"El saber de mis hijos hará mi grandeza"

UNIVERSIDAD DE SONORA

División de Ciencias Exactas y Naturales

Programa de Licenciatura en Matemáticas

Introducción al Bootstrap en Cursos de Estadística

T E S I S

Que para obtener el título de:

Licenciado en Matemáticas

Presenta:

Adilene Calderón Gómez

Directora de tesis: Dra. Gudelia Figueroa Preciado

Hermosillo, Sonora, México. Agosto 2022.

SINODALES

Dr. José Arturo Montoya Laos
Universidad de Sonora.

Dra. Rosalía Guadalupe Hernández Amador
Universidad de Sonora.

M. C. Martha Cecilia Palafox Duarte
Universidad de Sonora.

Dra. Gudelia Figueroa Preciado
Universidad de Sonora.

Dedicatoria

A mi persona favorita, mi querida hermana, por tu inmenso amor, confianza, y motivación para alcanzar mis anhelos.
A mis padres, por su paciencia, apoyo y gran amor incondicional.
A Dios, por nunca dejarme sola y darme valentía.

Agradecimientos

Agradezco principalmente a mi directora de tesis, la Dra. Gudelia Figueroa, por su constante apoyo, paciencia, dedicación, comprensión y por su tiempo incluso en vacaciones. Gracias por las enseñanzas, el conocimientos, los consejos que me ha dado y por permitirme trabajar con usted, ha sido un verdadero honor.

A mis sinodales, el Dr. José Montoya, la Dra. Rosalía Hernandez y la M.C. Cecilia Palafox por el apoyo que me han dado en la revisión de este trabajo, por sus consejos y sus enseñanzas. Gracias por motivarme a conocer más acerca del área.

A mi hermana, por acompañarme en esta etapa, por motivarme, apoyarme en todo. Has sido el impulso que he necesitado. Gracias por siempre confiar en mi y nunca dejarme sola. Te amo mucho.

A mis padres por su constante cariño y amor incondicional que siempre me motiva a seguir mis sueños. Su apoyo a lo largo de la licenciatura y en la tesis ha sido una parte fundamental en mi desarrollo académico. Los amo.

A Cinthia, por tu confianza hacia mí, por enseñarme a que hay belleza en vivir los momentos difíciles y afrontarlos. Estoy feliz de conocerte y que seas tú quien me esté ayudando a creer más en mí.
A mi hermosa perrita Geena, por estar cada noche conmigo, con solo verte dormir a un lado de mi me brindabas paz y amor.

A mis bellas amigas de la licenciatura. Martha, gracias porque tu amistad ha sido un abrazo constante, por todo tu amor conmigo y tu apoyo en cada parte de mi vida. A Ximena y María Fernanda por hacer mas alegres mis días en la universidad. Gracias por todas sus enseñanzas.
A mis amigos de matemáticas, Ivan, Germán, Juan Carlos, Carol y Gaby; por las tardes de estudio, por las muchas veces en la biblioteca y por todo lo que me enseñaron. Hicieron esta etapa de mi vida más bonita.

A mis mejores amigos que me han apoyado desde el primer día. Alondra, Juan Pablo y Génesis, simplemente gracias por todo su cariño y amor. Los quiero mucho, gracias por siempre estar en cada etapa de mi vida y apoyarme grandemente.

A mi mejor amiga Edith, porque tu amistad llegó a mi vida en el momento mas oportuno. Gracias por estar cada día conmigo. Por tu apoyo en todo y tu gran cariño siempre.
A mis grandes amigos, Nan, José Pablo, Carlos y Paola. Gracias por hacer mi vida mas divertida y placentera, por siempre motivarme y ser de gran inspiración.

Contenido

Introducción

Los cursos de estadística que se imparten a nivel licenciatura contienen, en la mayoría de los casos, aspectos que involucran el manejo descriptivo de datos, estimación puntual de parámetros, así como estimación por medio de intervalos de confianza y pruebas de hipótesis. Cuando se abordan los diferentes temas que integran la parte inferencial de estos cursos, se enfatiza que el uso de los intervalos de confianza y pruebas de hipótesis que se presentan requieren ciertos supuestos distribucionales para la variable en estudio. Aunque la teoría que sustenta la construcción de intervalos de confianza y las pruebas de hipótesis generalmente no se alcanza a cubrir en muchos de estos cursos, se explica la necesidad de siempre verificar dichos supuestos para poder utilizar las herramientas vistas en clase.

Dada la gran cantidad de material que generalmente integra los cursos de estadística y principalmente la forma en que éste se aborda, el tiempo no permite que pueda cubrirse lo relativo a remuestreo, procedimiento que tiene un enorme potencial en educación estadística. La facilidad de contar hoy en día con herramientas computacionales que se encuentran al alcance de todo estudiante, permite incluir temas como bootstrap, procedimiento por medio del cual se genera una gran cantidad de muestras simuladas de manera paramétrica o no paramétrica, y permite calcular intervalos de confianza y pruebas de hipótesis, para diversos estadísticos de interés.

En esta tesis se espera proporcionar ideas básicas con respecto a bootstrap, procedimiento que puede por ejemplo ser utilizado para estimar el error estándar y sesgo de un estimador, obtener intervalos de confianza o realizar pruebas de hipótesis, entre otras cosas. El enfoque que se presenta es relativamente simple y brinda un gran aporte pedagógico en cursos introductorios de estadística [5].

En el Capítulo 1 se presentan algunos conceptos estadísticos que sería muy conveniente se cubrieran en los cursos de estadística que se imparten, por ejemplo, en una licenciatura en matemáticas, ya que permitiría que los estudiantes comprendan la necesidad de cumplir con ciertos supuestos que requieren la construcción de intervalos de confianza y su interpretación frecuentista, la importancia del planteamiento de las hipótesis nula y alternativa, la interpretación de un p-valor, etcétera. De esta manera, con base en las características de la muestra bajo estudio, el estudiante tendría las herramientas necesarias para seleccionar adecuadamente la inferencia estadística a realizar y comprender las limitaciones de los temas vistos en clase. En el Capítulo 2 se presenta una introducción al bootstrap, tanto paramétrico como no paramétrico; mostrando primeramente la estimación del error estándar de un estimador, mediante ambos procedimientos y después la construcción de intervalos bootstrap.

Se inicia presentando los intervalos bootstrap de percentiles y enseguida se abordan los intervalos bootstrap$-t$, que se utilizan para estimar parámetros de localización. Se muestra también cómo podrían realizarse pruebas de hipótesis utilizando bootstrap$-t$. El Capítulo 3 está dedicado a presentar, primeramente, la construcción de intervalos de confianza a partir de una cantidad pivotal, tomando el caso más sencillo de intervalo para la media de una población distribuida normalmente, y después calculando un intervalo bootstrap para este parámetro. Se aborda después la realización de una prueba de hipótesis para una media, bajo el enfoque de Neyman-Pearson, y luego se efectúa esta prueba utilizando bootstrap. El caso de construir intervalos de confianza o realizar prueba de hipótesis para comparar las medias de dos poblaciones normales también se incluye en este capítulo, que finaliza con la construcción de un intervalo de confianza para el parámetro en una distribución exponencial, situación que se explora tanto analíticamente como por medio de bootstrap; ello con la finalidad de mostrar una situación que no podría resolverse con las herramientas que comúnmente se imparten en clase. Finalmente, en las Conclusiones se incluyen los puntos principales que resumen la experiencia del trabajo desarrollado, así como bondades y limitaciones del material aquí presentado.

Capítulo 1

Algunos conceptos estadísticos

En este capítulo se presenta diverso material teórico que actualmente no está dentro del contenido de los temarios de los cursos de estadística, incluso en área de ciencias exactas; material que por lo general sólo se enuncia en estos cursos, remitiendo al estudiante a bibliografía específica. En el caso de la Licenciatura en Matemáticas de la Universidad de Sonora, la diversidad de temas que incluye un curso básico de estadística no permite abordar, en forma adecuada, el material que aquí se presenta. Por otra parte, en cursos de estadística que el Departamento de Matemáticas imparte en las áreas de servicio no sería posible abordar estos temas, ya que requieren haber cubierto cierto material en cursos previos, como por ejemplo función generadora de momentos, teorema del límite central, distribución de sumas de variables aleatorias, etcétera.

1.1. Momentos y función generadora de momentos

Algunos conceptos que resultan necesarios en los cursos de estadística son los relativos a momentos y función generadora de momentos(fgm), ya que a través de éstos puede conocerse la distribución de algún estadístico en cuestión. Es particularmente importante conocer la fgm de una variable aleatoria $X \sim N(\mu, \sigma^2)$, como se mostrará en esta sección.

Las siguientes tres definiciones relativas a los conceptos de momento, momento central y función generadora de momentos fueron tomadas de [9, p. 73, 78].

Definición 1.1 (Momento) *Si X es una variable aleatoria y su esperanza existe, el n-ésimo momento de X, usualmente denotado por μ'_n, se define como*

$$\mu'_n = E[X^n]. \tag{1.1}$$

Puede observarse que $\mu'_1 = E[X]$ corresponde a la media de X.

Definición 1.2 (Momento central) *Si X es una variable aleatoria, el n-ésimo momento central de X, con respecto a, se define como $E[(X-a)^n]$.*

Si $a = \mu_X$, el n-ésimo momento central de X con respecto a μ_X, denotado generalmente por μ_n, es

$$\mu_n = E[(X - \mu_X)^n]. \tag{1.2}$$

Definición 1.3 (Función generadora de momentos) *Sea X una variable aleatoria con densidad $f_X(\cdot)$. El valor esperado de e^{tX} se define como la función generadora de momentos de X (fgm), si el valor esperado existe para cada valor de t en algún intervalo $-h < t < h$; $h > 0$.*
La función generadora de momentos, denotada por $M_X(t)$ o $M(t)$ es

$$M(t) = E[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx, \tag{1.3}$$

si la variable aleatoria X es continua, y

$$M(t) = E[e^{tX}] = \sum_{x} e^{tx} f_X(x), \tag{1.4}$$

en el caso que la variable aleatoria sea discreta.

