = T)$p.value

[1] 0.891251
```

Como 0,89 > 0,05 no se rechaza la hipótesis nula. Por lo tanto, no existen evidencias, al nivel del 5%, para decir que el nivel promedio de colesterol total es significativamente distinto por zona de residencia.

Ejemplo 5.5 Suponga que los datos del Ejemplo 5.3 son los que muestra la Tabla 5.3 con respecto a las notas promedio en el curso de Estadística pertenecientes a estudiantes procedentes de ambos tipos de universidades.

Tabla 5.3: Notas obtenidas en el curso de Estadística según el tipo de universidad

	1	2	3	4	5	6	7	8	9	10
Público	71	73	85	78	63	66	69	82	75	67
Privado	46	56	72	57	38	63	92	74	102	32

Se desea conocer, con un nivel de significancia del 3%, si una persona que pertenece a la universidad pública, rinde en promedio más que una persona procedente de la universidad privada en el curso de Estadística.

Ejemplo 5.6 Suponga que se tienen los salarios (en colones) de personas docentes que impartieron los cursos de Estadística en las universidades del Ejemplo 5.3, los cuales se observan en la Tabla 5.4.

Tabla 5.4: Salarios de personas docentes del curso de Probabilidad y Estadística según el tipo de universidad

	1	2	3	4	5	6	7	8
Público	652410	644870	655262	661082	625477	656929	636905	635530
Privado	725113	672358	744125	729287	719284	718231	678718	NA

Se desea conocer, con un nivel de significancia del 2%, si la diferencia en el salario medio de docentes que impartieron los cursos de Estadística en universidades privadas y públicas es a lo sumo ¢50000.

Ejemplo 5.7 Suponga que un grupo de estudiantes del curso de Cálculo I, se dividen en quienes cursaron Matemática General antes de llevar Cálculo I y los que no. Los datos de las notas finales del curso de Cálculo I por grupo se resumen en la Tabla 5.5.

Tabla 5.5: Notas obtenidas en el curso de Cálculo I

No cursaron Matemática General			Si cursaron Matemática General		
6,65	8,38	9,20	7,79	9,17	10,05
5,76	5,83	7,89	7,11	6,31	8,78
7,27	7,70	7,77	6,27	8,39	8,24
6,53	5,86	6,48	7,22	6,19	7,08
8,09	5,53	8,28	8,83	6,39	8,86
9,56	6,54	8,01	10,5	7,17	8,58

a) Con un nivel de significancia del 7%, podría decirse que los estudiantes que llevaron el curso de Matemática General tienen notas significativamente mayores en el curso de Cálculo I que los que no llevaron Matemática General.

b) Calcule el valor p para esta prueba.

c) Calcule la potencia de prueba que sea capaz de ver una diferencia de al menos 2 puntos.

Ejemplo 5.8 Se efectuó un estudio para comparar dos tratamientos que reducen los niveles de estrés en personas que laboran en el sector privado. Los resultados se midieron mediante un índice de salud percibida, que va de 0 a 10, donde 0 indica ninguna mejoría, determinado mediante un cuestionario. Se asignaron 8 pacientes de forma aleatoria a cada uno de los grupos de tratamiento y se obtuvieron los siguientes resultados:

Tabla 5.6: Índice de salud de personas trabajadoras

	1	2	3	4	5	6	7	8
Tratamiento A	6,7	6,8	7,0	6,0	5,3	6,3	5,1	6,0
Tratamiento B	7,3	7,3	9,3	10,9	8,4	8,2	9,8	9,0

Con un nivel de significancia del 5%, ¿cuál de los dos tratamientos es más efectivo para combatir el estrés?

5.2. Contraste de hipótesis para la diferencia de proporciones, para muestras grandes

Para muestras aleatorias independientes de tamaño n_X y n_Y, provenientes de poblaciones $Bernoulli(P_X)$ y $Bernoulli(P_Y)$, respectivamente.

$$\hat{p}_X - \hat{p}_Y \backsimeq \left(P_X - P_Y, \sqrt{\frac{P_X(1-P_X)}{n_X} + \frac{P_Y(1-P_Y)}{n_Y}} \right)$$

Es decir, la variable aleatoria

$$Z = \frac{(\hat{p}_X - \hat{p}_Y) - (P_X - P_Y)}{\sqrt{\frac{P_X(1-P_X)}{n_X} + \frac{P_Y(1-P_Y)}{n_Y}}}$$

tiene una distribución aproximadamente normal estándar.

Desafortunadamente, los valores de P_X y P_Y son desconocidos. Una forma de aproximarlos es tomando a $\hat{p}_X$ y $\hat{p}_Y$ como los estadísticos que aproximarán los respectivos valores poblacionales. Es decir,

$$Z = \frac{\hat{p}_X - \hat{p}_Y - \delta_0}{\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n_X} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{n_Y}}}$$

Considerando el supuesto de que H_0 es cierta, la variable aleatoria

$$Z = \frac{\hat{p}_X - \hat{p}_Y - \delta_0}{\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n_X} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{n_Y}}}$$

tiene una distribución aproximadamente normal estándar. Por lo tanto, el estadístico de prueba está dado por z, tal que:

$$z = \frac{\hat{p}_X - \hat{p}_Y - \delta_0}{\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n_X} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{n_Y}}}$$

De esta manera, la hipótesis nula se rechaza si:

Tabla 5.7: Región de rechazo para el contraste de hipótesis para la diferencia de dos proporciones considerando el estadístico z

H. Nula	**H. Alternativa**	**Región de rechazo**
$H_0 : P_X - P_Y = \delta_0$	$H_1 : P_X - P_Y \neq \delta_0$	$z < -z_{\frac{\alpha}{2}}$ ó $z > z_{\frac{\alpha}{2}}$
$H_0 : P_X - P_Y \leq \delta_0$	$H_1 : P_X - P_Y > \delta_0$	$z > z_\alpha$
$H_0 : P_X - P_Y \geq \delta_0$	$H_1 : P_X - P_Y < \delta_0$	$z < -z_\alpha$

Considerando que $\chi^2 = Z^2$ es una variable aleatoria con distribución aproximadamente chi-cuadrada con un grado de libertad, el estadístico de prueba queda definido por χ^2 y para un nivel de significancia α, la hipótesis nula se rechaza:

Tabla 5.8: Región de rechazo para el contraste de hipótesis para la diferencia de dos proporciones considerando el estadístico chi-cuadrado

H. Nula	**H. Alternativa**	**Región de rechazo**
$H_0 : P_X - P_Y = \delta_0$	$H_1 : P_X - P_Y \neq \delta_0$	$\chi^2 > \chi^2_\alpha$
$H_0 : P_X - P_Y \leq \delta_0$	$H_1 : P_X - P_Y > \delta_0$	$\chi^2 > \chi^2_\alpha$
$H_0 : P_X - P_Y \geq \delta_0$	$H_1 : P_X - P_Y < \delta_0$	$\chi^2 > \chi^2_\alpha$

Sin embargo, para el caso en que δ_0 sea cero, lo cual es muy común, se acostumbra a usar un estadístico ponderado para estimar la proporción poblacional P, pues en este caso se tendría que $P_X = P_Y = P$. El estadístico ponderado de P, denotado como $\hat{p}$, es:

$$\hat{p} = \frac{X + Y}{n_X + n_Y}$$

lo que representa una simple estimación de la proporción total de éxitos. En este caso el estadístico de prueba sería:

$$Z = \frac{\hat{p}_X - \hat{p}_Y}{\sqrt{\hat{p}(1-\hat{p}) \cdot \frac{1}{n_X} + \frac{1}{n_Y})}}$$

Cuando $|\hat{p}_X - \hat{p}_Y| > \frac{1}{2}\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)$, se acostumbra usar la corrección por continuidad, que se aplica de manera automática con el comando `prop.test()` de R.

Ejemplo 5.9 La compañía BIOMAC. S.A. fabrica productos medicinales a base de drogas extraídas de los bosques tropicales y está probando dos nuevos compuestos destinados a reducir los síntomas producidos por la diabetes tipo II. Los compuestos se suministran, de manera independiente, a dos conjuntos de animales en el laboratorio; en el primer grupo, 71 de 100 animales respondieron positivamente a la droga 1, mientras que 58 de 90 animales respondieron positivamente al suministrarles la droga 2. La compañía desea probar con un nivel de significancia del 5% si existe diferencia entre la eficacia de ambas drogas, ¿realmente la droga 1 será más efectiva que la droga 2?

Ejemplo 5.10 El municipio de una ciudad utiliza dos métodos para cobrar los impuestos relacionados con la carga tributaria según el nivel de renta de las personas. El primero requiere que cada persona se presente a proporcionar la información; el segundo método permite a las personas brindar la información por medio de correo electrónico, accediendo a la página de la municipalidad. En el municipio se piensa que el primer método produce menos errores en la recolección de la información que el segundo y autoriza un estudio de 50 personas que fueron personalmente a declarar los impuestos y 75 que lo hicieron vía Internet. El 10% de las fórmulas recolectadas por el primer método tenían errores y el $13,3\%$ de las recolectadas por el segundo método también poseían errores. En el municipio se desea determinar con un nivel de significancia del 5%, si el primer método produce una proporción de errores menor que el segundo, ¿qué podría decirse en este caso?

Ejemplo 5.11 Durante el proceso de selección para el ingreso a una prestigiosa universidad durante el año 2019, se observa que de 100 personas residentes en la zona A que solicitaron ingreso, solo 50 fueron admitidas en la carrera que anotaron como primera opción, mientras tanto de 100 personas residentes en la zona B que solicitaron ingreso, solo 21 fueron admitidas en la carrera de su preferencia, ¿representan estos datos evidencia suficiente, con un nivel de significancia de un 3%, para afirmar que las personas residentes en la zona A presentan una proporción de éxito mayor al 10% que las de la zona B a la hora de ser admitidas en la carrera de su preferencia?

Capítulo 6

Medidas de asociación y modelos de regresión lineal

6.1. Medidas de asociación

6.1.1. Medidas de asociación para variables categóricas

Algunas veces es necesario realizar análisis de datos binomiales que dan lugar a tablas de clasificación de r filas por c columnas. En este tipo de problemas interesa contrastar la hipótesis nula:

H_0: Las variables aleatorias son independientes,

contra la hipótesis alternativa:

H_1: Las variables aleatorias no son independientes.

O bien, contrastar la hipótesis nula.

$H_0 = p_{i1} = p_{i2} = \cdots = p_{ic}$: La probabilidad de obtener una observación en la i-ésima fila es la misma en cada categoría de las columnas.

Contra la hipótesis alternativa

H_1: Al menos una probabilidad es significativamente diferente.

Estas pruebas pueden aplicarse, por ejemplo, cuando personas pertenecientes a distintos grupos (por zona, por sexo, por grupo de edad, etc.) se clasifican según las categorías de otro grupo (grupo sanguíneo, peso, categoría de enfermedad, etc.). Es decir, la clasificación de los datos se presenta en tablas de r filas por c columnas. Algunas medidas para realizar este tipo de contraste se mencionan a continuación.

Estadístico χ^2 (chi-cuadrado)

Este estadístico de prueba para el análisis de tablas de tamaño $r \times c$ es dado por:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

donde $E_{ij} = \dfrac{n_i \cdot n_j}{n}$, siendo

n_i: el número de elementos de la categoría de la fila i.

n_j: el número de elementos de la categoría de la columna j.

n: el total de elementos.

De esta manera, la hipótesis nula es rechazada si el valor del estadístico χ^2 excede el valor χ^2_α con $r-1$ por $c-1$ grados de libertad, es decir:

$$\chi^2 > \chi^2_{\alpha,\,(r-1)(c-1)}$$

Ejemplo 6.1 Se desea investigar si el tipo de colegio de procendencia influye en el rendimiento de un curso. Considere los datos de la Tabla 6.1 y utilice un nivel de significancia de 2% para determinar si el tipo de colegio tiene efecto en el rendimiento del curso.

Tabla 6.1: Distribución de personas por tipo de colegio y condición de aprobación

	Público	Privado	Subvencionado
Aprobación	21	64	17
No aprobación	16	49	14

Solución

H_0: el tipo de colegio de procedencia y el rendimiento en el curso son independientes.

H_1: el tipo de colegio de procedencia y el rendimiento en el curso no son independientes.