A continuación se presenta el cálculo de la función generadora de momentos de una variable aleatoria que sigue una distribución normal, que aunque generalmente no se cubre en los cursos de estadística, su conocimiento facilita la presentación de diversos temas del curso.

Ejemplo 1.1 *Sea X una variable aleatoria distribuida normalmente con media μ y varianza σ^2, lo cual denotaremos como $X \sim N(\mu, \sigma^2)$. La función de densidad de X está dada por $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, $-\infty < x < \infty$. Calcular su función generadora de momentos.*

$$\begin{aligned} M_X(t) = E(e^{tX}) &= \int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx = \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2} + tx}\, dx. \end{aligned}$$

A continuación se trabaja el término de la exponencial

$$\begin{aligned} &-\frac{(x-\mu)^2 - 2\sigma^2 tx}{2\sigma^2} = -\frac{x^2 - 2x(\mu + \sigma^2 t) + \mu^2}{2\sigma^2} \\ &= -\frac{x^2 - 2x(\mu + \sigma^2 t) + \mu^2 + (\mu + \sigma^2 t)^2 - (\mu + \sigma^2 t)^2}{2\sigma^2} \\ &= -\frac{(x - (\mu + \sigma^2 t))^2}{2\sigma^2} - \frac{\mu^2 - (\mu + \sigma^2 t)^2}{2\sigma^2}. \end{aligned}$$

$$
\begin{aligned}
M_X(t) &= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-(\mu+\sigma^2 t))^2}{2\sigma^2} - \frac{\mu^2-(\mu+\sigma^2 t)^2}{2\sigma^2}}\, dx \\
&= e^{-\frac{\mu^2-(\mu+\sigma^2 t)^2}{2\sigma^2}} \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-(\mu+\sigma^2 t))^2}{2\sigma^2}}\, dx \\
&= e^{-\frac{\mu^2-(\mu+\sigma^2 t)^2}{2\sigma^2}} \\
&= e^{-\frac{\mu^2-(\mu^2+2\mu\sigma^2 t+\sigma^4 t^2)}{2\sigma^2}}.
\end{aligned}
$$

Con lo cual se obtiene que

$$M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}; \quad \forall t \in \mathbb{R}. \tag{1.5}$$

Lo anterior puede ser consultado en [9, P. 109]. Es importante recordar esta expresión que resulta para la función generadora de momentos de una variable aleatoria que sigue una normal con media μ y varianza σ^2 ya que es muy sencillo inferir otras situaciones, como la siguiente.

Observación 1.1 *Si Z es una v.a. que tiene una distribución normal con media $\mu = 0$ y varianza $\sigma^2 = 1$, es decir $Z \sim N(0,1)$, su función generadora de momentos es*

$$M_Z(t) = e^{\frac{t^2}{2}}. \tag{1.6}$$

Es importante recordar esta expresión pues permitirá identificar fácilmente la distribución de ciertas variables aleatorias.

1.2. Distribución muestral de $\bar{X}$, población normal y σ^2 conocida.

En los cursos de estadística es de suma importancia conocer cómo se distribuyen ciertos estadísticos, particularmente el caso de la media muestral $\bar{X}$. Por simplicidad, es muy recurrido suponer que la variable aleatoria X bajo análisis sigue una distribución normal con varianza conocida; por ello, ésta será la primer situación que se analizará, para lo cual resulta necesario presentar primeramente algunos resultados.

Teorema 1.2 *Sean X y Y variables aleatorias independientes.*

- *Para cada $A \subset \mathbb{R}$ y $B \subset \mathbb{R}$, $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$, es decir, los eventos $\{X \in A\}$ y $\{Y \in B\}$ son eventos independientes.*
- *Sea $g(x)$ una función solo de x y $h(y)$ una función solo de y. Entonces*

$$E(g(X)h(Y)) = Eg(X)Eh(Y). \tag{1.7}$$

Teorema 1.3 *Para cada suma de variables aleatorias independientes $Y = X_1 + X_2$,*

$$M_Y(t) = E(e^{tY}) = M_{X_1}(t)M_{X_2}(t). \tag{1.8}$$

Demostración. Aplicamos el teorema anterior con $g(X_1) = e^{tX_1}$ y $h(X_2) = e^{tX_2}$ para obtiene:

$$M_Y(t) = E(e^{tY}) = E(e^{t(X_1+X_2)}) = E(e^{tX_1} \cdot e^{tX_2}) = E(e^{tX_1})E(e^{tX_2}).$$

Con lo cual queda demostrado el teorema anterior.

■

Teorema 1.4 *Si X tiene función generadora de momentos $M_X(t)$, entonces*

$$E(X^n) = M_X^{(n)}(0),$$

donde

$$M_X^{(n)}(0) = \frac{d^n}{dt^n} M_X(t)\bigg|_{t=0}, \tag{1.9}$$

es decir, el n-ésimo momento es igual a la n-ésima derivada de $M_X(t)$ evaluada en $t = 0$.

Demostración. Asumiendo que podemos diferenciar bajo el signo integral, tenemos

$$\begin{aligned}
\frac{d}{dt} M_X(t) &= \frac{d}{dt} \int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx \\
&= \int_{-\infty}^{\infty} \left(\frac{d}{dt} e^{tx} \right) f_X(x)\, dx \\
&= \int_{-\infty}^{\infty} (x e^{tx}) f_X(x)\, dx \\
&= E(X e^{tX}).
\end{aligned}$$

Así

$$\frac{d}{dt} M_X(t)\bigg|_{t=0} = E(X e^{tX})|_{t=0} = E(X).$$

De manera análoga, podemos establecer que

$$\frac{d^n}{dt^n} M_X(t)\bigg|_{t=0} = E(X^n e^{tX})|_{t=0} = E(X^n).$$

■

Teorema 1.5 *Sean $X_1, \ldots, X_n$ una muestra aleatoria de una población con función generadora de momento $M_X(t)$. Entonces la función generadora de momentos de la media muestral $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ es*

$$M_{\bar{X}_n}(t) = [M_X(t/n)]^n = (E[e^{tX/n}])^n. \tag{1.10}$$

Teorema 1.6 *Sean $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una población con distribución normal $N(\mu, \sigma^2)$, entonces la media muestral*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \tag{1.11}$$

se distribuye normalmente con media μ y varianza $\frac{\sigma^2}{n}$, esto es, $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.

Demostración. La media muestral, que está dada por

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + X_2 + \ldots + X_n}{n},$$

es combinación lineal de variables aleatorias independientes; luego

$$\begin{aligned} M_{\bar{X}_n}(t) &= E(e^{t\bar{X}_n}) = E(e^{\frac{t}{n}\sum_{i=1}^{n} X_i}) \\ &= \prod_{i=1}^{n} E(e^{tX_i/n}) = [M_X(t/n)]^n = [e^{(\mu t/n)+\sigma^2 t^2/(2n^2)}]^n \\ &= e^{\mu t + \frac{\sigma^2 t^2}{2n}}. \end{aligned}$$

La ecuación anterior resulta la función generadora de momentos de la media muestral, y de acuerdo con (1.5) la media muestral está distribuida normalmente con media μ y varianza σ^2/n.

■

Ahora, podemos utilizar la función generadora de momentos obtenida anteriormente para mostrar que $E(\bar{X}) = \mu$ y que $V(\bar{X}) = \frac{\sigma^2}{n}$.

$$\begin{aligned} M'_{\bar{X}}(t) &= \frac{d}{dt} e^{\mu t + \frac{t^2}{2}\left(\frac{\sigma^2}{n}\right)} = \left(\frac{\sigma^2 t}{n} + \mu\right) e^{\mu t + \frac{t^2}{2}\left(\frac{\sigma^2}{n}\right)}. \\ M'_{\bar{X}}(0) &= \mu. \\ M''_{\bar{X}}(t) &= \frac{d}{dt}\left[\left(\frac{\sigma^2 t}{n} + \mu\right) e^{\mu t + \frac{t^2}{2}\left(\frac{\sigma^2}{n}\right)}\right] \\ &= \left(\frac{\sigma^2}{n}\right) e^{\frac{\sigma^2 t^2}{2n} + \mu t} + \left(\frac{\sigma^2 t}{n} + \mu\right)\ e^{\frac{\sigma^2 t^2}{2n} + \mu t}\left(\frac{\sigma^2 t}{n} + \mu\right) \\ &= \left(\frac{\sigma^2 t}{n} + \mu\right)^2 e^{\frac{\sigma^2 t^2}{2n} + \mu t} + \frac{\sigma^2}{n} e^{\frac{\sigma^2 t^2}{2n} + \mu t}. \\ M''_{\bar{X}}(0) &= \mu^2 + \frac{\sigma^2}{n}. \\ Var(\bar{X}) &= E(\bar{X}^2) - [E(\bar{X})]^2 = \mu^2 + \frac{\sigma^2}{n} - \mu^2 = \frac{\sigma^2}{n}. \end{aligned}$$

Por lo tanto, a partir de la función generadora de momentos de la media muestral es posible conocer que su media es μ y varianza es $\frac{\sigma^2}{n}$.

1.3. Distribución muestral de $\frac{(n-1)S^2}{\sigma^2}$ en una población normal

En esta sección es de interés mostrar que la cantidad pivotal $\frac{(n-1)S^2}{\sigma^2}$ se distribuye como una ji-cuadrada con $n-1$ grados de libertad, pues esta cantidad permitirá construir intervalos de confianza para la varianza σ^2 en poblaciones distribuidas normalmente, al igual que realizar pruebas de hipótesis para este parámetro. Para conocer la distribución del estadístico $\frac{(n-1)S^2}{\sigma^2}$ es de utilidad conocer la función de probabilidad de una variable aleatoria que se distribuye como una gamma, y la función generadora de momentos de esta variable aleatoria, ya que la distribución ji-cuadrada es un caso especial de la distribución gamma.

Una variable aleatoria X tiene una distribución gamma con parámetros $\alpha > 0, \beta > 0$, $\Gamma(\alpha) > 0$ si su función de densidad esta dada por

$$f(x) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} & 0 < x < \infty \\ 0 & \text{en otro caso} \end{cases} \tag{1.12}$$

con

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

En Hogg & Craig [6] puede consultarse el desarrollo para obtener su función generadora de momentos, la cual está dada por

$$M_X = \frac{1}{(1-\beta t)^\alpha}, \quad t < \frac{1}{\beta}. \tag{1.13}$$

Para obtener sus momentos se calculan las siguientes derivadas:

$$M'(t) = (-\alpha)(1-\beta t)^{-\alpha-1}(-\beta),$$
$$M''(t) = (-\alpha)(-\alpha-1)(1-\beta t)^{-\alpha-2}(-\beta)^2.$$

Por lo que su media y su varianza resultan:

$$\mu = M'(0) = \alpha\beta, \tag{1.14}$$
$$\sigma^2 = M''(0) - \mu^2 = \alpha(\alpha+1)\beta^2 - \alpha^2\beta^2 = \alpha\beta^2. \tag{1.15}$$

(véase en [6, P. 151]).

Teorema 1.7 *Sea $X_1, \ldots, X_n$ variables aleatorias independientes. Supóngase para $i = 1, \ldots, n$, que X_i tiene una distribución $\Gamma(\alpha_i, \beta)$. Sea $Y = \sum_{i=1}^n X_i$. Entonces Y tiene distribución $\Gamma(\sum_{i=1}^n \alpha_i, \beta)$. (Ver en [6, P. 154].*

Demostración. Suponiendo independencia y usando la función generadora de momentos de una distribución gamma, tenemos para $t < 1/\beta$,

$$M_Y(t) = E\left[e^{t\sum_{i=1}^n X_i}\right] = \prod_{i=1}^n E\left[e^{tX_i}\right]$$
$$= \prod_{i=1}^n (1-\beta t)^{-\alpha_i} = (1-\beta t)^{-\sum_{i=1}^n \alpha_i},$$

el cual es la función generadora de momento de una distribución $\Gamma(\sum_{i=1}^{n}\alpha_i, \beta)$.

■

Consideremos ahora un caso especial de la distribución gamma, donde sus parámetros son $\alpha = n/2$, con n un entero positivo, y $\beta = 2$. Una variable aleatoria con tales características se dice tiene distribución ji-cuadrada con n grados de libertad, con función de densidad dada por (véase [9, P.241-242]):

$$f(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{\frac{n}{2}-1} e^{\frac{-x}{2}}. \tag{1.16}$$

Con base en (1.13) se puede ver que su función generadora de momentos resulta:

$$M_X = \frac{1}{(1-2t)^{n/2}}, \quad t < \frac{1}{2}. \tag{1.17}$$

Por otra parte, sus momentos son:

$$\begin{aligned} M'(t) &= n, \\ M''(t) &= n^2 + 2n, \end{aligned}$$

y su media y su varianza resultan:

$$\begin{aligned} \mu &= M'(0) = n, \\ \sigma^2 &= M''(0) - n^2 = n^2 + 2n - n^2 = 2n. \end{aligned}$$

Ahora, en caso de no presentar durante clase que una variable con distribución ji-cuadrada es un caso especial de la gamma y obtener fácilmente su función generadora de momentos, no es difícil llegar a la fgm de una variable aleatoria con distribución ji-cuadrada, como se muestra a continuación.

$$\begin{aligned} M_X(t) = E\left(e^{tX}\right) &= \int_0^\infty \frac{1}{2^{n/2}\Gamma(n/2)} e^{tx} x^{\frac{n}{2}-1} e^{\frac{-x}{2}}\, dx \\ &= \frac{1}{2^{n/2}\Gamma(n/2)} \int_0^\infty e^{tx} x^{\frac{n}{2}-1} e^{\frac{-x}{2}}\, dx \\ &= \frac{1}{2^{n/2}\Gamma(n/2)} \int_0^\infty x^{\frac{n}{2}-1} e^{(t-\frac{1}{2})x}\, dx. \end{aligned}$$

Para el caso donde $t < \frac{1}{2}$, usamos el cambio de variable $u = (\frac{1}{2} - t)x$, tenemos:

$$\begin{aligned} M_X(t) &= \left(\frac{1}{2} - t\right)^{\frac{-n}{2}} \cdot \frac{1}{2^{n/2}\Gamma(n/2)} \int_0^\infty u^{\frac{n}{2}-1} e^{(-r)}\, dr \\ &= (1-2t)^{\frac{-n}{2}} \cdot \frac{1}{\Gamma(n/2)} \int_0^\infty u^{\frac{n}{2}-1} e^{(-r)}\, dr \\ &= \frac{1}{(1-2t)^{n/2}}. \end{aligned}$$

Es de gran utilidad presentar el siguiente resultado, pues se analiza la distribución de una suma de variables aleatorias independientes. En este caso particular, las variables aleatorias se distribuyen normal estándar y se suma el cuadrado de éstas, resultando una distribución ji-cuadrada.

Teorema 1.8 *Sean $Z_1, Z_2, \ldots, Z_n$ con n entero positivo, variables aleatorias independientes que se distribuyen normal estándar, entonces la variable aleatoria*

$$\chi^2(n) = \sum_{i=1}^{n} Z_i^2, \tag{1.18}$$

sigue una distribución ji-cuadrada con n grados de libertad.

La demostración que a continuación se presenta, puede ser consultada en [9, P.242].

Demostración.

$$\begin{aligned} M(t) &= E\left(e^{t\sum Z_i^2}\right) \\ &= E\left(\prod_{i=1}^{n} e^{tZ_i^2}\right) = \prod_{i=1}^{n} E(e^{tZ_i^2}), \end{aligned}$$

donde

$$\begin{aligned} E\left(e^{tz_i^2}\right) &= \int_{-\infty}^{\infty} e^{tz^2}\left(\frac{1}{\sqrt{2\pi}}\right) e^{-\frac{1}{2}z^2}\, dz \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(1-2t)z^2}\, dz \\ &= \frac{1}{\sqrt{1-2t}} \int_{-\infty}^{\infty} \frac{\sqrt{1-2t}}{\sqrt{2\pi}} e^{-\frac{1}{2}(1-2t)z^2}\, dz \\ &= \frac{1}{\sqrt{1-2t}}, \quad para \quad t < \frac{1}{2}, \end{aligned}$$

pues lo contenido en la integral es igual a uno, ya que representa el área bajo la curva de una variable aleatoria que se distribuye normalmente con varianza $\frac{1}{1-2t}$. Por lo tanto

$$\prod_{i=1}^{n} E\left(e^{tZ_i^2}\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{1-2t}} = \left(\frac{1}{1-2t}\right)^{n/2}, \quad \text{para} \quad t < \frac{1}{2},$$

de donde se obtiene que esta función generadora de momentos es la correspondiente a la de una distribución ji-cuadrada con n grados de libertad.

■

Corolario 1 *Si $Z \sim N(0,1)$, Z^2 tiene una distribución ji-cuadrada con un grado de libertad.*

Demostración.

La demostración es inmediata usando el teorema anterior.

■

Por otra parte, la distribución del estadístico $\frac{(n-1)S^2}{\sigma^2}$ es ahora sencillo obtenerlo, de la siguente manera.

Teorema 1.9 *Si* $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ *entonces*

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1), \tag{1.19}$$

sigue una distribución ji-cuadrada con $n-1$ *grados de libertad.*

Demostración.

Definimos $U = \frac{(n-1)S^2}{\sigma^2}$ y $T = \left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2$ variables independientes, tenemos

$$\begin{aligned}(n-1)S^2 &= \sum_i (X_i - \bar{X})^2 = \sum_i (X_i - \mu)^2 - n(\bar{X} - \mu)^2, \\ \frac{(n-1)S^2}{\sigma^2} &= \sum \left(\frac{X_i - \mu}{\sigma}\right)^2 - \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2,\end{aligned}$$

con $U = V - T$, $V = \sum \left(\frac{X_i-\mu}{\sigma}\right)^2$, y donde U y T independientes.
La fgm M_{U+T} y M_V resulta la misma y además

$$M_{U+T} = M_U M_T = M_V,$$

donde V y T siguen una distribución χ^2, con n y 1 grados de libertad, respectivamente. Así

$$M_U = \frac{M_V}{M_T} = \frac{(1-2t)^{-n/2}}{(1-2t)^{1/2}} = (1-2t)^{-(n-1)/2},$$

resulta ser la fgm de la variable de $\frac{(n-1)S^2}{\sigma^2}$, la cual coincide con la función generadora de momentos de una variable aleatoria ji-cuadrada con $n-1$ grados de libertad.

■

1.4. Distribución muestral de $\bar{X}$, población normal y σ^2 desconocida

En de importancia conocer cómo se distribuye la media muestral cuando se trabaja con variables aleatorias que se distribuyen normalmente con varianza σ^2 desconocida, pues en la práctica es mucho más común que se presente una situación de este tipo.

Definición 1.4 (Distribución t-Student) *Si* X *es una variable aleatoria con función de densidad dada por*

$$f(x) = \frac{\Gamma[(n+1)/2]}{\Gamma(n/2)} \frac{1}{\sqrt{n\pi}} \frac{1}{(1 + x^2/n)^{(n+1)/2}},$$

se dice que X *tiene una distribución* t*-Student con* n *grados de libertad.*

Cuando una variable aleatoria X tiene una función generadora de momentos, entonces el $k-$ésimo momento de X existe, y es finito, para cualquier $k \in \mathbb{N}$. Dado que en esta distribución el $k-$ésimo momento sólo está bien definido si $k < n$, se dice que la fgm de esta distribución no existe. Como alternativa, en los cursos de estadística generalmente se utiliza el resultado que se muestra a continuación, con la finalidad de estudiar la distribución muestral de la media, en el caso de población normal con varianza desconocida.

Teorema 1.10 *Si $Z \sim N(0,1)$, si $U \sim \chi^2_n$, con Z y U independientes, entonces $Z/\sqrt{U/n}$ tiene una distribución $t-$Student con n grados de libertad.*

La demostración de este teorema puede consultarse en [9, P. 150]. Aunque este material está fuera del alcance de los temas de un curso de estadística, es de importancia presentarlo pues facilita la comprensión del siguiente corolario, que permitirá resolver una diversidad de problemas que si se abordan en los cursos de estadística.