```
> alpha <- 0.02; r <- 2; c <- 3
> eje1<-matrix(c(21,64,17,16,49,14), byrow = T, ncol=3)
> chi2 <- chisq.test(eje1)$statistic
> chialpha <- qchisq(1 - alpha, (r - 1)*(c - 1))
> c(chialpha, chi2)

            X-squared
7.82404601 0.03506294
```

Como 0,035 < 7,824, la hipótesis nula no se rechaza. No existen evidencias, al nivel de 2%, para indicar que las variables tipo de colegio y rendimiento del curso no son independientes. Es decir, el tipo de colegio no tiene efecto en el rendimiento del curso.

En la Figura 6.1, la región destacada con color rojo representa la zona de rechazo para el contraste del Ejemplo 6.1. Un código de R para la construcción de la Figura 6.1 puede verse en el Anexo 4.

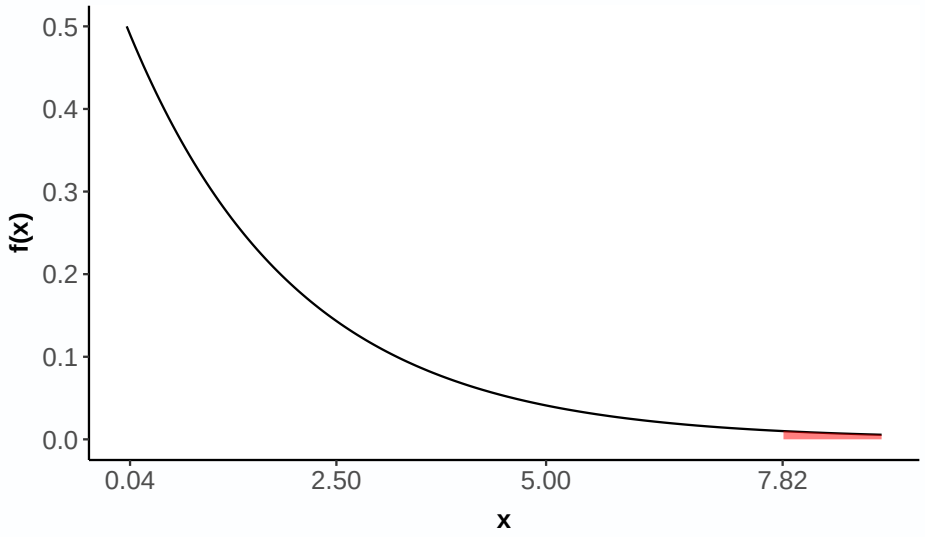


Figura 6.1: Región de rechazo para la prueba de independencia $\chi^2_{0,02}$ con 1×2 grados de libertad

Fuente: Elaboración propia

Ejemplo 6.2 Se ha mencionado que las personas con sobrepeso tienden a presentar niveles elevados de presión arterial. Con el fin de verificar esta relación considere los datos de la Tabla 6.2 e investigue si el peso de la persona tiene efecto en el nivel de presión arterial utilizando un nivel de significancia de 1%.

Tabla 6.2: Distribución de personas por peso y condición de hipertensión

	Normal	Sobrepeso	Obesidad
No hipertensión	15	12	8
Hipertensión	10	13	28

Solución

H_0: el peso de la persona y el nivel de presión arterial son independientes.

H_1: el peso de la persona y el nivel de presión arterial no son independientes.

La prueba chi-cuadrado se muestra a continuación.

```
> alpha <- 0.01; r <- 2; c <- 3
> eje2 <- matrix(c(15,12,8,10,13,28), 2, 3, byrow = TRUE)
> chi2 <- chisq.test(eje2)$statistic
> chialpha <- qchisq(1 - alpha, (r - 1)*(c - 1))
> c(chialpha, chi2)

          X-squared
 9.210340  9.503308
```

Como 9,5 > 9,21, la hipótesis nula se rechaza. Existen evidencias, al nivel de 1%, para indicar que el peso de la persona y el nivel de presión arterial no son independientes. La región de rechazo se representa gráficamente en la Figura 6.2.

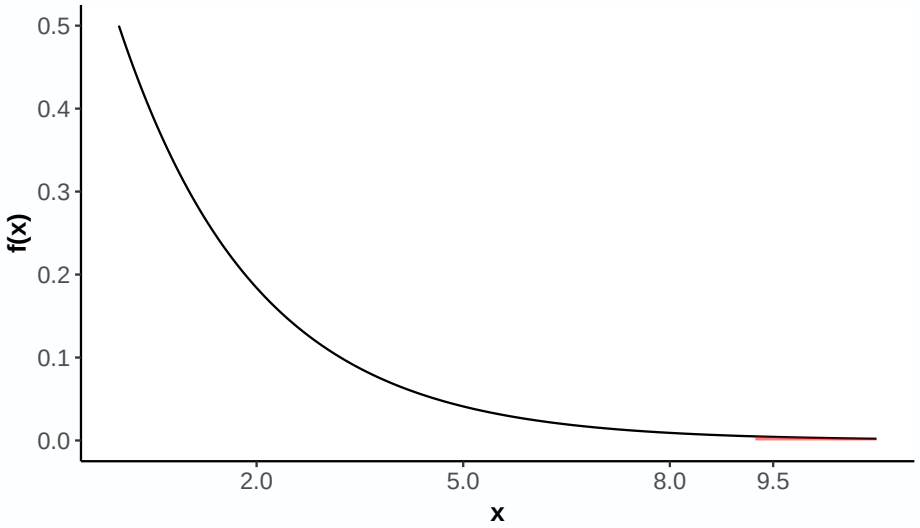


Figura 6.2: Región de rechazo para la prueba de independencia $\chi^2_{0,01}$ con 1×2 grados de libertad

Fuente: Elaboración propia

Si se considera el valor p, la hipótesis nula se rechaza si $p < \alpha$. En este caso se tiene que:

```
> alpha <- 0.01; r <- 2; c <- 3
> eje2 <- matrix(c(15,12,8,10,13,28), 2, 3, byrow = TRUE)
> p <- chisq.test(eje2)$p.value; p

[1] 0.008637398
```

Como 0,0086 < 0,01, la hipótesis nula se rechaza. Existen evidencias, al nivel de 1%, para indicar que el peso de la persona y el nivel de presión arterial no son independientes.

En algunas ocasiones la prueba chi-cuadrado puede resultar no apropiada, esto ocurre cuando las muestras son pequeñas (se tienen frecuencias esperadas menores a 5 o la proporción de celdas con frecuencias esperadas menores a 5 supera el 20%).

Cuando esto ocurre, suele utilizarse el test de independencia de Fisher o Fishers test. Esta prueba puede llevarse a cabo en R empleando la función `fisher.test()`, la cual utiliza un nivel de significancia de 5% por defecto.

Ejemplo 6.3 Considere que los datos de sobrepeso y condición de hipertensión son los que se muestran en la Tabla 6.3. Investigue si el peso de la persona tiene efecto en el nivel de presión arterial utilizando un nivel de significancia del 5%.

Tabla 6.3: Distribución de personas por peso y condición de hipertensión

	Normal	Sobrepeso	Obesidad
No hipertensión	7	5	3
Hipertensión	3	5	12

Solución

```
> alpha <- 0.05; r <- 2; c <- 3
> eje3 <- matrix(c(7,5,3,3,5,12), 2, 3, byrow = TRUE)
> chi2 <- chisq.test(eje3)$statistic; chi2

X-squared
 6.416667
```

La prueba chi-cuadrado muestra una advertencia, por lo que es necesario aplicar la prueba de independencia de Fischer.

```
> alpha <- 0.05; r <- 2; c <- 3
> eje3 <- matrix(c(7,5,3,3,5,12), 2, 3, byrow = TRUE)
> fisher.test(eje3, conf.int = T)

        Fisher's Exact Test for Count Data

data:  eje3
p-value = 0.05313
alternative hypothesis: two.sided
```

Como 0,0531 > 0,05, la hipótesis nula no se rechaza. No existen evidencias, al nivel de 5%, para indicar que el peso de la persona y el nivel de presión arterial no son independientes.

Razón de verosimilitud

La razón de verosimilitud (Fischer, 1924; Neyman y Pearson, 1928) es un estadístico asintóticamente equivalente al χ^2, esto es, se distribuye e interpreta de la misma forma que el χ^2. Este estadístico se denota por L^2 y se define por:

$$L^2 = 2\sum_i \sum_j \Big[O_{ij} \cdot ln\Big(\frac{O_{ij}}{E_{ij}}\Big)\Big]$$

Es muy utilizado en el estudio de variables categóricas, sobre todo en el contexto de los modelos log-lineales.

6.1.2. Intensidad de la asociación entre variables categóricas

Entre las medidas más utilizadas para medir la intensidad de la relación entre dos variables categóricas se tienen: el cuadrado medio de contingencia Phi, el coeficiente V de Cramer, el coeficiente de contingencia de Pearson, entre otras.

Cuadrado medio de contingencia Phi

Para un tamaño de muestra n dado, el coeficiente de contingencia phi, denotado por ϕ, puede calcularse para tablas 2×2 mediante la expresión:

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

Coeficiente V de Cramer

El coeficiente V de Cramer es una medida que permite obtener la fuerza de asociación entre dos variables categóricas. Puede utilizarse en tablas de contingencias de tamaño $r \times c$. Su valor nunca excede a 1 y en tablas 2×2 coincide con el coeficiente Phi. El coeficiente V de Cramer se define por:

$$V_{cramer} = \sqrt{\frac{\chi^2}{n \cdot m}}$$

La fuerza o el nivel de asociación que identifica este coeficiente puede clasificarse según la siguiente escala:

$0,00 \leq V \leq 0,30$: Ninguna o baja asociación entre las variables.

$0,30 < V \leq 0,50$: Asociación moderada baja entre las variables.

$0,50 < V \leq 0,75$ Asociación moderada alta entre las variables.

$0,75 < V \leq 0,75$ Alta o perfecta asociación entre las variables.

Coeficiente de contingencia de Pearson

Este coeficiente se denota por C y se define por:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

Este coeficiente nunca alcanza el valor de 1, pero sí un valor máximo que depende de la cantidad de filas y columnas de la tabla en estudio. Este valor máximo se denota por C_{max} y se define por

$$C_{max} = \sqrt{\frac{q-1}{q}}$$

,

donde $q = min\{r, c\}$ para una tabla de dimensión $r \times c$.

En R, estas medidas pueden calcularse empleando la función `assocstats()` del paquete `vcdExtra` (Friendly, 2017).

Ejemplo 6.4 Considere que los datos de aprobación y tipo de colegio de procedencia son los que se muestran en la Tabla 6.4. Investigue si el tipo de colegio de procedencia de la persona tiene efecto en la condición de aprobación utilizando un nivel de significancia del 5%.

Tabla 6.4: Distribución de personas por tipo de colegio y condición de aprobación

	Público	Privado
No aprobación	7	5
Aprobación	5	12

Solución

```
> #library(vcdExtra)
> eje3 <- matrix(c(7,5,5,12), 2, 2, byrow = TRUE)
> assocstats(eje3)