Corolario 2 *Si $X_1, \dots X_n$ es una muestra aleatoria de una distribución normal con media μ y varianza σ^2, $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ y $U = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2} \sim \chi^2_{(n-1)}$, con Z y U independientes, entonces,*

$$\frac{(\bar{X}-\mu)/(\sigma/\sqrt{n})}{\sqrt{(1/\sigma^2)\sum(X_i-\bar{X})^2/(n-1)}} = \frac{(\bar{X}-\mu)\sqrt{n}}{S} = \frac{\bar{X}-\mu}{S/\sqrt{n}},$$

tiene una distribución t-Student con n-1 grados de libertad, donde $S^2 = \frac{1}{n-1}\sum(X_i-\bar{X})^2$.

En la Figura 1.1 se presenta la densidad que resulta, al variar los grados de libertad, de una variable aleatoria que sigue una distribución $t-$*Student*. En este caso se trabaja con $n = \{10, 20, 30\}$ grados de libertad. El código utilizado para realizar la Figura 1.1 se encuentra en el Apéndice B.1. Puede observarse que conforme aumentan los grados de libertad, la densidad tiende a parecerse a la de una variable normal estándar. De hecho, si X es una variable aleatoria que se distribuye como una $t-$Student con n grados de libertad (véase [9])

$$E(X) = 0 \quad si \quad n > 1 \quad y \quad V(X) = n/n-1 \quad si \quad n > 2. \tag{1.20}$$

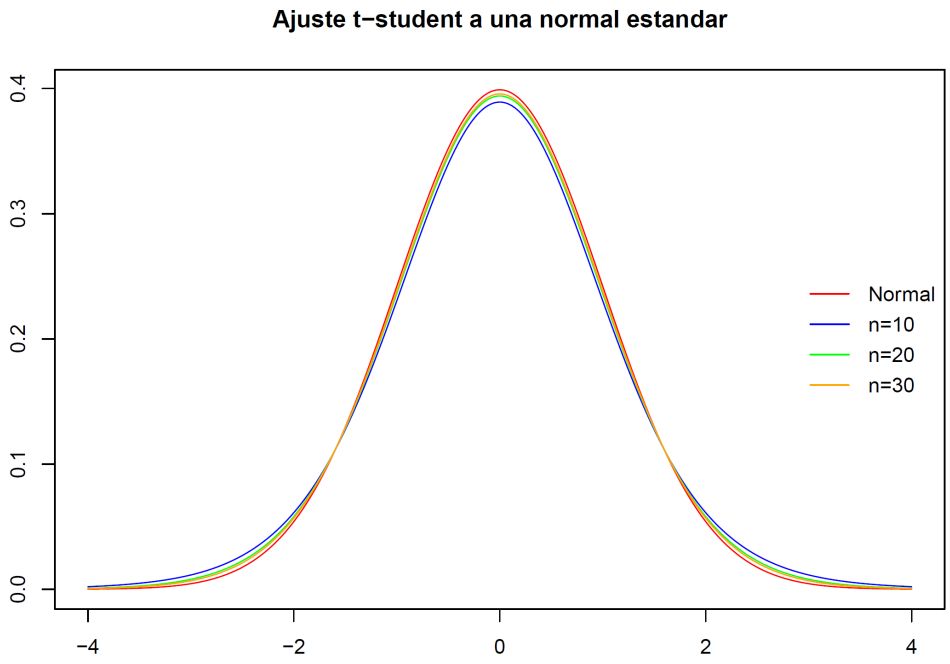

Figura 1.1: Densidades para una variable aleatoria $t - Student$ con $n = \{10, 20, 30\}$ g.l. y una variable aleatoria normal estándar.

1.5. Teorema de límite central

En esta sección se presenta un resultado muy utilizado en la resolución de problemas en los cursos de estadística: el teorema del limite central (TLC). Por lo general no se demuestra este teorema, ya que al momento que los estudiantes cursan estadística, no han visto convergencia en distribución. Sin embargo, en el Apéndice A se define convergencia en distribución y se presenta una demostración que puede ser de interés.

Teorema 1.11 (Teorema del límite central) *Sea $f(\cdot)$ una densidad con media μ y varianza finita σ^2. Sea $\bar{X}_n$ la media de una muestra aleatoria de tamaño n tomada de $f(\cdot)$. Sea la variable aleatoria Z_n definida como*

$$Z_n = \frac{\bar{X}_n - E(\bar{X}_n)}{\sqrt{Var[\bar{X}_n]}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}},$$

entonces, la distribución de Z_n se aproxima a la distribución normal estándar conforme n tiende a infinito.

Este teorema también se puede enunciar en términos de una suma de variables aleatorias independientes $S_n = \sum_{i=1}^{n} X_i$, y donde la variable Z_n sería

$$Z_n = \frac{S_n - E(S_n)}{\sqrt{Var[S_n]}} = \frac{S_n - n\mu}{\sqrt{n}\sigma},$$

Tan sólo con la finalidad de profundizar en la importancia y aplicaciones del TLC, se muestra su utilidad con el caso de la aproximación normal a la distribución binomial, que se presenta a continuación.

1.5.1. Aproximación normal a la distribución binomial

Consideremos una variable aleatoria X que se distribuye binomialmente con parámetros n y p, lo cual podemos denotar por $X \sim Bin(n,p)$, donde $X = \sum_{j=1}^{n} X_j$, y $X_1, X_2, \ldots X_n$ son variables aleatorias independientes Bernoulli con $P(X_j = 1) = p$ y $P(X_j = 0) = 1 - p$, con p la probabilidad de éxito. Su función de probabilidad está dada por

$$f(x) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k \in \{0, 1, \ldots n\},$$

donde $n \geq 1$ es el número de ensayos, $0 < p < 1$, y su función de distribución acumulada resulta

$$F_X(k) = P(X \leq k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i}.$$

Aunque en el curso de probabilidad, que se lleva previo a un curso de estadística, se demuestra que $E(X) = np$ y $V(X) = npq$, con $q = 1 - p$, no es difícil llegar a esos resultados a partir de su función generadora de momentos, la cual es

$$M_X(t) = E(e^{tX}) = (pe^t + (1-p))^n, \quad \forall t,$$

donde diferenciando con respecto a t se llega a que

$$E(X) = np(q+p)^{n-1} = np,$$

$$E(X^2) = np(q + np),$$

$$V(X) = E(X^2) - [E(X)]^2 = np(q + np) - n^2p^2 = npq.$$

De esta manera, utilizando el TLC, se puede formar la variable Z_n de la siguiente manera

$$Z_n = \frac{X - E(X)}{\sqrt{Var[X]}} = \frac{X - np}{\sqrt{np(1-p)}}.$$

Un aspecto importante que en los cursos se muestra a través de ejemplos, es qué tan buena resulta la aproximación normal a la binomial utilizando el resultado anterior y calculando la probabilidad de que la variable X se encuentre, digamos, entre dos valores a y b.

$$P(a \leq X \leq b) = P\left(\frac{a - np}{\sqrt{np(1-p)}} \leq Z \leq \frac{b - np}{\sqrt{np(1-p)}} \right). \tag{1.21}$$

Los resultados de ejemplos que se desarrollan en el curso, muestran que es necesario introducir una corrección por continuidad, la cual es un ajuste que se realiza cuando una distribución

discreta es aproximada por medio de una distribución continua. Dado que la distribución binomial es discreta y la distribución normal es continua, es común utilizar la corrección de continuidad y la más popular es la corrección de Yates definida como

$$F_X(k) \approx \Phi\left(\frac{k - np + 0.5}{\sqrt{np(1-p)}}\right), \quad k \in \{0, 1, \ldots, n\}.$$

Esto es, la Ecuación en (1.21) resulta

$$P(a \leq X \leq b) = P\left(\frac{a - 0.5 - np}{\sqrt{np(1-p)}} \leq Z \leq \frac{b + 0.5 - np}{\sqrt{np(1-p)}}\right). \tag{1.22}$$

Para mostrar la necesidad de esta corrección por continuidad, pueden utilizarse situaciones ilustrativas como la del ejemplo que a continuación se muestra, donde se observa la diferencia entre considerar o no esta corrección por continuidad. El código con el que se realizó la Figura 1.2 se encuentra en el Apéndice B.2.

Ejemplo 1.2 *Supongamos que* $X \sim Bin(10, 0.5)$*. Calculemos la* $P(X \leq 6)$*, con y sin el uso de la corrección de continuidad.*

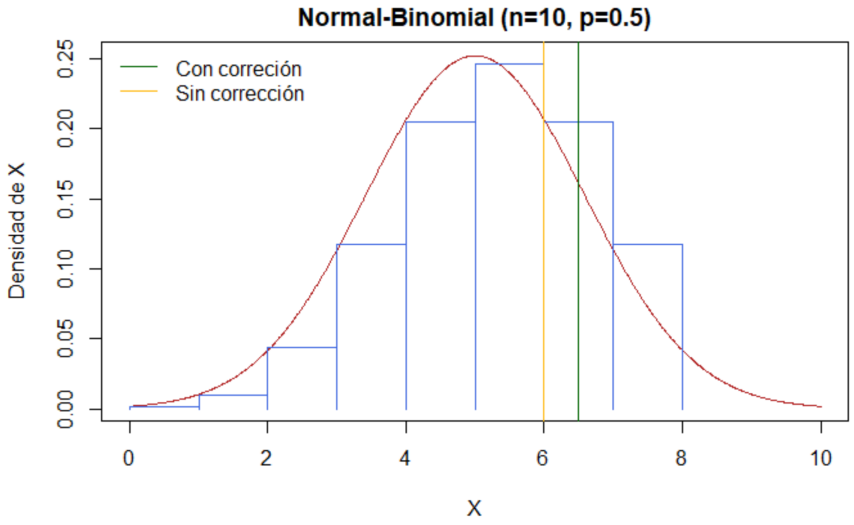

Figura 1.2: Una aproximación normal a la probabilidad $P(X \leq 6)$ con y sin corrección por continuidad, con $n = 10$, $p = 0.5$.