                    X^2 df P(> X^2)
Likelihood Ratio 2.4384  1  0.11840
Pearson          2.4257  1  0.11936

Phi-Coefficient   : 0.289
Contingency Coeff.: 0.278
Cramer's V        : 0.289
```

Ejemplo 6.5 Considere los datos la Tabla 6.2 y determine las medidas de asociación estudiadas con el fin de determinar la intensidad de la relación.

Solución

```
> #library(vcdExtra)
> eje3 <- matrix(c(15,12,8,10,13,28), 2, 3, byrow = TRUE)
> assocstats(eje3)

                    X^2 df  P(> X^2)
Likelihood Ratio 9.8204  2 0.0073711
Pearson          9.5033  2 0.0086374

Phi-Coefficient   : NA
Contingency Coeff.: 0.315
Cramer's V        : 0.332
```

6.1.3. Medidas de asociación para variables continuas

6.1.4. Coeficiente de correlación

En el caso de variables continuas medidas a partir de muestras aleatorias, el nivel y la dirección de asociación lineal entre dos variables pueden medirse por medio del coeficiente de correlación r de Pearson. Para la medición de este nivel de asociación se considera que las variables presentan una distribución aproximadamente normal. El coeficiente r de Pearson es tal que $-1 \leq r \leq 1$ en donde:

$r = -1$ indica relación lineal perfecta e inversa, es decir, el aumento de una variable implica la disminución de la otra.

$r = 1$ indica relación lineal perfecta directa, esto es, el aumento de una variable implica el aumento de la otra.

$r = 0$ indica ausencia de relación lineal entre las variables.

Este coeficiente puede calcularse mediante la expresión:

$$\begin{aligned} r &= \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} \\ &= \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}} \end{aligned}$$

Diagrama de dispersión

Una idea inicial de la posible correlación o asociación lineal entre dos variables puede tenerse gráficamente con la ayuda de un diagrama de dispersión, el cual es un gráfico en dos dimensiones donde el eje de las abscisas representa a la variable independiente y el eje de las ordenadas a la variable dependiente, de tal forma que cada valor se dibuja como un par ordenado (X, Y).

Ejemplo 6.6 Se tiene que uno de los factores que altera la presión arterial de las personas es el sobrepeso u obesidad. Considere el siguiente conjunto de datos sobre el índice de masa corporal (IMC) y nivel de presión arterial sistólica (PS) de la persona. Construya un diagrama de dispersión e investigue la existencia de una posible relación lineal entre las variables.

Tabla 6.5: IMC y presión arterial.

	1	2	3	4	5	6	7
IMC	24	26	27	29	31	33	35
PS	121	125	133	129	135	137	130

Solución

La idea gráfica de una posible relación entre las variables mencionadas puede verse en la Figura 6.3.

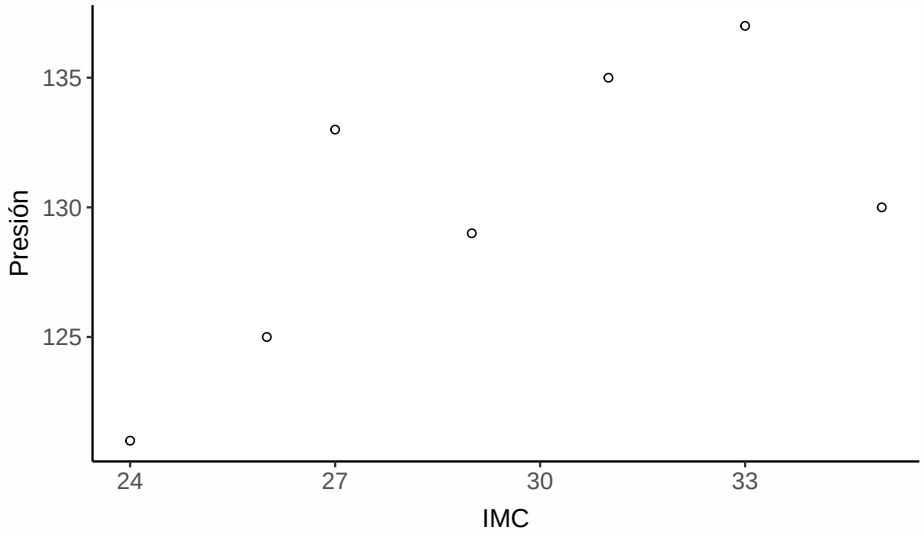


Figura 6.3: Diagrama de dispersión para la relación entre el IMC y la presión sistólica

Fuente: Elaboración propia

El cálculo del coeficiente de correlación de Pearson puede realizarse en R con ayuda de la función `cor()`.

```
> cor(imc, ps)

[1] 0.6905119
```

El valor de 0,6905 indica una relación moderada y directa, por lo que puede decirse que si el IMC aumenta entonces el nivel de presión arterial también aumenta.

Uno de los supuestos para obtener el coeficiente de correlación de Pearson es que las distribuciones de las variables deben ser aproximadamente normales. Para evaluarlo, puede utilizarse la prueba Shapiro-Wilk que considera la hipótesis nula de que la distribucción de la variable es normal. En R puede obtenerse empleando la función `shapiro.test()`.

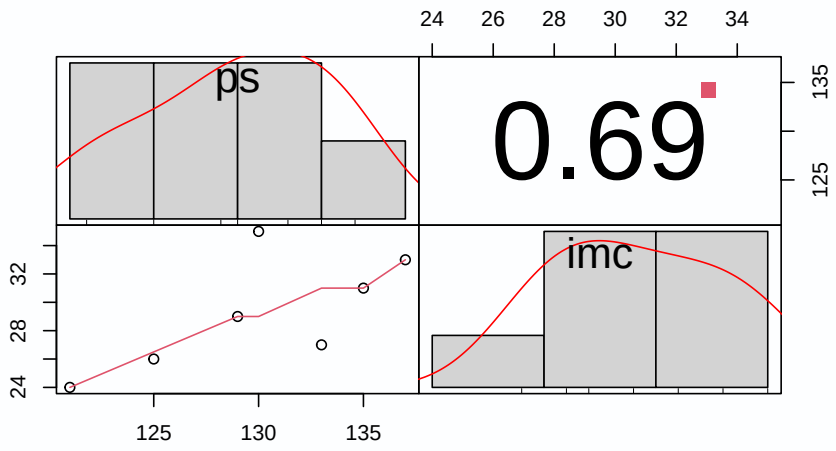
Ejemplo 6.7 Considere la función `shapiro.test()` y los datos del Ejemplo 6.6 para verificar la normalidad de las variables de estudio.

```
> shapiro.test(imc); shapiro.test(ps)

        Shapiro-Wilk normality test

data:  imc
W = 0.97409, p-value = 0.9263

        Shapiro-Wilk normality test

data:  ps
W = 0.96839, p-value = 0.8866
```

Ambas variables presentan una distribución normal, por lo que el coeficiente de Pearson es adecuado.

Utilizando la función `chart.Correlation()` del paquete `PerformanceAnalytics` (Peterson y Carl, 2020) también puede estudiarse la correlación entre dos variables.

```
> #install.packages("PerformanceAnalytics")
> PerformanceAnalytics::chart.Correlation(cbind(ps, imc), method = "pearson")
```



Si el supuesto de normalidad no se cumple, la correlación entre las variables puede obtenerse aplicando el coeficiente de correlación de Spearman, denotado por ρ, el cual permite medir el nivel de asociación lineal (independencia) entre dos variables cuantitativas, donde al menos una presenta una distribución asimétrica. Su interpretación es similar a la del coeficiente de Pearson y su valor se encuentra en el rango de -1 a 1 ($-1 \leq \rho \leq 1$).

El coeficiente de correlación de Spearman puede obtnerse en R con ayuda de la función `cor()` utilizando el argumento `method =` "`spearman`".

Ejemplo 6.8 Otro de los factores que altera la presión arterial de las personas es la edad. Considere el siguiente conjunto de datos sobre la edad y nivel de presión arterial sistólica (PS) de la persona e investigue la existencia de una posible relación lineal entre las variables.

Tabla 6.6: Edad y presión arterial.

	1	2	3	4	5	6	7
Edad	24	26	34	33	35	33	24
PS	121	125	133	129	135	137	130

Una idea gráfica de la relación entre las variables puede observarse en la Figura 6.4.

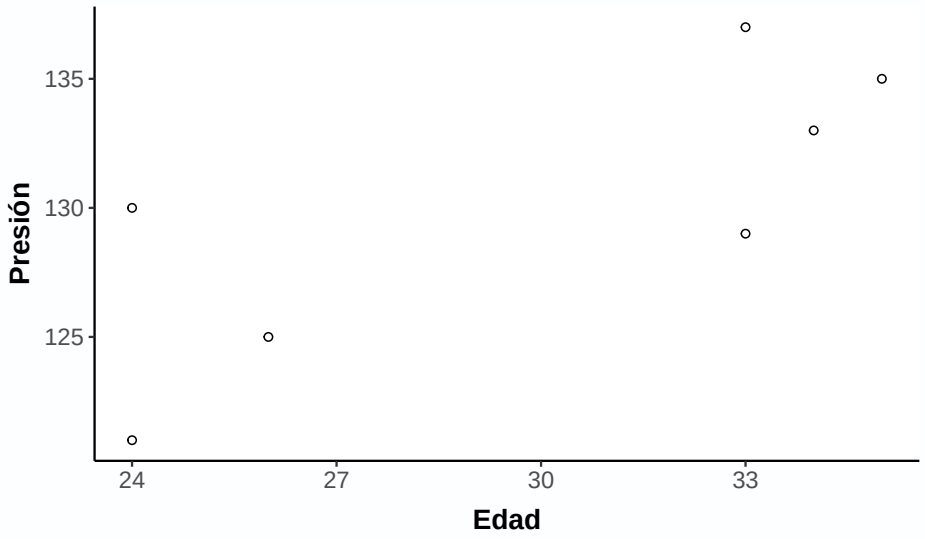


Figura 6.4: Diagrama de dispersión para la relación entre la edad y la presión sistólica

Fuente: Elaboración propia

Evaluando el supuesto de normalidad se tiene que:

```
> shapiro.test(edad)

        Shapiro-Wilk normality test

data:  edad
W = 0.80491, p-value = 0.04579
```

Como la variable edad no es normal, entonces procede calcular el coeficiente de correlación de Spearman.

```
> cor(edad, ps, method = "spearman")

[1] 0.6728385
```

Cuando se calcula un coeficiente de correlación puede evaluarse la significancia de dicho coeficiente. En este caso se contrasta la hipótesis nula que indica que el coeficiente de correlación es igual a cero. Este contraste puede realizarse en R con ayuda de la función `cor.test()` e indicando como argumento cuál coeficiente se está contrastando a través del argumento `method`.

Ejemplo 6.9 Considere los datos del Ejemplo 6.8 y evalúe la significancia del coeficiente de correlación que se ha calculado.

Solución

H_0: El coeficiente de correlación real entre la variable edad y la presión sistólica es igual a cero.

H_1: El coeficiente de correlación real entre la variable edad y la presión sistólica es significativamente distinto de cero.

```
> p <- cor.test(edad, ps, method = "spearman", exact = F)$p.value; p

[1] 0.09764906
```

Como $0{,}0976 > 0{,}05$, puede concluirse que la hipótesis nula no se rechaza. Es decir, no existe evidencias, al nivel del 5%, para decir que el coeficiente de correlación de Spearman entre la variable edad y presión sistólica es significativamente distinto de cero.

En este caso existe una asociación moderada entre las variables de estudio.

6.2. Regresión lineal

El análisis de regresión es utilizado para modelar relaciones entre una variable Y, llamada variable respuesta (dependiente, explicada, predicha o endógena) y una o más variables conocidas como variables explicativas (independientes, predictoras o exógenas) $x_1, x_2, \ldots, x_{p-1}$. En el análisis de regresión lineal, la variable Y debe ser continua, pero las predictoras pueden ser continuas, discretas o categóricas.

El término regresión fue introducido en estadística por el científico británico Sir Francis Galton cuando realizaba estudios sobre la estatura de ciertas poblaciones. Galton encontró una tendencia en que los padres de estatura alta tenían hijos altos y los padres de estatura baja tenían hijos de estatura baja, sin embargo, la estatura promedio de los niños nacidos de padres de una estatura dada tendían a moverse o ***regresar*** hacia la estatura promedio de la población total, de ahí el término regresión.

No obstante, la interpretación moderna del término de regresión es mucho más amplia (Gujarati, 2009).

El análisis de regresión trata del estudio de la dependencia de la variable respuesta, respecto a una o más variables explicativas con el objetivo de estimar o predecir la media o valor promedio poblacional de la primera en términos de los valores conocidos o fijos (en muestras repetidas) de las últimas.

Por ejemplo, considere el experimento de Galton, él estaba interesado en averiguar las razones por las cuales existía estabilidad en la distribución de las estaturas de una cierta población. Con el enfoque moderno, lo que interesa es hallar la forma en que cambia la estatura promedio de los hijos dada la estatura de los padres, es decir, predecir la estatura promedio de los hijos conociendo la estatura de sus padres.

6.2.1. Modelos probabilísticos vrs modelos determinísticos

En los modelos de regresión estudiados interesa analizar la dependencia estadística entre las variables, no la dependencia funcional o determinística, ya que en una relación determinística las variables NO son aleatorias o estocásticas. Por ejemplo, si la tarifa de autobús entre San José y Heredia es de ₡300 puede generarse un modelo entre la variable de interés (dígase cantidad de dinero recolectada en una carrera) y la variable independiente o de predicción (número de personas que pagan su pasaje), es decir, se forma una ecuación lineal que determina un valor exacto de y (cantidad de dinero recolectada en una carrera) para cada valor x de X (cantidad de dinero recolectada en una carrera).

Para nuestro ejemplo la ecuación sería $y = 300x$, de tal manera que si se montan al autobús 15 pasajeros, el dinero recolectado es de $y = 300 \cdot 15 = 4500$.

Por otra parte, cuando se trabaja con modelos estadísticos o probabilísticos las variables que intervienen son aleatorias y tienen distribuciones de probabilidad, por lo que no es posible determinar exactamente el valor de la variable dependiente y se debe agregar a la ecuación un componente de error aleatorio.

6.2.2. Modelo de regresión

El modelo de regresión lineal simple puede escribirse de la siguiente manera:

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{6.1}$$

Donde β_0 y β_1 son el intercepto y la pendiente, respectivamente, y ϵ es el componente de error que puede usarse para comprender la relación en el modelo lineal entre las dos variables.

Por ejemplo, suponga que quiere predecirse la nota final de un estudiante que está matriculado en el curso de Estadística. En la nota final de un curso intervienen muchos factores tales como asistencia, cantidad de horas de estudio que se dedican, motivación, entre otros. Dado que es imposible considerar todos los aspectos que intervienen, se describe un modelo probabilístico lineal de dos variables: $Y = \beta_0 + \beta_1 x + \epsilon$, donde la variable dependiente Y es la nota final del curso de Estadística y la variable explicativa X es el número de horas de estudio semanal que se le dedica al curso y ϵ es el componente de error aleatorio. Los datos se resumen a continuación (Tabla 6.7).

Tabla 6.7: UNA: Nota en el curso de Estadística según número de horas de estudio independiente.

	1	2	3	4	5	6	7	8	9	10
Horas	2,1	3,5	4,0	5,5	0,0	8,4	9,3	1,0	2,5	4,5
Nota Final	65,4	60,3	75,6	80,4	38,9	87,6	95,2	50,4	60,7	75,8

Considere, por el momento, que el modelo resultante para nuestro ejemplo es el siguiente $Y = 4{,}3X + 45{,}4$. Suponga que una persona estudia 5 horas, según el modelo puede decirse que esta persona obtendría una nota de $y = 4{,}3 \cdot 5 + 45{,}4 = 66{,}9$. No obstante, eso no significa que si la persona dedica 5 horas semanales a la materia su nota final será de 66.9, es tan solo una aproximación y esta es la diferencia con respecto al modelo determinístico donde sí se da un valor exacto. Como las variables que intervienen en los modelos probabilísticos son aleatorias, no es posible predecir exactamente el valor de la variable dependiente. Para efectos de este curso se trabaja con modelos probabilísticos, a menos que se indique lo contrario.

Los diagramas de dispersión son muy útiles ya que es un método informal de apreciar si existe algún tipo de relación entre las variables en estudio. La naturaleza de la relación puede tomar muchas formas, como tendencias lineales, cuadráticas o exponenciales, entre otras. El diagrama de dispersión para los datos de la Tabla 6.7 se muestra en la Figura 6.5.

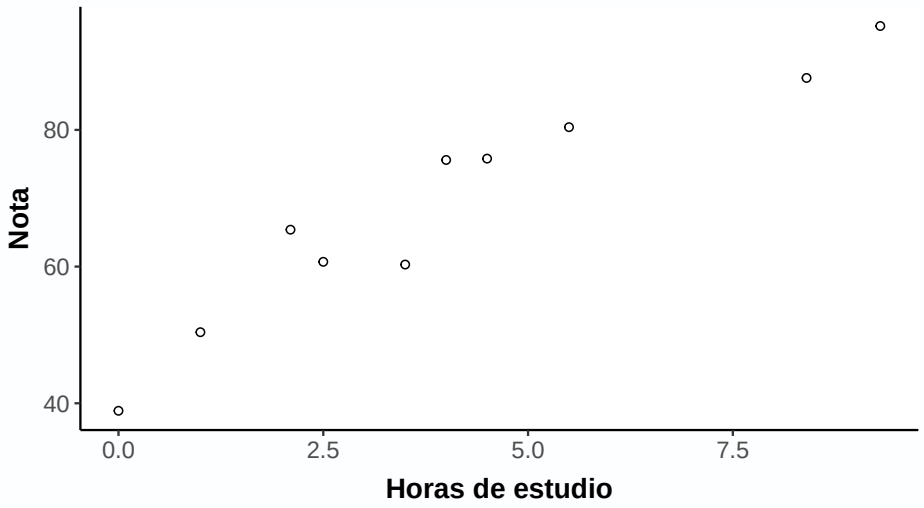


Figura 6.5: Diagrama de dispersión que muestra la relación entre el tiempo dedicado al estudio y la nota obtenida en el curso

Fuente: Elaboración propia

6.2.3. Método de mínimos cuadrados ordinarios

Para calcular el valor de los estimadores para los parámetros β_0 y β_1 se utiliza el método de mínimos cuadrados ordinarios, el cual permite minimizar los errores aleatorios del modelo lineal a partir del cumplimiento de determinados supuestos.

El método estima los parámetros minimizando la suma de las desviaciones al cuadrado entre los Y observados con respecto a sus valores esperados (predichos), es decir, minimiza la suma de los residuos al cuadrado. Esta suma se denota por SSR (por sus siglas en inglés) y se define por:

$$SSR = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} [y_i - E(y_i|x_i)]^2 = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 \tag{6.2}$$

Si se toma la Ecuación 6.2 y se deriva con respecto a β_0 se tiene que:

$$\begin{aligned}\frac{\partial SSR}{\partial \beta_0} &= \frac{\partial\left(\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2\right)}{\partial \beta_0}\\ &= -2\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)\\ &= -2\left(\sum_{i=1}^{n} y_i - n\beta_0 - \beta_1\sum_{i=1}^{n} x_i\right)\end{aligned}$$

Considerando la Ecuación 6.2 y derivando con respecto a β_1 se tiene que:

$$\begin{aligned}\frac{\partial SSR}{\partial \beta_1} &= \frac{\partial\left(\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2\right)}{\partial \beta_1}\\ &= -2\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)\cdot x_i\\ &= -2\sum_{i=1}^{n}(y_i\cdot x_i-\beta_0\cdot x_i-\beta_1 x_i^2)\\ &= -2\left(\sum_{i=1}^{n} y_i\cdot x_i - \beta_0\sum_{i=1}^{n} x_i - \beta_1\sum_{i=1}^{n} x_i^2\right)\end{aligned}$$

Igualando las derivadas anteriores a cero se tiene que las ecuaciones

$$\frac{\partial SSR}{\partial \beta_0} = 0, \quad \frac{\partial SSR}{\partial \beta_1} = 0$$

se denominan ecuaciones de mínimos cuadrados, las cuales permitirán estimar los parámetros de la recta de regresión. Las ecuaciones de mínimos cuadrados son lineales en β_0 y β_1 y por lo tanto pueden resolverse simultáneamente.

La solución del sistema de ecuaciones anterior es:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} \tag{6.3}$$

Además,

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{6.4}$$

Estimadores de mínimos cuadrados ordinarios para el modelo de regresión lineal simple

Antes de realizar el cálculo de los estimadores de mínimos cuadrados, es conveniente enumerar algunas expresiones relacionadas con los valores muestrales X, Y que se utilizan con frecuencia en el contexto de modelos de regresión, que son:

1. $S_{XX} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$, o bien, $S_{XX} = \sum_{i=1}^{n} (x_i - \bar{x})^2$
2. $S_{YY} = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$, o bien, $S_{YY} = \sum_{i=1}^{n} (y_i - \bar{y})^2$
3. $S_{XY} = \sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}$, o bien, $S_{XY} = \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})$
4. $SSR = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$
5. $\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}$
6. $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
7. $SSR = SSY - SSM$
8. $SSM = \hat{\beta}_1 \cdot S_{XY}$

El modelo de regresión lineal simple puede estimarse en R con ayuda de la función `lm()`. El modelo debe expresarse de la forma $y \sim x$, donde y es la variable **dependiente** y x es la variable **independiente**.

Ejemplo 6.10 Se tiene que uno de los factores que altera la presión arterial de las personas es el sobrepeso u obesidad. Considere el siguiente conjunto de datos sobre el índice de masa corporal (IMC) y nivel de presión arterial sistólica (PS) de la persona e investigue la existencia de una posible relación lineal entre las variables.

Tabla 6.8: IMC y presión arterial.

	1	2	3	4	5	6	7
IMC	24	26	27	29	31	33	35
PS	121	125	133	129	135	137	130

Solución

El diagrama de dispersión ayudar a dar una una idea gráfica de la relación. El código de R para la contrucción de la Figura 6.6 puede observarse en el Anexo 5.

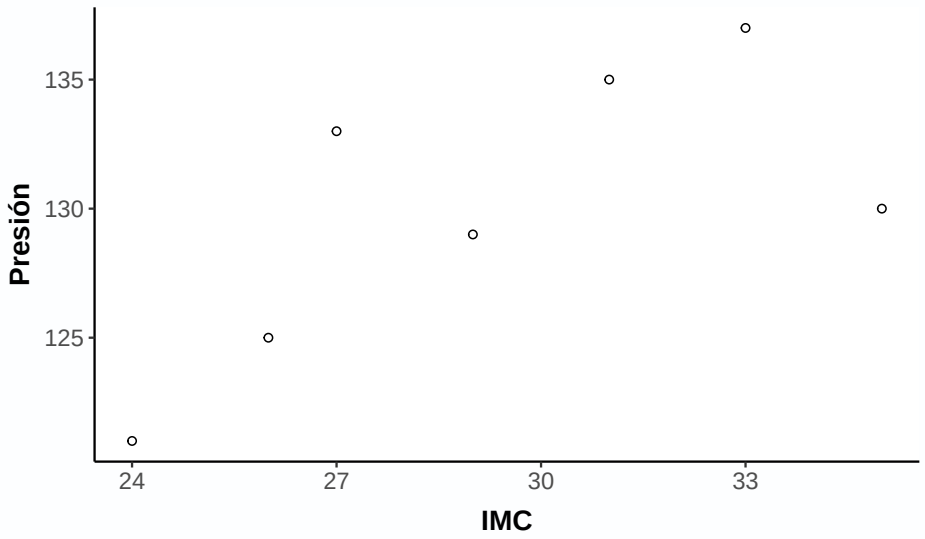


Figura 6.6: Diagrama de dispersión para la relación entre el IMC y la presión sistólica

Fuente: Elaboración propia

El cálculo del coeficiente de correlación de Pearson puede realizarse en R con ayuda de la función `cor()`.