En la Figura 1.2 puede observarse que el área comprendida entre la línea amarilla y la línea verde, no es considerada cuando no se utiliza la corrección por continuidad.

El uso de software permite mostrar el comportamiento de la variable Z_n, para el caso de trabajar con una variable aleatoria X binomial, cuando se consideran diferentes tamaños de muestra. Esto puede apreciarse en la Figura 1.3, donde se construyen gráficas cuantil-cuantil, cuya construcción se explica en el Apéndice A.2.1.

Ejemplo 1.3 *Construcción de gráficas cuantil-cuantil considerando valores de* $n = 10, n = 50$ *y* $n = 100; p = 0.1, p = 0.3, p = 0.5$, *en una distribución binomial con parámetros* n *y* p.

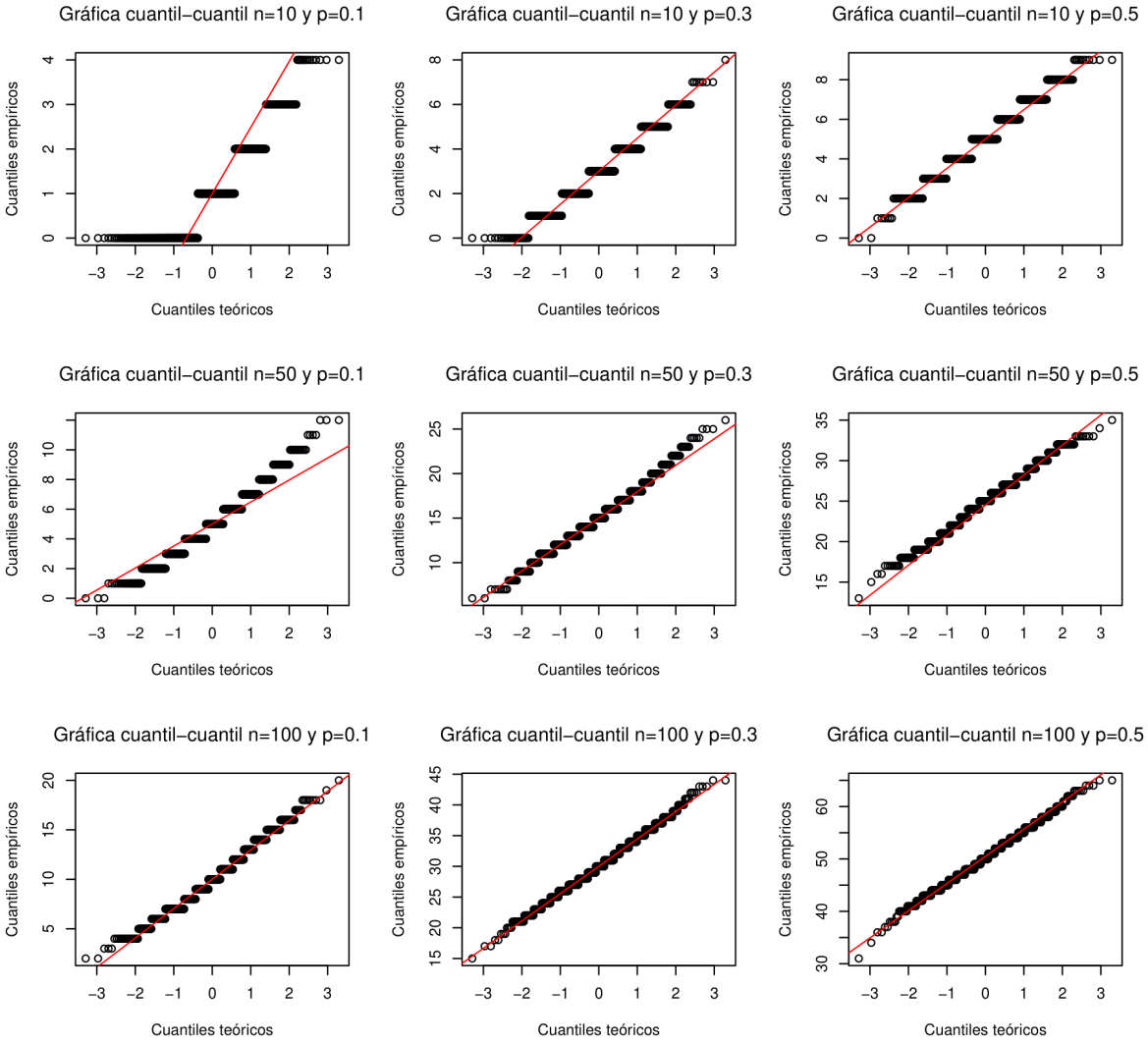

Figura 1.3: Comparación de cuantiles binomiales y normales.

Las gráficas que se ilustran en la Figura 1.3, cuyo código se encuentra en el Apéndice B.3, muestra que cuando n es pequeña y alejada de $p = 0.5$, la aproximación normal a la binomial no será muy adecuada, sin embargo, conforme n aumenta, aún con valores alejados de $p = 0.5$, la aproximación es buena.

1.6. Intervalos de confianza

Un tema de gran importancia en los cursos de estadística es el relativo a intervalos de confianza. Por lo general, la construcción de intervalos de confianza en los cursos de estadística se basa

en el uso de cantidades pivotales que usualmente se presentaron con anterioridad, mismas que contienen un estimador, generalmente insesgado, el cual permite estimar algún parámetro de un modelo, con base en los valores observados de ciertas variables aleatorias. A continuación se presentan algunos conceptos necesarios en la construcción de intervalos de confianza.

Definición 1.5 (Cantidad pivotal) *Sea $X_1, \ldots, X_n$ una muestra aleatoria de una densidad $f(\cdot;\theta)$. Sea $Q = q(X_1, \ldots, X_n; \theta)$; esto es, sea Q una función de $X_1, \ldots, X_n$ y θ. Si Q tiene una distribución que no depende de θ, entonces se dice que Q es una cantidad pivotal.*

Cuando el interés principal consiste en estimar el centro de una distribución, podría utilizarse un estimador puntual del valor esperado de la variable aleatoria bajo estudio. Dado que por lo general se utilizan estimadores insesgados, se incluye la definición de éste.

Definición 1.6 (Estimador insesgado) *Un estimador $\widehat{\theta}$ de un parámetro θ, es insesgado si*

$$E(\widehat{\theta}) = \theta.$$

La propiedad de ser insesgado no es suficiente pues una distribución puede estar centrada en θ y estar muy dispersa, por lo que también se requiere información acerca de su dispersión, por lo cual se suele analizar la varianza o bien la desviación estándar del estimador. Una manera usual de presentar ambas cantidades utilizando

$$\text{estimación} \pm \text{desviación estándar del estimador},$$

con lo cual se obtiene una estimación por intervalo, la cual puede definirse como:

Definición 1.7 (Estimación por intervalo) *Una estimación por intervalo para un parámetro $\theta \in \mathcal{R}$, es cualquier par de funciones $L(x_1, x_2, \ldots, x_n)$ y $U(x_1, x_2, \ldots, x_n)$ de una muestra que satisface $L(X) \leq U(X)$ para toda $x \in X$. Una vez que $X = x$ es observada, el intervalo aleatorio $[L(x), U(x)]$ es llamado estimador por intervalo.*

Nótese que en la definición anterior no se habla de nivel de confianza, pues es simplemente una estimación por intervalo. La definición de intervalo de confianza requiere el concepto de coeficiente de confianza y éste, el de probabilidad de cobertura; ambos se presentan a continuación.

Definición 1.8 (Probabilidad de cobertura) *Para un estimador de intervalo $[L(X), U(X)]$ de un parámetro θ, la probabilidad de cobertura de $[L(X), U(X)]$ es la probabilidad que el intervalo aleatorio $[L(X), U(X)]$ cubra el verdadero parámetro θ; ello puede ser denotado por $P_\theta(\theta \in [L(X), U(X)])$.*

Es muy común hablar del nivel de confianza de un intervalo, pero ¿qué es ello? Cuando se habla por ejemplo, de un 95 % de confianza, significa que si tomáramos repetidamente muestras aleatorias, del mismo tamaño, de la población en estudio y calculáramos el intervalo de confianza que se obtiene a partir de cada una de estas muestras, a la larga observaríamos que aproximadamente el 95 % de estos intervalos capturan el parámetro bajo estudio. Este concepto podemos definirlo como sigue.

Definición 1.9 (Coeficiente de confianza) *Para un estimador de intervalo* $[L(X), U(X)]$ *de un parámetro* θ*, el coeficiente de confianza de* $[L(X), U(X)]$*, es el ínfimo de probabilidades de cobertura,* $\inf_\theta P_\theta(\theta \in [L(X), U(X)])$.

A continuación se presenta una definición de intervalo de confianza.

Definición 1.10 (Intervalo de confianza) *Sea* $X_1, \ldots, X_n$ *una muestra aleatoria de una densidad* $f(\cdot;\theta)$*. Sea* $T_1 = t_1(X_1, \ldots, X_n)$ *y* $T_2 = t_1(X_1, \ldots, X_n)$ *dos estadísticos que satisfacen* $T_1 \leq T_2$ *tal que* $P_\theta[T_1 < \tau(\theta) < T_2] \equiv (1-\alpha)$*, donde* $(1-\alpha)$ *no depende de* θ*. El intervalo aleatorio* (T_1, T_2) *es llamado un intervalo al* $(1-\alpha)100\,\%$ *de confianza para* $\tau(\theta)$*, y* $(1-\alpha)$ *es llamado nivel de confianza, y* T_1 *y* T_2 *son los límites de confianza inferior y superior, respectivamente, para* $\tau(\theta)$*. Los valores* (t_1, t_2) *del intervalo aleatorio* (T_1, T_2) *es llamado intervalo de confianza del* $(1-\alpha)100\,\%$ *para* $\tau(\theta)$.

En el Capítulo 3 se presentará la construcción de diversos intervalos de confianza, para uno o dos parámetros, que generalmente se cubren en los cursos básicos de estadística.

1.7. Prueba de hipótesis

En ocasiones se está interesado no sólo en una estimación puntual o bien una estimación por intervalo. Puede ser relevante el poder decidir si la muestra proporciona evidencia a favor o en contra de una cierta afirmación con respecto a un parámetro. Es entonces cuando realizamos prueba de hipótesis; es decir utilizamos la información de la muestra aleatoria para decidir a favor de cuál de dos hipótesis da evidencia dicha muestra observada.

En prueba de hipótesis se establecen dos hipótesis complementarias que se denotan como hipótesis nula H_0, e hipótesis alternativa H_1. De acuerdo con el enfoque de Neyman y Pearson, la decisión de rechazar H_0 a favor de H_1 se toma sobre la base de un estadístico $T(X)$, que se evalúa en la muestra observada $X = x$. El conjunto de valores de $T(x)$ para los cuales se rechaza H_0, se denomina región de rechazo y ésta se determina con base a lo que se conoce como nivel de significancia α. Si θ denota un parámetro de una población, el formato general de las hipótesis nula y alternativa es $H_0 : \theta \in \Theta_0$ y $H_1 : \theta \in \Theta_0^c$, donde Θ_0 es un subconjunto del espacio de parámetros y Θ_0^c es su complemento. Cuando se conduce una prueba de hipótesis H_0 contra una hipótesis alternativa H_1, al nivel de significancia α, se tiene un conjunto R de valores del estadístico de prueba para el cual se rechazará la hipótesis nula. A R se le conoce generalmente como región de rechazo o la región critica y los valores en los puntos finales son llamados los valores críticos.

Se puede incurrir en dos tipos de errores al aplicar este paradigma, los cuales se señalan en la siguiente tabla: H_0 puede ser rechazada cuando es verdadera; tal error se conoce como Error Tipo I y su probabilidad se denota por α, donde α=P(Rechazar $H_0|H_0$ es cierta), y se conoce como nivel de significancia de la prueba. Ahora, H_0 puede no rechazarse cuando es falsa; tal error se conoce como Error Tipo II y la probabilidad de cometerlo se denota por β, donde β=P(No rechazar $H_0|H_0$ es falsa). Por otra parte, la probabilidad de no rechazar la hipótesis nula cuando ésta es verdadera es $1-\alpha$, o sea, el nivel de confianza y la probabilidad de rechazar H_0 cuando ésta es falsa es $1-\beta$, y se conoce como la potencia de la prueba. Es posible que no se utilice el criterio de Neyman y Pearson de prueba de hipótesis y se

	H_0 es verdadera	H_0 es falsa
Rechazar H_0	Error tipo I (α)	Decisión correcta $(1-\beta)$
No rechazar H_0	Decisión correcta $(1-\alpha)$	Error tipo II (β)

Tabla 1.1: Decisiones posibles en una prueba de hipótesis.

realice lo que se conoce como prueba de significancia, en la cual no se plantea una hipótesis alternativa, y la decisión se basa en lo que se conoce como p-valor, el cual puede describirse como la probabilidad de observar un resultado muestral o algo más extremo, bajo el supuesto que la hipótesis nula es cierta. Con base en el resultado del $p-$valor se toma una decisión relativa a rechazar o no la hipótesis nula. Por lo general se rechaza H_0 cuando el $p-$valor es menor que un cierto umbral o nivel de significancia.

Las siguiente definiciones de $p-$valor y $p-$valor valido pueden encontrarse en [1].

Definición 1.11 ($p-$valor) *Un $p-$valor $p(\mathbf{X})$ es un estadístico de prueba que satisface $0 \leq p(\mathbf{x}) \leq 1$ para cada punto muestral $\mathbf{x}$. Valores pequeños de $p(\mathbf{X})$ dan evidencia a favor de H_1. Un $p-$valor es válido si, para cada $\theta \in \Theta_0$ y cada $0 \leq \alpha \leq 1$,*

$$P_\theta(p(\mathbf{X}) \leq \alpha) \leq \alpha. \tag{1.23}$$

Definición 1.12 ($p-$valor) *Sea $W(\mathbf{X})$ un estadístico de prueba tal que valores grandes de W proporcionan evidencia a favor de H_1. Para cada valor muestral $\mathbf{x}$, se define*

$$p(\mathbf{x}) = \sup_{\theta \in \Theta_0} P_\theta(W(\mathbf{X}) \geq W(\mathbf{x})). \tag{1.24}$$

Entonces, $p(\mathbf{x})$ es un $p-$valor válido.

En el Capítulo 3 se realizarán algunas pruebas de hipótesis tanto para uno como dos parámetros. Se plantearán hipótesis desde el enfoque de Neyman y Pearson, pero los resultados también pueden analizarse en términos de un $p-$valor.

Capítulo 2

Una introducción al bootstrap

El bootstrap es una técnica estadística que forma parte de los métodos de remuestreo, los cuales tienen un enorme potencial en educación estadística y en la práctica estadística. En el artículo Bootstrap Methods: Another Look at the Jackknife en Annals of Statistics, Brad Efron [3] propuso un procedimiento de remuestreo al que llamó bootstrap, mismo que supone que los datos con los que se trabajan son representativos de la población bajo estudio, por lo que tomar muestras aleatorias a partir de esos datos resultaría equivalente a muestrear de la población misma.

En este capítulo se presentan algunos procedimientos que utilizan bootstrap y que permiten obtener intervalos de confianza o bien realizar pruebas de hipótesis de interés. Se trabajará principalmente con los intervalos conocidos como bootstrap$-t$, tomando en cuenta ciertas investigaciones como la realizada por por Hesterberg [5], donde muestra que este procedimiento funciona mucho mejor cuando se tienen muestras pequeñas, como es el caso de los diversos ejemplos que se presentan el Capítulo 3; mientras que para muestras grandes resulta conveniente el calcular los intervalos llamados *percentile*-bootstrap, o intervalos basados en percentiles, donde simplemente se toman ciertos percentiles de la distribución bootstrap del estimador de un parámetro de interés.

La parte inferencial que se enseña en los cursos universitarios de estadística está basada en distribuciones muestrales, las cuales pueden obtenerse cuando se seleccionan todas las muestras aleatorias de la población bajo estudio, se calcula el estadístico de interés en cada una de las muestras y se construye la distribución muestral del estadístico. Sin embargo, en la práctica solamente se cuenta con una muestra. La idea de bootstrap es seleccionar muestras de un estimado de la población, en lugar de la población misma, y calcular el estadístico de interés en cada una de las muestras, con lo cual se construirá la distribución bootstrap. La idea detras de este procedimiento es lo que llaman principio *plug-in*, el cual establece que si algo es desconocido, entonces se sustituye por una estimación.

Este capítulo se enfoca principalmente en cómo construir los intervalos conocidos como bootstrap$-t$, aunque se explica brevemente otras opciones de construccion de intervalos, como el intervalo bootstrap de percentiles y el intervalo basado en la distribución $t-$Student. En las siguientes secciones se expone cómo realizar tanto un bootstrap paramétrico como uno

no-paramétrico, iniciando con el procedimiento de calcular el error estándar del estimador bajo estudio. Finalmente, se expone una forma general de cómo construir los intervalos bootstrap, que se ejemplifican en el Capítulo 3.

2.1. Bootstrap no paramétrico.

El procedimiento bootstrap más común y sencillo es el conocido como bootstrap no-paramétrico, pues los datos aleatorios son creados por medio de un remuestreo con reemplazamiento, a partir de la muestra original. Con el fin de explicar lo anterior con mayor detalle, se presenta primeramente el concepto de función de distribución empírica que en lo sucesivo será de mucha utilidad.

Definición 2.1 *Función de distribución empírica. Dada una muestra aleatoria* $(X_1, X_2, \ldots, X_n)$ *de tamaño* n*, de una distribución de probabilidad* F*, la función de distribución empírica* $\hat{F}_n$ *es una distribución discreta que asigna una probabilidad de* $\frac{1}{n}$ *a cada valor de* X_i*,* $i = 1, 2, \ldots, n$*.*

En otras palabras, $\hat{F}_n$ asigna a un conjunto A en el espacio muestral de X su probabilidad empírica

$$\widehat{P}\{A\} = \#\{x_i \in A\}/n,$$

la proporción de la muestra observada $(x_1, x_2, \ldots, x_n)$ que ocurre en A.

Con la finalidad de ilustrar cómo se realiza el remuestreo en un bootstrap no-paramétrico, Efron y Tibshirani, en su libro [4, p. 10] utilizan un ejemplo que a continuación se retoma. Los datos son relativos a un tratamiento destinado a prolongar la supervivencia después de una cirugía de prueba. Se seleccionaron 16 ratones para el pequeño experimento, en el que se tomaron aleatoriamente 7 de ellos para recibir un nuevo tratamiento médico, mientras que los 9 restantes fueron asignados al grupo control (sin tratamiento). Indiquemos por $x_1, x_2, \ldots, x_7$ el tiempo de vida, en días, del grupo con tratamiento; donde $x_1 = 94, x_2 = 197, x_3 = 16, x_4 = 38, x_5 = 99, x_6 = 141$ y $x_7 = 23$. Igualmente sean $y_1, y_2, \ldots, y_9$, el tiempo de vida del grupo de control, con valores resultantes que se muestran en la Tabla 2.1. Básicamente la idea es ver si el tratamiento prolongó la supervivencia. Una comparación de las medias muestrales proporciona una idea al respecto.

Grupo	Datos	Tamaño de muestra	Media
Tratamiento	$x_1 = 94, x_2 = 197, x_3 = 16,$ $x_4 = 38, x_5 = 99, x_6 = 141,$ $x_7 = 23$	7	$\bar{x} = \sum_{i=1}^{7} x_i/7 = 86.86$
Control	$y_1 = 52, y_2 = 104, y_3 = 146,$ $y_4 = 10, y_5 = 51, y_6 = 30,$ $y_7 = 40, y_8 = 27, y_9 = 49$	9	$\bar{y} = \sum_{i=1}^{9} y_i/9 = 56.2$

Tabla 2.1: Tabla de comparación de medias muestrales.