```
> cor(imc, ps)

[1] 0.6905119
```

El modelo de regresión estimado puede obtenerse como sigue:

```
> reg1 <- lm(ps ~ imc); reg1

Call:
lm(formula = ps ~ imc)

Coefficients:
(Intercept)          imc
   101.1621       0.9847
```

Una vez que el modelo ha sido estimado, pueden obtenerse diferentes expresiones que complementan el análisis, entre ellos, el conjunto de datos que dieron origen al modelo, los valores estimados, los residuos, los residuos estandarizados, etc.

Para visualizar el conjunto de datos que dan origen al modelo puede escribirse la siguiente instrucción.

```
> lm(ps ~ imc)$model

   ps imc
1 121  24
2 125  26
3 133  27
4 129  29
5 135  31
6 137  33
7 130  35
```

Los valores estimados del modelo pueden obtenerse como sigue:

```
> round(lm(ps ~ imc)$fitted, 2)

     1      2      3      4      5      6      7
124.80 126.76 127.75 129.72 131.69 133.66 135.63
```

Recuerde que la diferencia entre los valores observados y los valores estimados son los residuos del modelo, los cuales pueden obtenerse de la siguiente manera:

```
> round(lm(ps ~ imc)$residuals, 2)

    1     2     3     4     5     6     7
-3.80 -1.76  5.25 -0.72  3.31  3.34 -5.63
```

Los grados de libertad del modelo (n menos el número de coeficientes a estimar) pueden obtenerse de la siguiente manera:

```
> lm(ps ~ imc)$df

[1] 5
```

La función `summary()` permite obtener un resumen de la estimación del modelo.

```
> summary(reg1)

Call:
lm(formula = ps ~ imc)

Residuals:
      1       2       3       4       5       6       7
-3.7951 -1.7645  5.2508 -0.7187  3.3119  3.3425 -5.6269

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 101.1621    13.6142   7.431 0.000696 ***
imc           0.9847     0.4613   2.135 0.085898 .
---
Signif. codes:  0
```

La recta de regresión estimada puede representarse gráficamente mediante el empleo de la función `visreg()` del paquete `visreg` (Breheny y Burchett, 2017).

```
> #install.packages("visreg")
> visreg::visreg(reg1, xlab="IMC", ylab="Presión")
```

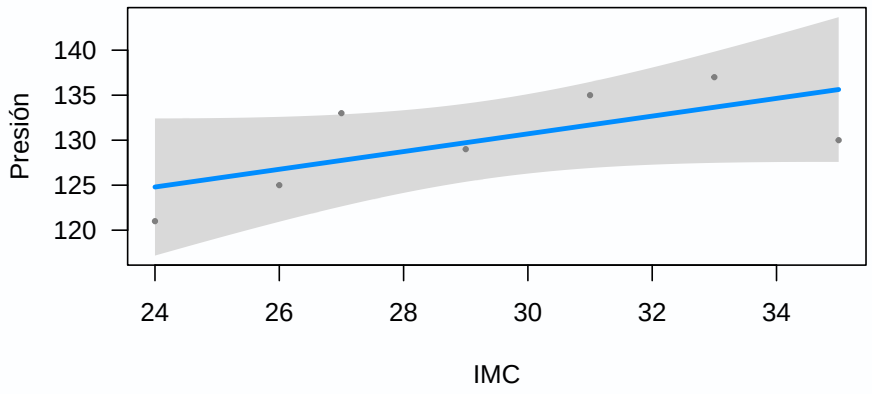


Figura 6.7: Recta de mínimos cuadrados que muestra la relación entre el IMC y la presión sistólica

Fuente: Elaboración propia

Otra forma es utilizar la función `ggplot()` y la opción `geom_smooth()`. El código de R para la contrucción de la Figura 6.8 puede observarse en el Anexo 6.

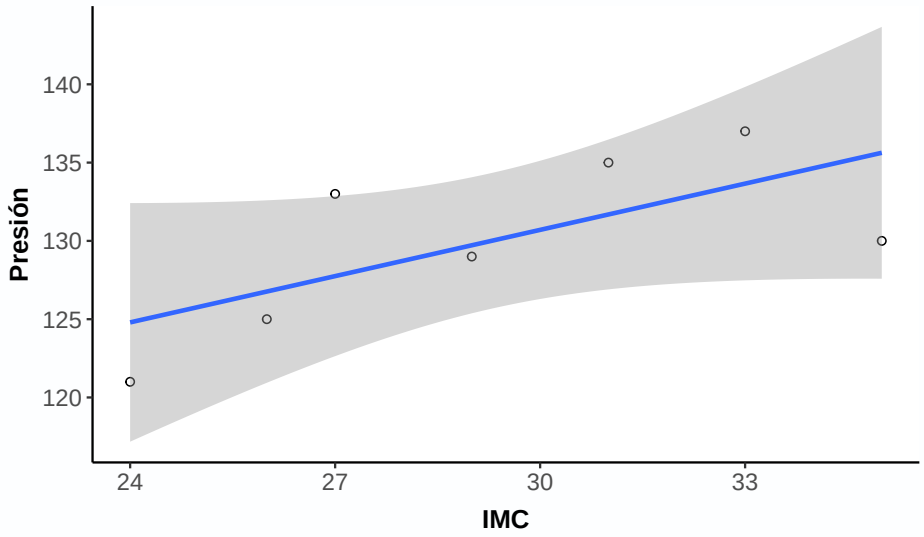


Figura 6.8: Recta de mínimos cuadrados que muestra la relación entre el IMC y la presión sistólica

Fuente: Elaboración propia

Puede obtenerse cada coeficiente de regresión por medio de la instrucción `modelo$coef[1]` y `modelo$coef[2]`, donde la palabra `modelo` hace referencia al nombre del modelo de regresión que ha sido estimado.

```
> c(reg1$coef[1], reg1$coef[2])

(Intercept)         imc
101.1620795   0.9847095
```

También puede obtenerse el error estándar residual, denotado por S_e y definido por:

$$S_e = \sqrt{\frac{\sum_{i=1}^{n} Y_i^2 - \hat{\beta}_0 \sum_{i=1}^{n} Y_i - \hat{\beta}_1 \sum_{i=1}^{n} (X_i Y_i)}{n-2}}$$

O bien,

$$S_e = \sqrt{\frac{\sum_{i=1}^{n} \epsilon^2}{n-2}}$$

```
> summary(reg1)$sigma

[1] 4.45885
```

Se dice que el modelo de regresión es significativo si existen evidencias para indicar que la pendiente de la recta de regresión verdadera es distinta de cero. Esto es, $H_0 : \beta_1 = 0$.

Si $\beta_1 = 0$, entonces el modelo tiene la forma:

$$Y = \beta_0 + \epsilon$$

Esto significa que la variación en Y no es afectada por los cambios ocurridos en X y está únicamente sujeta a cambios aleatorios en torno a la constante β_0.

Si $\beta_1 \neq 0$, entonces una porción de la variación de Y se explica por el hecho de que se observa Y con diversos valores de X.

Esto hace que el modelo de regresión sea útil para predecir valores de Y dado X.

Una hipótesis relacionada con la pendiente real β_1 del modelo de regresión puede plantearse de tres formas distintas, considerando a β_{1_0} para denotar el valor nulo de la pendiente se tiene que:

H. Nula	**H. Alternativa**	**Valor p**
$H_0 : \beta_1 = \beta_{1_0}$	$H_1 : \beta_1 \neq \beta_{1_0}$	$2 \cdot \mathbb{P}(T \geq \lvert t_{obs} \rvert)$
$H_0 : \beta_1 \leq \beta_{1_0}$	$H_1 : \beta_1 > \beta_{1_0}$	$\mathbb{P}(T \geq \lvert t_{obs} \rvert)$
$H_0 : \beta_1 \geq \beta_{1_0}$	$H_1 : \beta_1 < \beta_{1_0}$	$\mathbb{P}(T \geq \lvert t_{obs} \rvert)$

Tabla 6.9: Cálculo del valor p para el contraste de hipótesis para la pendiente de la recta de regresión

El **estadístico de prueba** para llevar a cabo el contraste de hipótesis para la pendiente real de la recta de regresión está dado por t, tal que:

$$T = \frac{(\hat{\beta}_1 - \beta_1)}{S_e} \cdot \sqrt{S_{xx}}$$

donde $T \sim t_{n-2}$.

Una hipótesis relacionada con la intersección real β_0 del modelo de regresión puede plantearse de tres formas distintas; considerando a β_{0_0} para denotar el valor nulo de la intersección se tiene que:

H. Nula	**H. Alternativa**	**Valor p**
$H_0 : \beta_0 = \beta_{0_0}$	$H_1 : \beta_0 \neq \beta_{0_0}$	$2 \cdot \mathbb{P}(T \geq \lvert t_{obs} \rvert)$
$H_0 : \beta_0 \leq \beta_{0_0}$	$H_1 : \beta_0 > \beta_{0_0}$	$\mathbb{P}(T \geq \lvert t_{obs} \rvert)$
$H_0 : \beta_0 \geq \beta_{0_0}$	$H_1 : \beta_0 < \beta_{0_0}$	$\mathbb{P}(T \geq \lvert t_{obs} \rvert)$

Tabla 6.10: Cálculo del valor p para el contraste de hipótesis para la intersección de la recta de regresión

El **estadístico de prueba** para llevar a cabo el contraste de hipótesis para la intersección real de la recta de regresión está dado por t, tal que:

$$T = \frac{(\hat{\beta}_0 - \beta_0)}{S_e} \cdot \sqrt{\frac{nS_{xx}}{S_{xx} + n(\bar{x})^2}}$$

donde $T \sim T_{n-2}$.

Ejemplo 6.11 Determine la significancia del modelo de regresión que plantea la relación entre el IMC y la presión arterial sistólica según los datos del Ejemplo 6.10.

Solución

$H_0 : \beta_1 = 0$: La pendiente de la recta de regresión real es cero. No existe relación lineal entre el IMC y la presión arterial sistólica.

$H_1 : \beta_1 \neq 0$: La pendiente de la recta de regresión real es significativamente distinta de cero. Existe relación lineal significativa entre el IMC y la presión arterial sistólica.