Una muestra bootstrap, que podemos denotar por $x^* = (x_1^*, x_2^*, \ldots, x_n^*)$ se obtiene tomando muestras aleatorias de tamaño n, con reemplazamiento, de los datos originales x_i. Así, por ejemplo, $x^* = (x_1, x_4, x_7, x_1, x_2, x_5, x_7)$ es una muestra bootstrap de tamaño 7, extraída con reemplazamiento, donde, $x_1^* = x_1, x_2^* = x_4, x_3^* = x_7, x_4^* = x_1, x_5^* = x_2, x_6^* = x_5, x_7^* = x_7$. De igual manera, $y^* = (y_2, y_3, y_1, y_5, y_5, y_6, y_7, y_9, y_5)$ sería una muestra bootstrap de tamaño 9, extraída de la muestra obtenida para el grupo de control, y efectuada con reemplazamiento. Así, $y_1^* = y_2, y_2^* = y_3, y_3^* = y_1, y_4^* = y_5, y_5^* = y_5, y_6^* = y_6, y_7^* = y_7, y_8^* = y_9, y_9^* = y_5$.

Las medias de estas muestras bootstrap x_i^* y y_j^* resultan:

$$\bar{x}^* = \sum_{i=1}^{7} x_i^*/7 = 81.14 \quad y \quad \bar{y}^* = \sum_{i=1}^{9} y_i^*/9 = 63.44. \tag{2.1}$$

Para construir una distribución bootstrap efectuaríamos este procedimiento un número m de veces, que por lo general suele ser 1000, 5000 o $10,000$.

2.1.1. Estimación bootstrap del error estándar

Como puede observarse en el ejemplo presentado en la sección anterior, propuesto por Efron y Tibshirani [4, p. 10], la diferencia de $\bar{x} - \bar{y} = 86.86 - 56.2 = 30.66$, pareciera sugerir que hay un efecto considerable de prolongación de la supervivencia cuando se toma el tratamiento. Ahora, cuando se analizan las medias obtenidas mediante remuestreo $\bar{x}^* - \bar{y}^* = 81.14 - 63.44 = 17.7$, la diferencia ya no es tan grande. Sin embargo, puede observarse que hay mediciones muy extremas en ambos grupos y las medias se calcularon con base a muestras pequeñas de tan solo 7 y 9 ratones, por lo que Efron y Tibshirani [4] aprovechan estas características del problema para proponer llevar a cabo una estimación de la precisión de las medias muestrales de $\bar{x}$ y $\bar{y}$, para lo cual proponen un algoritmo bootstrap que les permite tener una aproximación al error estándar de un estimador.

Así, si $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ es una muestra aleatoria de una distribución de probabilidad desconocida F y θ es un parámetro de interés que puede estimarse mediante $\hat{\theta}$, el error estándar de $\hat{\theta}$ es tan sólo la desviación estándar del estimador y puede denotarse por $se_F(\hat{\theta})$. Luego entonces, por el principio *plug-in* el estimador bootstrap de este error estándar es

$$se_{\hat{F}}(\hat{\theta}^*), \tag{2.2}$$

donde $\hat{\theta}^* = s(\mathbf{x}^*)$, y $\mathbf{x}^* = (x_1^*, x_2^*, \ldots, x_n^*)$ es una muestra bootstrap, es decir una muestra aleatoria de tamaño n, tomada con reemplazamiento de $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ y $s(\mathbf{x}^*)$ es el resultado de aplicar una función $s(\cdot)$ a $\mathbf{x}^*$, así como fue aplicada en $\mathbf{x}$. Por ejemplo, si $s(\mathbf{x})$ es la media muestral $\bar{x}$, entonces $s(\mathbf{x}^*)$ es la media de la muestra bootstrap $\bar{x}^* = \sum_{i=1}^{n} x_i^*/n$.

Es muy común denotar el estimador bootstrap del error estándar de un estadístico por $\hat{se}_B$, donde B es el número de muestras bootstrap utilizadas. Para cada una de las muestras se calcula el valor del estadístico $s(\mathbf{x})$, obteniendo así $s(\mathbf{x}^{*1}), s(\mathbf{x}^{*2}), \ldots s(\mathbf{x}^{*B})$. La desviación

estándar de las estimaciones $s(\mathbf{x}^{*1}), s(\mathbf{x}^{*2}), \ldots s(\mathbf{x}^{*B})$ resulta la estimación bootstrap del error estándar $se_F(\hat{\theta})$.

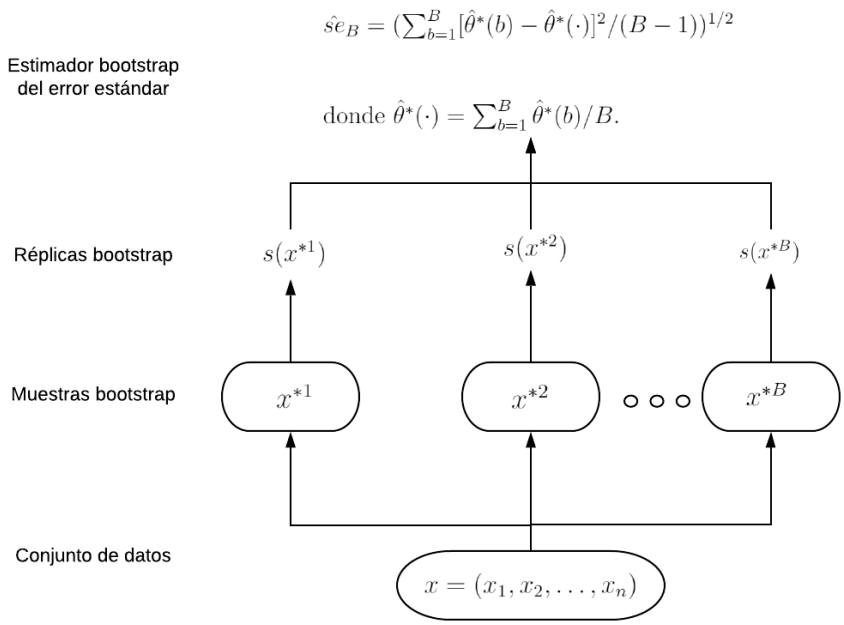

Figura 2.1: Esquema del proceso para estimar el error estándar de un estadístico $s(\mathbf{x})$.

Lo descrito a continuación se ilustra en la Figura 2.1, donde se resume el algoritmo bootstrap no-paramétrico para estimar errores estándar, que se describe más ampliamente en los siguientes puntos:

- Seleccionar B muestras bootstrap independientes $\mathbf{x}^{*1}, \mathbf{x}^{*2}, \ldots, \mathbf{x}^{*B}$. Cada una de estas muestras consiste de n valores de datos extraídos con remplazamiento de la muestra $\mathbf{x}$.
- Evaluar el estadístico de interés en cada replicación de muestra bootstrap

$$\hat{\theta}^*(b) = s(\mathbf{x}^{*b}), \quad b = 1, 2, \ldots, B. \tag{2.3}$$

- Estimar el error estándar $se_F(\hat{\theta})$ por medio de la desviación estándar muestral de las B estimaciones bootstrap del estadístico de interés.

$$\hat{se}_B = \left\{ \sum_{b=1}^{B} [\hat{\theta}^*(b) - \hat{\theta}^*(\cdot)]^2/(B-1) \right\}^{1/2}, \tag{2.4}$$

  donde $\hat{\theta}^*(\cdot) = \sum_{b=1}^{B}\hat{\theta}^*(b)/B$.

Retomando el caso en que $s(\mathbf{x})$ sea la media muestral $\bar{x}$, entonces $s(\mathbf{x}^{*b})$ es la media de la $b-$*ésima* muestra bootstrap $\bar{x}^{*b} = \sum_{i=1}^{n} x_i^{*b}/n$ y el error estándar por medio de bootstrap se

estimaría de la siguiente manera:

$$\hat{se}_B = \left\{ \sum_{b=1}^{B} [\bar{x}^{*b} - \bar{x}^*(\cdot)]^2/(B-1) \right\}^{1/2}, \tag{2.5}$$

donde $\bar{x}^*(\cdot) = \sum_{b=1}^{B} \bar{x}^{*b}/B$.

Como ejemplo al cálculo del error estándar de la media muestral, en la Tabla 2.2 se muestran diversos cálculos del error estándar, a través de un bootstrap no-paramétrico que se realiza a partir de la muestra de tamaño 7 del grupo de tratamiento, del ejemplo incluido en Efron y Tibshirani [4, p. 10], utilizando diferentes valores de B, obteniendo estimaciones muy similares a las mostradas en dicho ejemplo.

B:	50	100	250	500	1000
Media	19.64	24.03	22.54	23.63	22.89

Tabla 2.2: Estimación bootstrap del error estándar para la media.

2.2. Bootstrap paramétrico.

La diferencia fundamental entre el bootstrap no-paramétrico y el paramétrico, es la manera en que se obtienen las muestras bootstrap. En el bootstrap paramétrico se asume que la función de distribución poblacional pertenece a una familia paramétrica F_θ, con $\theta \in \Theta$, por lo cual resulta lógico estimar θ a partir de un estimador $\hat{\theta}$ y obtener remuestras a partir de $F_{\hat{\theta}}$ y no de $\hat{F}_n$. Su nombre pues, reside en que la distribución estimada a partir de la cual se simulan los datos, se obtiene de un modelo paramétrico.

2.2.1. Estimación bootstrap del error estándar

El algoritmo para estimar el error estándar de un estadístico, cuando se trabaja con bootstrap paramétrico, es muy similar al del bootstrap no-paramétrico, en varios de sus puntos. La diferencia principal estriba en la manera que se generan las muestras bootstrap. A continuación se presenta este algoritmo.

- Se tiene $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, una muestra aleatoria de tamaño n proveniente de $F(x;\theta)$, con θ un parámetro desconocido.
- Se cuenta con un estimador $\hat{\theta}$ de θ.
- Se extraen, de manera aleatoria, B muestras independientes $\mathbf{x}^{*1}, \mathbf{x}^{*2}, \ldots, \mathbf{x}^{*B}$ de la distribución parametrizada $F(x;\hat{\theta})$.
- Se estima el error estándar $se_F(\hat{\theta})$ por medio de la desviación estándar muestral de las B muestras bootstrap

$$\hat{se}_B(\hat{\theta}^*) = (\sum_{b=1}^{B} [\hat{\theta}^*(b) - \hat{\theta}^*(\cdot)]^2/(B-1))^{1/2}, \tag{2.6}$$

donde $\hat{\theta}^*(\cdot) = \sum_{b=1}^{B} \hat{\theta}^*(b)/B$.