```
> n <- length(imc); xbarra <- mean(imc); alpha <- 0.05
> se <- summary(reg1)$sigma
> sxx <- sum((imc - mean(imc))^2)
> tca <- ((reg1$coef[1] - 0)/se)*sqrt(n*sxx/(sxx + n*xbarra^2))
> tcb <- ((reg1$coef[2] - 0)/se)*sqrt(sxx)
> c(pa <- 2*(1 - pt(abs(tca), n - 2)), pb <- 2*(1 - pt(abs(tcb), n - 2)))

(Intercept)         imc
0.000695648 0.085897815
```

O bien,

```
> coef(summary(reg1))[c(7,8)]

[1] 0.000695648 0.085897815
```

Como 0,0859 > 0,05, no se rechaza la hipótesis nula, es decir, no existen evidencias, al nivel del 5%, para indicar que existe relación lineal significativa entre el IMC y la presión arterial sistólica.

Dado el modelo regresión lineal estimado $Y = \hat{\beta}_0 + \hat{\beta}_1 X$, un intervalo de confianza del $100(1-\alpha)\%$ para los coeficientes de regresión viene dado por:

$$\beta_0 : \hat{\beta}_0 \pm t_{\frac{\alpha}{2},\, n-2} \cdot S_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

$$\beta_1 : \hat{\beta}_1 \pm t_{\frac{\alpha}{2},\, n-2} \cdot S_e \sqrt{\frac{1}{S_{xx}}}$$

Ejemplo 6.12 Determine el intervalo de confianza del 95% para los coeficientes del modelo de regresión que plantea la relación entre el IMC y la presión arterial sistólica según los datos del Ejemplo 6.10.

Solución

```
> alpha <- 0.05; alpham <- alpha/2; n <- length(imc)
> se <- summary(reg1)$sigma; talpham <- qt(1 - alpham, n - 2)
> sxx <- sum((imc - mean(imc))^2)
> lia <- reg1$coef[1] - talpham*se*sqrt(1/n + mean(imc)^2/sxx)
> lsa <- reg1$coef[1] + talpham*se*sqrt(1/n + mean(imc)^2/sxx)
> lib <- reg1$coef[2] - talpham*se/sqrt(sxx)
> lsb <- reg1$coef[2] + talpham*se/sqrt(sxx)
> c(lia, lsa); c(lib, lsb)

(Intercept) (Intercept)
   66.16567   136.15849

       imc         imc
-0.2010988  2.1705177
```

O bien,

```
> confint(reg1, level = 0.95)

                 2.5 %     97.5 %
(Intercept) 66.1656656 136.158493
imc         -0.2010988   2.170518
```

6.2.4. Coeficiente de determinación

Cuando se aplica un modelo de regresión, es importante saber qué tan bien ajusta el modelo la regresión y para ello existe una medida llamada coeficiente de determinación R^2 que mide la proporción o el porcentaje de la variación total en Y, que es explicada por el modelo de regresión que se está utilizando.

$$\begin{aligned} R^2 &= 1 - \frac{\sum_{i=1}^{n} \epsilon_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \\ &= 1 - \frac{SSR}{S_{YY}} \end{aligned}$$

Ejemplo 6.13 Determine el coeficiente de determinación para el modelo de regresión estimado con los datos del Ejemplo 6.10.

Solución

En R el coeficiente de determinación puede obtenerse de la siguiente manera:

```
> summary(reg1)[8]

$r.squared
[1] 0.4768067
```

El 47,68% de la variabilidad en la presión sistólica es explicado por la variabilidad del IMC.

6.2.5. Supuestos del modelo de regresión simple

Para que las estimaciones de los parámetros del modelo de regresión lineal (usando el método de mínimos cuadrados) sean adecuadas, es necesario el cumplimiento de una serie de supuestos que garanticen una interpretación válida de la información. Dichos supuestos son:

1. El modelo de regresión es lineal en los parámetros, es decir, $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$.
2. Los valores de X son fijos en el muestreo repetido.
3. El valor esperado de las discrepancias aleatoria es cero, es decir, $E(\epsilon_i|X = x_i) = 0$.
4. Homocedasticidad o igualdad de las variancias de las ϵ_i, es decir, $Var[Y|X = x] = Var(\epsilon) = \sigma^2$.
5. No existe autocorrelación entre las perturbaciones o errores aleatorios, es decir, la covariancia es cero, $Cov(\epsilon_i, \epsilon_j|X = x_i, X = x_j) = 0$.
6. $Cov(\epsilon_i|X = x_i) = 0$.
7. El número de observaciones n debe ser mayor que el número de parámetros por estimar.
8. Variabilidad en los valores de X, es decir, no todos los valores de X en una muestra deben ser iguales.
9. El modelo de regresión está correctamente especificado, es decir, que los datos se ajustan para trabajar un modelo lineal.
10. No hay multicolinealidad perfecta (en el caso de modelos de regresión múltiples).

Como consecuencia de los supuestos, dado un valor x_i de la variable aleatoria X, se tiene que:

Además, dado que el término de error se distribuye normalmente, el modelo de regresión lineal simple viene dado por:

$$Y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i$$

donde

Y_i: es el valor de la variable respuesta para el i-ésimo ensayo.

β_0 y β_1: son los parámetros del modelo.

x_i: es una constante conocida para el i-ésimo ensayo.

ϵ_i: es el término de error aleatorio, que se asume, tiene una distribución $N(0, \sigma^2)$, donde σ^2 es la variancia que usualmnete es desconocida.

Además, el valor esperado y la variancia para el modelo están dados por:

$$E[Y|X = x] = \beta_0 + \beta_1 x \tag{6.5}$$

$$Var[Y|X = x] = Var(\epsilon) = \sigma^2 \tag{6.6}$$

Ejemplo 6.14 Verifique el supuesto de normalidad de los residuos del modelo de regresión que plantea la relación entre el IMC y la presión arterial sistólica según los datos del Ejemplo 6.10.

Solución

Uno de los supuestos del modelo indica que los residuos se distribuyen de manera normal con media 0 y variancia σ^2, esto es,

$$\epsilon \sim N(0,\ \sigma^2)$$

Para corroborar este supuesto puede utilizarse la prueba de normalidad **Shapiro-Wilk**. Las hipótesis de la prueba se plantean como sigue:

H_0: Los residuos se distribuyen de manera normal.

H_1: Los residuos no se distribuyen de manera normal.

```
> shapiro.test(reg1$residuals)

        Shapiro-Wilk normality test

data:  reg1$residuals
W = 0.93819, p-value = 0.6224
```

Como $0,6224 > 0,05$ no se rechaza la hipótesis nula. Puede concluirse que no existen evidencias, al nivel del 5%, para indicar que los residuos no se distribuyen de manera normal.

Por otro lado, los **residuos estandarizados** permiten eliminar cualquier influencia generada por la presencia de valores extremos. Para obtenerlos se utiliza la siguiente expresión:

$$r_{est} = \frac{e_i}{\sigma \cdot \sqrt{1 - h_i}}$$

donde,

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$$

Los residuos estandarizados pueden obtenerse en R mediante el uso de la función `rstandar()`.

Ejemplo 6.15 Obtenga los residuos estandarizados del modelo de regresión que plantea la relación entre el IMC y la presión arterial sistólica según los datos del Ejemplo 6.10.

Solución

```
> round(rstandard(reg1), 2)

    1     2     3     4     5     6     7
-1.14 -0.46  1.32 -0.17  0.82  0.89 -1.77
```

Gráficamente, la **normalidad** de los residuos puede verse por medio de un gráfico **Q−Q normal** (Q−Q Normal Plot), el cual dibuja la distribución normal teórica contra los residuos estandarizados.

Para que exista normalidad los puntos deben estar distribuidos cercanos a una línea recta. Esta gráfica puede construirse en R con ayuda de la función `qqnorm()`.

Ejemplo 6.16 Dibuje el **gráfico Q−Q normal** para verificar el supuesto de normalidad de lo residuos del modelo de regresión que plantea la relación entre el IMC y la presión arterial sistólica según los datos del Ejemplo 6.10.

Solución

```
> qqnorm(rstandard(reg1), ylab = "standardized residual")
> qqline(rstandard(reg1), col= "blue")
```

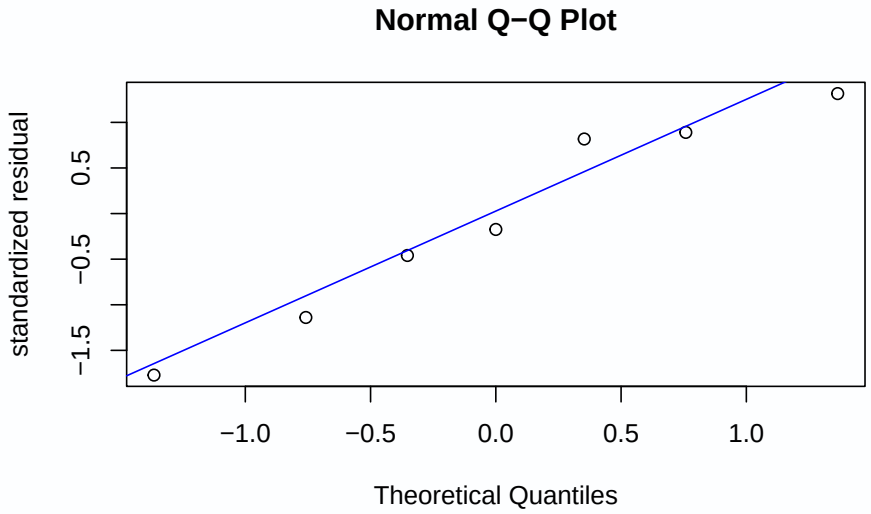


Figura 6.9: Región de rechazo para la prueba de independencia $\chi^2_{0,05}$ con 1×2 grados de libertad

Fuente: Elaboración propia

Otro supuesto del modelo de regresión es el principio de varianza constante, también conocido como **homocedasticidad**. En el caso de que la hipótesis nula se rechace se dice que la varianza de los errores no es constante, o bien, que el modelo de regresión presenta **heterocedasticidad**.

Ejemplo 6.17 Verifique el supuesto de homocedasticidad del modelo de regresión que plantea la relación entre el IMC y la presión arterial sistólica según los datos del Ejemplo 6.10.

Solución

Las hipótesis del contraste son las siguientes:

H_0: La varianza de los residuos es constante.

H_1: La varianza de los residuos no es constante.

Este contraste puede realizarse en R mediante el uso de la función `ncvTest()` (non constant variance Test) del paquete `car` (Fox y Weisberg, 2019).

```
> #install.packages("car")
> car::ncvTest(reg1)

Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.2231585, Df = 1, p = 0.63664
```

Como $0,6366 > 0,05$ no se rechaza la hipótesis nula. No existe evidencia, al nivel del 5%, para indicar que la varianza de los residuos no es constante.

Bondad del ajuste, el ANOVA

Otra forma de evaluar la bondad del ajuste es por medio del análisis de varianza (ANOVA), el cual compara la variabilidad explicada por el modelo (SSM: sum of square of the model) con la que no es explicada por el modelo (SSR: sum of square of the residuals). Si la primera suma es mayor que la segunda para valores pequeños de $p, p < 0,05$ se dice que el modelo tiene capacidad predictiva significativa.

En caso contrario, se dice que el modelo no presenta capacidad predictiva, o bien, la variable X no aporta capacidad predictiva sobre la variable respuesta Y. El análisis de varianza puede llevarse a cabo en R utilizando la función `anova()`.

Ejemplo 6.18 Determine la bondad del ajuste del modelo de regresión que plantea la relación entre el IMC y la presión arterial sistólica según los datos del Ejemplo 6.10.

Solución

```
> anova(reg1)

Analysis of Variance Table

Response: ps
          Df Sum Sq Mean Sq F value Pr(>F)
imc        1 90.593  90.593  4.5567 0.0859 .
Residuals  5 99.407  19.881
---
Signif. codes:  0
```

Como puede observarse, la suma de cuadrados del modelo es menor que la suma de cuadrados de los residuos, por lo que puede considerarse que el modelo no aporta capacidad predictiva.

El modelo y las predicciones

Una vez aceptado el modelo, pueden realizarse predicciones de valores de la variable respuesta para nuevas observaciones utilizando la ecuación de la recta de regresión. En R, esta acción puede efectuarse con ayuda de la función `predict()`.

Ejemplo 6.19 Determine el valor promedio de la presión sistólica para personas con un índice de masa corporal de $25 kg/m^2$ utilizando el modelo de regresión que plantea la relación entre el IMC y la presión arterial sistólica según los datos del Ejemplo 6.10.

Solución

Para predecir el valor promedio de la presión sistólica para personas con un índice de masa corporal de 25 se ejecuta la siguiente instrucción:

```
> predict(reg1, data.frame(imc = 25))[1]

       1
125.7798
```

Se espera que, en promedio una persona con un IMC de $25 kg/m^2$ presenta una presión arterial sistólica de 125,78 $mmHg$.

Pueden obtenerse además, intervalos de confianza para las predicciones de las nuevas observaciones. En R estos intervalos se obtienen utilizando el argumento `interval =` "`prediction`" en la función `predict()`.

Ejemplo 6.20 Obtenga el intervalo de confianza del 95% para el valor promedio de la presión sistólica para una persona con un índice de masa corporal de $25 kg/m^2$ utilizando el modelo de regresión que plantea la relación entre el IMC y la presión arterial sistólica según los datos del Ejemplo 6.10.

Solución

```
> predict(reg1, data.frame(imc = 25), interval = "prediction")

       fit      lwr      upr
1 125.7798 112.5145 139.0451
```

Con una confianza del 95%, la verdadera presión sistólica para un IMC de $25 kg/m^2$ se encuentra entre 112,51 y 139,05 $mmHg$.

También puede obtenerse el intervalo de confianza para el valor promedio de las predicciones de las nuevas observaciones. En R estos intervalos se obtienen utilizando el argumento `interval =` "`confidence`" en la función `predict()`.

Ejemplo 6.21 Obtenga el intervalo de confianza del 95% para el valor promedio de la presión sistólica para un índice de masa corporal promedio de $25 kg/m^2$ utilizando el modelo de regresión que plantea la relación entre el IMC y la presión arterial sistólica según los datos del Ejemplo 6.10.

Solución

```
> predict(reg1, data.frame(imc = 25), interval = "confidence")

       fit      lwr      upr
1 125.7798 119.1019 132.4577
```

Con una confianza del 95%, la verdadera presión sitólica promedio se encuentra entre 119,1 y 132,46 $mmHg$.

Ejemplo 6.22 Otro de los factores que altera la presión arterial de las personas es la edad. Considere el siguiente conjunto de datos sobre la edad y el nivel de presión arterial sistólica (PS) de la persona e investigue la existencia de una posible relación lineal entre las variables.

Tabla 6.11: Edad y presión arterial.

	1	2	3	4	5	6	7
Edad	24	26	34	33	35	36	28
PS	121	125	133	129	135	137	130

Identifique la existencia de una posible asociación lineal entre las variables y modele dicha relación.

Ejemplo 6.23 En la empresa ALECO S.A el gerente de producción sospecha que existe una relación entre la edad de los trabajadores y el número de días que, en promedio, se ausentan del trabajo. Se propone seleccionar una muestra de 10 trabajadores, cuyos datos se presentan en la Tabla 6.12.

Tabla 6.12: ALECO S.A: Número promedio de días de ausentismo de las personas colaboradoras según edad.

	1	2	3	4	5	6	7	8	9	10
Edad	28	62	38	24	45	59	30	37	64	41
Ausentismo	14,2	7,5	11,3	19,7	10,1	8,7	14,4	12,1	5,9	6,3

a) Suponiendo un modelo lineal, utilice el método de mínimos cuadrados ordinarios para calcular los coeficientes de regresión.

b) ¿Cuántos días (como promedio) se esperaría que una persona de 39 años se ausente?

c) Trace la recta de regresión estimada.

d) Determine e interprete el coeficiente de correlación y de determinación, para estos datos.

e) Determine la bondad del ajuste del modelo por medio del ANOVA.

Ejemplo 6.24 Un ingeniero agrónomo quiere determinar el efecto que tendría el abono orgánico en el rendimiento de una plantación de chile. Se utilizan cuatro diferentes cantidades de abono orgánico en 10 lotes de terreno equivalentes a 0, 15, 25 y 30 kilogramos por cada 50 metros cuadrados de sembrado. Los niveles de abono se asignan de manera aleatoria a los lotes y los resultados se presentan en la Tabla 6.13.

Tabla 6.13: Rendimiento por lote según cantidad de fertilizante

	1	2	3	4	5	6	7	8	9	10	11	12
Cantidad	0	0	0	15	15	15	25	25	25	30	30	30
Rendimiento	6	9	10	14	19	22	24	28	31	35	32	33

a) Suponiendo un modelo lineal, utilice el método de mínimos cuadrados ordinarios para calcular los coeficientes de regresión.

b) Interprete el significado de la ordenada al origen y la pendiente para este problema.

c) Prediga el rendimiento promedio de chiles para un lote al que se le han aplicado 10 kilos de abono orgánico por cada 50 m^2 de sembrado.

d) Trace la recta de regresión estimada.

e) Determine e interprete el coeficiente de correlación y de determinación, para estos datos.

f) Determine la bondad del ajuste del modelo por medio del ANOVA.

Ejemplo 6.25 La Tabla 6.14 muestra los datos sobre la pesca de atún, en miles de toneladas, en la península de Nicoya y los precios de la harina de pescado, en dólares, durante los últimos 12 meses.

Tabla 6.14: Precio de la harina de pescado de acuerdo con el volumen de pesca de atún en la península de Nicoya

	1	2	3	4	5	6	7	8	9	10	11	12
Pesca	7,23	8,53	9,82	10,26	8,96	12,27	10,28	4,45	1,78	4,00	3,30	4,3
Precio	190	160	134	129	172	197	167	239	542	372	245	376

a) Suponiendo un modelo lineal, utilice el método de mínimos cuadrados ordinarios para calcular los coeficientes de regresión.

b) Interprete el significado de la ordenada al origen y la pendiente para este problema.

c) Prediga el precio promedio de la harina para un volumen de pesca de atún de 11 toneladas.

d) Trace la recta de regresión estimada.

e) Determine e interprete el coeficiente de correlación y de determinación, para estos datos.

f) Determine la bondad del ajuste del modelo por medio del ANOVA.

Referencias

Agresti, A., y Caffo, B. (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *The American Statistician*, *54*(4), 280–288. Descargado de https://estadistica-dma.ulpgc.es/estadFCM/pdf/agresti_caffo_2000.pdf

Agresti, A., y Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, *52*(2), 119–126. Descargado de https://www.ime.usp.br/~rbrito/docs/2014-05-06/2685469.pdf

Aguilar, E., y Zamora, J. A. (2020). *Introducción a la estadística descriptiva con R*. Costa Rica: EUNA. Descargado de https://www.euna.una.ac.cr/index.php/EUNA/catalog/book/255

Arnholt, A. T., y Evans, B. (2017). Bsda: Basic statistics and data analysis [Manual de software informático]. Descargado de https://CRAN.R-project.org/package=BSDA (R package version 1.2.0)

Breheny, P., y Burchett, W. (2017). Visualization of regression models using visreg. *The R Journal*, *9*(2), 56–71. Descargado de https://journal.r-project.org/archive/2017/RJ-2017-046/RJ-2017-046.pdf

Fisher, R. A. (1932). Inverse probability and the use of likelihood. En *Mathematical proceedings of the cambridge philosophical society* (Vol. 28, pp. 257–261).

Fisher, R. A. (1934). Two new properties of mathematical likelihood. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, *144*(852), 285–307. Descargado de https://royalsocietypublishing.org/doi/pdf/10.1098/rspa.1934.0050

Fox, J., y Weisberg, S. (2019). *An R companion to applied regression*. Thousand Oaks CA: Sage. Descargado de https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Freund, J. E., Miller, I., y Miller, M. (2004). *John E. Freund's mathematical statistics: With applications*. India: Pearson Education.

Friendly, M. (2017). vcdextra: 'vcd' extensions and additions [Manual de software informático]. Descargado de https://CRAN.R-project.org/package=vcdExtra (R package version 0.7-1)

Gujarati, D. N. (2009). *Basic econometrics*. New Delhi: Tata McGraw-Hill Education.

Peterson, B. G., y Carl, P. (2020). Performanceanalytics: Econometric tools for performance and risk analysis [Manual de software informático]. Descargado de https://CRAN.R-project.org/package=PerformanceAnalytics (R package version 2.0.4)

Pruim, R. (2018). fastr2: Foundations and applications of statistics using R (2nd edition) [Manual de software informático]. Descargado de https://CRAN.R-project.org/package=fastR2 (R package version 1.2.1)

R Core Team. (2021). R: A language and environment for statistical computing [Manual de software informático]. Vienna, Austria. Descargado de https://www.R-project.org/

RStudio Team. (2015). Rstudio: Integrated development environment for r [Manual de software informático]. Boston, MA. Descargado de http://www.rstudio.com/

Signorell, A., Aho, K., Alfons, A., Anderegg, N., Aragon, T., Arachchige, C., ... Zeileis, A. (2021). DescTools: Tools for descriptive statistics [Manual de software informático]. Descargado de https://cran.r-project.org/package=DescTools (R package version 0.99.40)

Subbiah, M., y Rajeswaran, V. (2017). Proportion: Inference on single binomial proportion and bayesian computations [Manual de software informático]. Descargado de https://CRAN.R-project.org/package=proportion (R package version 2.0.0)

Ugarte, M. D., Militino, A. F., y Arnholt, A. T. (2008). *Probability and statistics with R*. Florida, CRC Press.

Wackerly, D. D., Muñoz, R., y Humberto, J. (2010). *Estadística matemática con aplicaciones*. México: Cengage Learning.

Wasserstein, R. L., y Lazar, N. A. (2016). The asa statement on p-values: context, process, and purpose. *The American Statistician*, *70*(2), 129-133. doi: https://doi.org/10.1080/00031305.2016.1154108

Anexos

Anexo 1 Código de R para obtener la Figura 1.1.

Se realiza la simulación de la variable aleatoria con distribución uniforme, para $n = 100$ y se crea una base de datos (data frame) con los resultados de la simulación.

```
> set.seed(1002)
> du <- runif(n = 100, min = 2, max = 5)
> du1 <- data.frame(du)
> colnames(du1) <- "x"
```

Se cargan los paquetes `reshape2` y `ggplot2`.

```
> library(reshape2)
> library(ggplot2)
```

Se grafica los resultados de la simulación de la variable aleatoria X y se escriben las etiquetas de los ejes.

```
> hunif <- ggplot(du1, aes(x=x))
> hunif +
+   geom_histogram(binwidth=0.5, col='black', fill='green', alpha=0.3) +
+   xlab('X') + ylab('Frecuencia')
```

La función `theme()` de ggplot2 permite personalizar la apariencia de la gráfica. Esta función ejerce control sobre 3 tipos principales de componentes:

`Axis`: controla el título, la etiqueta, la línea y las marcas de los ejes.

`Background`: controla el color de fondo y las líneas de cuadrícula mayores y menores.

`Legend`: controla la posición, el texto, los símbolos y otros.

Se asigna el blanco como el color de fondo de la gráfica y se eliminan el fondo, las líneas de cuadrícula y los bordes de la gráfica.

```
> hunif + geom_histogram(binwidth=0.5, col='black', fill='green', alpha=0.3) +
+   xlab(expression(bar(X))) + ylab('Frecuencia') +
+   #tema con fondo blanco
+      theme_bw() +
+   #elimina el fondo, las líneas de cuadrícula y el borde
+      theme(
+        plot.background = element_blank()
+       ,panel.grid.major = element_blank()
+       ,panel.grid.minor = element_blank()
+       ,panel.border = element_blank()
+      )
```

Se dibujan las líneas de los ejes de la gráfica y se le da formato a sus respectivos elementos.

```
> hunif + geom_histogram(binwidth=0.5, col='black', fill='green', alpha=0.3) +
+   xlab(expression(bar(X))) + ylab('Frecuencia') +
+   #tema con fondo blanco
+      theme_bw() +
+   #elimina el fondo, las líneas de cuadrícula y el borde
+      theme(
+        plot.background = element_blank()
+       ,panel.grid.major = element_blank()
+       ,panel.grid.minor = element_blank()
+       ,panel.border = element_blank()
+      ) +
+   #dibuja las líneas y da formato a los elementos de los ejes
+     theme(axis.line = element_line(color = 'black')) +
+     theme(axis.title.x = element_text(face="bold", vjust=-0.5,
+     colour="black", size=rel(1))) +
+     theme(axis.title.y = element_text(face="bold", vjust=1.5,
+     colour="black", size=rel(1))) +
+     scale_x_continuous(breaks=c(1.75, 2.25, 2.75, 3.25, 3.75, 4.25, 4.75,
+                                 5.25, 5.75)) +
+     theme(axis.text.x = element_text(size = 10)) +
+     theme(axis.text.y = element_text(size = 10))
```

Anexo 2 Código de R para obtener la gráfica de la Figura 1.4.

Se carga el paquete `ggplot2`.

```
> library(ggplot2)
```

Se define el promedio y desviación estándar poblacional, el tamaño de muestra, el promedio muestral así como los límites de la distribución.

```
> mu = 280; sigma <- 50; n <- 36
> li <- 250; ls <- 310
```

Se determina la función de densidad normal para la distribución de la media muestral con parámetros $\mu = 280$ y $\sigma_{\bar{x}} = \frac{50}{\sqrt{36}}$.

```
> x_1 <- seq(li, ls, 0.05)
> y_1 <- dnorm(x_1, mean = mu, sd = sigma/sqrt(n))
```

Se grafica la función de densidad para la distribución de la media muestral con parámetros $\mu = 280$ y $\sigma_{\bar{x}} = \frac{50}{\sqrt{36}}$.

```
> p <- ggplot(data=data.frame(x=x_1, y=y_1), aes(x=x, y=y)) + geom_line()
```

Se define un valor específico para el promedio muestral de la distribución en estudio ($\bar{x} = 260$) y se crea una función que determine la función de densidad de la distribución de la media muestral con parámetros $\mu = 280$ y $\sigma_{\bar{x}} = \frac{50}{\sqrt{36}}$, para valores de $\bar{x} < 260$.

```
> xbarra <- 260
> func1 <- function(x) {
+     y <- dnorm(x, mean = mu, sd = sigma/sqrt(n))
+     y[x >=  xbarra] <- NA
+     return(y)
+ }
```

Se dibuja sobre la gráfica de la función de densidad de la distribución de la media muestral en estudio el área bajo la curva que representa el valor de la probabilidad que $\bar{x} < 260$.

```
> p + stat_function(fun=func1, geom="area", fill = "#84CA72", alpha=0.6)
```

Se asigna el blanco como el color de fondo de la gráfica.

```
> p + stat_function(fun=func1, geom="area", fill = "#84CA72", alpha=0.6) +
+   theme_bw()
```

Luego se elimina el fondo, las líneas de cuadrícula y los bordes de la gráfica.

```
> p + stat_function(fun=func1, geom="area", fill = "#84CA72", alpha=0.6) +
+     #tema con fondo blanco
+      theme_bw() +
+
+     #elimina el fondo, las líneas de cuadrícula y el borde
+      theme(
+        plot.background = element_blank()
+       ,panel.grid.major = element_blank()
+       ,panel.grid.minor = element_blank()
+       ,panel.border = element_blank()
+      )
```

Se agregan las líneas de los ejes de la gráfica.

```
>  p + stat_function(fun=func1, geom="area", fill="#84CA72", alpha=0.6) +
+     # tema con fondo color blanco
+       theme_bw() +
+     # elimina el fondo, las líneas de cuadrícula y el borde
+       theme(
+        plot.background=element_blank()
+       ,panel.grid.major=element_blank()
+       ,panel.grid.minor=element_blank()
+       ,panel.border=element_blank()
+      ) +
+     # dibuja las líneas de los ejes
+     theme(axis.line=element_line(color = 'black'))
```

Finalmente, se le da formato a los ejes de la gráfica.

```
> p + stat_function(fun=func1, geom="area", fill="#84CA72", alpha=0.6) +
+     # tema con fondo color blanco
+       theme_bw() +
+     # elimina el fondo, las líneas de cuadrícula y el borde
+       theme(
+        plot.background=element_blank()
+       ,panel.grid.major=element_blank()
+       ,panel.grid.minor=element_blank()
+       ,panel.border=element_blank()
+      ) +
+     # dibuja las líneas de los ejes
+     theme(axis.line=element_line(color = 'black')) +
+     # dar formato a elementos de la gráfica
+     scale_x_continuous(breaks=c(li, xbarra, mu, ls)) +
+     theme(axis.title.y=element_text(face="bold", vjust=1.5, colour="black",
+     size=rel(1))) +
+     theme(axis.title.x=element_text(face="bold", vjust=-1.5, colour="black",
+     size=rel(1))) +
+     labs(x=expression(bar(X)), y=expression(f(bar(X))))
```

Anexo 3 Código de R para obtener la Figura 1.5.

```
> library(ggplot2)
> n <- 36; mu = 280; sigma <- 50;
> li <- 250; ls <- 310
> x_1 <- seq(li, ls, 0.05)
> y_1 <- dnorm(x_1, mean = mu, sd = sigma/sqrt(n))
> linf <- 270; lsup <- 290
> p <- ggplot(data=data.frame(x = x_1, y = y_1), aes(x=x, y = y)) + geom_line()
> func1 <- function(x) {
+     y <- dnorm(x, mean = mu, sd = sigma/sqrt(n))
+     y[x <=  linf | x >= lsup] <- NA
+     return(y)}
>     p + stat_function(fun=func1, geom="area", fill = "#84CA72", alpha=0.6) +
+     # tema con fondo color blanco
+      theme_bw() +
+     # elimina el fondo, las líneas de cuadrícula y el borde
+      theme(
+        plot.background = element_blank()
+       ,panel.grid.major = element_blank()
+       ,panel.grid.minor = element_blank()
+       ,panel.border = element_blank()
+      ) +
+     # dibuja las líneas de los ejes
+      theme(axis.line = element_line(color = 'black')) +
+     # dar formato a elementos de la gráfica
+      scale_x_continuous(breaks=c(li, linf, mu, lsup, ls)) +
+      theme(axis.title.y = element_text(face="bold", vjust=1.5,
+            colour="black", size=rel(1))) +
+      theme(axis.title.x = element_text(face="bold", vjust=-1.5,
+            colour="black", size=rel(1))) +
+      labs(x = expression(bar(X)), y = expression(f(bar(X))))
```

Anexo 4 Código de R para obtener la Figura 6.1.

```
> library(ggplot2)
> alpha <- 0.02; r <- 2; c <- 3
> eje1<-matrix(c(21,64,17,16,49,14), byrow = T, ncol=3)
> chi2 <- chisq.test(eje1)$statistic
> chialpha <- qchisq(1 - alpha, (r - 1)*(c - 1))
> chi2 <- as.numeric(chi2)
> x_1 <- seq(0, 9, 0.05)
> y_1 <- dchisq(x_1, (r - 1)*(c - 1))
> ls <- chialpha
> p <- ggplot(data=data.frame(x=x_1, y=y_1), aes(x=x, y=y)) + geom_line()
> func1 <- function(x) {
+     y <- dchisq(x, (r - 1)*(c - 1))
+     y[x < ls] <- NA
+     return(y)}
>     p + stat_function(fun=func1, geom="area", fill="red", alpha=0.5) +
+     # tema con fondo color blanco
+     theme_bw() +
+     # elimina el fondo, las líneas de cuadrícula y el borde
+     theme(
+        plot.background = element_blank()
+       ,panel.grid.major = element_blank()
+       ,panel.grid.minor = element_blank()
+       ,panel.border = element_blank()
+     ) +
+     # dibuja las líneas de los ejes
+     theme(axis.line = element_line(color = 'black')) +
+     # dar formato a elementos de la gráfica
+     theme(axis.title.x = element_text(face="bold", vjust=-0.5,
+            colour="black", size=rel(1.1))) +
+     theme(axis.title.y = element_text(face="bold", vjust=1.5,
+            colour="black", size=rel(1.1))) +
+     labs(x = "x", y = "f(x)") +
+     scale_x_continuous(breaks=c(2.5, 5, round(chialpha,2), round(chi2,2))) +
+     theme(axis.text.x = element_text(size = 12)) +
+     theme(axis.text.y = element_text(size = 12))
```

Anexo 5 Código de R para obtener la Figura 6.6.

```
> imc <- c(24, 26, 27, 29, 31, 33, 35)
> ps <- c(121, 125, 133, 129, 135, 137, 130)
> ej <- data.frame(imc, ps)
> library(ggplot2)
> ggplot(ej, aes(x=imc, y=ps)) +
+   # genera círculos en la gráfica
+      geom_point(shape=1) +
+   # tema con fondo color blanco
+      theme_bw() +
+   # elimina el fondo, las líneas de cuadrícula y el borde
+      theme(
+        plot.background = element_blank()
+       ,panel.grid.major = element_blank()
+       ,panel.grid.minor = element_blank()
+       ,panel.border = element_blank()
+      ) +
+   # dibuja las líneas de los ejes
+      theme(axis.line = element_line(color = 'black')) +
+   # dar formato a elementos de la gráfica
+      theme(axis.title.x = element_text(face="bold", vjust=-0.5,
+            colour="black", size=rel(1.2))) +
+      theme(axis.title.y = element_text(face="bold", vjust=1.5,
+            colour="black", size=rel(1.2))) +
+      labs(x = "IMC",y = "Presión") +
+      theme(axis.text.x = element_text(size = 11)) +
+      theme(axis.text.y = element_text(size = 11))
```

Anexo 6 Código de R para obtener la Figura 6.8.

```
> library(ggplot2)
> imc <- c(24, 26, 27, 29, 31, 33, 35)
> ps <- c(121, 125, 133, 129, 135, 137, 130)
> reg1 <- lm(ps ~ imc)
> ggplot(reg1, aes(x=imc, y=ps)) + geom_point(shape=1) +
+   # adjuntar la línea de regresión
+     geom_smooth(method=lm) +
+   # tema con fondo color blanco
+     theme_bw() +
+   # elimina el fondo, las líneas de cuadrícula y el borde
+     theme(
+        plot.background = element_blank()
+       ,panel.grid.major = element_blank()
+       ,panel.grid.minor = element_blank()
+       ,panel.border = element_blank()
+      ) +
+   # dar formato a elementos de la gráfica
+     theme(axis.line = element_line(color = 'black')) +
+     theme(axis.title.x = element_text(face="bold", vjust=-0.5,
+           colour="black", size=rel(1.1))) +
+     theme(axis.title.y = element_text(face="bold", vjust=1.5,
+           colour="black", size=rel(1.1))) +
+     labs(x = "IMC",y = "Presión") +
+     theme(axis.text.x = element_text(size = 11)) +
+     theme(axis.text.y = element_text(size = 11))
```

La versión electrónica en formato PDF Interactivo se realizó en el Programa de Publicaciones e Impresiones de la Universidad Nacional, en el 2022.

2904-22—P.UNA