2.3. Construcción de intervalos bootstrap

La mayoría de las estimaciones por intervalo que se enseñan en los cursos de estadística están basadas en trabajar con una cantidad pivotal y suponer un comportamiento asintótico que en la práctica pudiera resultar inexacto. Una aplicación muy útil del bootstrap es la construcción de intervalos de confianza, para lo cual hay diversas maneras de construirlos.

Cuando se desea una estimación por intervalo para un parámetro θ, y se cuenta con un estimador $\hat{\theta}$, que tiene una distribución aproximadamente normal alrededor de θ, hemos visto que puede construirse intervalo de la siguiente manera,

$$\hat{\theta} \pm t_{1-\alpha/2,n-1} \times se(\hat{\theta}), \tag{2.7}$$

donde $se(\hat{\theta})$ es el error estándar del estimador.

En caso que supuestos de normalidad no se cumplan, o bien se desee trabajar con bootstrap, la construcción de estos intervalos puede realizarse de diversas maneras. Una de ellas es calculando los percentiles del estadístico de interés, que se obtiene a través de remuestreo. Esto es, si se tiene un estimador $\hat{\theta}$, se realizan B simulaciones bootstrap, y para los valores ordenados de $\hat{\theta}^*$ se identifican los que se encuentran en los lugares $(\alpha/2 \times B)$ y $(1 - \alpha/2 \times B)$. Tendremos así que $\hat{\theta}^*_{\alpha/2}$ y $\hat{\theta}^*_{1-\alpha/2}$, cuantiles de los valores simulados $\hat{\theta}^*$, que dejan un área $\alpha/2$ y $1 - \alpha/2$ a su izquierda, respectivamente, nos permiten construir un intervalo de confianza para el parámetro θ de interés. De esta manera es posible construir un intervalo de la forma

$$(\hat{\theta}^*_{\alpha/2}, \hat{\theta}^*_{1-\alpha/2}), \tag{2.8}$$

Este intervalo resulta muy intuitivo para los estudiantes, sin embargo, tal como muestra Hesterberg [5], estos intervalos resultan ser muy angostos cuando se trabaja con muestras pequeñas; de hecho, hace una analogía para estos intervalos, explicando que su uso equivale a utilizar $Z_{1-\alpha/2} * \sigma/\sqrt{n}$ en lugar de $t_{1-\alpha/2,n-1} * s/\sqrt{n}$ en el cálculo de un intervalo de confianza.

Otro tipo de intervalo bootstrap es conocido como intervalo-t, o intervalo bootstrap-t. En este tipo de intervalo se estima la distribución del estadístico t que se muestra a continuación

$$T = \frac{\hat{\theta} - \theta}{se(\hat{\theta})}, \tag{2.9}$$

y donde $se(\hat{\theta})$ es el error estándar calculado de la muestra original. Para estimar la distribución de (2.9) se trabaja la distribución bootstrap de

$$T^* = \frac{\hat{\theta}^* - \hat{\theta}}{\hat{se}_B(\hat{\theta}^*)}. \tag{2.10}$$

Por ejemplo, si θ fuera la media poblacional μ, podemos utilizar $\bar{x}$ como estimador de μ, y la Ecuación (2.10) sería la distribución bootstrap de

$$T^* = \frac{\bar{X}^* - \bar{X}}{\hat{se}_B(\bar{X}^*)}. \tag{2.11}$$

Como muestra [5], el estadístico t, como el mostrado en (2.9) no sigue una distribución t-Student cuando la población es sesgada. El intervalo de confianza bootstrap-t está basado en el estadístico t pero estima los cuantiles de la distribución actual usando los datos en lugar de una tabla o una distribución $t-$Student. De esta manera, si se denota por q_α al cuantil que deja un área de α a su izquierda, en la distribución bootstrap-t, entonces

$$\alpha/2 = P\left(\frac{\hat{\theta}^* - \hat{\theta}}{\hat{se}_B(\hat{\theta}^*)} < q_{\alpha/2}\right) \approx P\left(\frac{\hat{\theta} - \theta}{se(\hat{\theta})} < q_{\alpha/2}\right) = P(\hat{\theta} - q_{\alpha/2} se(\hat{\theta}) < \theta). \tag{2.12}$$

Similarmente puede hacerse para el cuantil $q_{1-\alpha/2}$, con lo cual, el intervalo resultante es:

$$(\hat{\theta} + q_{\alpha/2} se(\hat{\theta}), \hat{\theta} + q_{1-\alpha/2} se(\hat{\theta})). \tag{2.13}$$

2.4. Bootstrap t-test

Existe una dualidad entre los intervalos de confianza y las pruebas de hipótesis. Un intervalo al $(1-\alpha)\,\%$ de confianza para un parámetro θ, incluye valores de θ para los cuales no se rechazará la hipótesis nula

$$H_0 : \theta = \theta_0, \tag{2.14}$$

al nivel de significancia α. Ahora, es posible efectuar una prueba de hipótesis al nivel de significancia α, simplemente no rechazando la hipótesis nula $H_0 : \theta = \theta_0$, si θ_0 está contenido en el intervalo del $(1-\alpha)\,\%$ confianza y rechazando la hipótesis nula cuando θ_0 esté fuera del intervalo de confianza.

Aunque existen sugerencias de calcular 2.10 sustituyendo $\hat{\theta}$ por $\hat{\theta_0}$, también es válido lo propuesto, esto es, utilizar el intervalo de confianza obtenido para realizar la prueba de hipótesis correspondiente. Así, al plantear hipótesis del tipo:

$$H_0 : \theta = \theta_0, \quad H_1 : \theta \neq \theta_0, \tag{2.15}$$

simplemente verificaremos si el valor θ_0 se encuentra fuera o dentro del intervalo de confianza, como evidencia para rechazar o no la hipótesis nula.

Capítulo 3

Intervalos de confianza y prueba de hipótesis

En este capítulo se presenta la construcción de algunos intervalos de confianza para uno y dos parámetros, ello a partir de una cantidad pivotal. De igual manera se realizan pruebas de hipótesis para uno y dos parámetros. Los casos que se abordan son relativos a la media y diferencia de medias de una poblaciones distribuidas normalmente. Todo ello se presenta con el enfoque frecuentista que generalmente se enseña en los cursos de estadística, donde se verifican supuestos necesarios para utilizar las pruebas adecuadas y finalizar analizando los resultados obtenidos. Los ejemplos que se incluyen en este capítulo también son abordados mediante bootstrap$-t$ y sus resultados son comparados con los obtenidos mediante los procedimientos que usualmente se incluyen en los cursos de estadística. Todos los programas fueron realizados en el software R versión 4.2.0, y se encuentran disponibles en el Apéndice B.

3.1. Intervalo de confianza y prueba de hipótesis para un parámetro

3.1.1. Intervalo de confianza para una media

Para el caso de construir un intervalo de confianza para la media μ de una población normal con varianza σ^2 conocida, se considera que $X_1, \ldots, X_n$ son variables aleatorias independientes, idénticamente distribuidas $X_i \sim N(\mu, \sigma^2)$. De esta manera, el estadístico $(\bar{X} - \mu)/(\sigma/\sqrt{n})$, que se distribuye normal estándar, es una cantidad pivotal que puede utilizarse para calcular un intervalo de confianza para μ.

Para construir un intervalo de confianza para μ, procedemos de la siguiente manera:

$$P\left(-Z_{1-\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq Z_{1-\alpha/2}\right) = 1 - \alpha,$$

donde $Z_{1-\alpha/2}$ es el cuantil en la distribución normal estándar, que deja un área de $1 - \alpha/2$ a su izquierda. Ahora, con un poco de manipulación algebraica se obtiene que los extremos de

un intervalo al $(1-\alpha)100\,\%$ de confianza para la media μ es el siguiente:

$$\left\{\mu : \bar{X} - Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right\}. \tag{3.1}$$

En el caso que σ^2 sea desconocida, puede utilizarse la siguiente cantidad pivotal para obtener un intervalo de confianza para μ,

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1},$$

la cual se distribuye como una t-Student con $n-1$ grados de libertad.

$$P\left(-t_{1-\alpha/2,n-1} \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_{1-\alpha/2,n-1}\right) = 1-\alpha.$$

De donde se obtiene que al nivel de confianza $1-\alpha$, el intervalo para μ resulta

$$\left\{\mu : \bar{X} - t_{1-\alpha/2,n-1}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{1-\alpha/2,n-1}\frac{S}{\sqrt{n}}\right\}. \tag{3.2}$$

Los datos del siguiente ejemplo, el cual fue tomado de [7, p. 228] permiten mostrar la construcción de un intervalo de confianza para una media, en el cual se utilizará la fórmula presentada en Ecuación (3.2) y también se trabajará por medio de bootstrap no-paramétrico y paramétrico.

Ejemplo 3.1 *Bajo una dieta especial, doce ratas lograron los siguientes aumentos de peso (en gramos) desde el nacimiento hasta los tres meses de edad: 55.3, 54.8, 65.9, 60.7, 59.4, 62.0, 62.1, 58.7, 64.5, 62.3, 67.6, 61.2. Calcular un intervalo de confianza para el aumento de peso promedio de las ratas, durante este periodo.*

Si desea utilizarse el intervalo de confianza presentado en (3.2), es necesario verificar que los datos satisfacen el supuesto de normalidad, lo cual puede realizarse mediante una prueba de normalidad. La que utilizaremos es la prueba de Shapiro-Wilk, y el procedimiento se realizará con el software R. Documentación sobre esta prueba se brinda en el Apéndice A.2.2.

Antes de realizar una prueba de normalidad analizamos los datos de manera descriptiva. A continuación se presenta un resumen de algunas medidas principales y en la Figura 3.1 se muestra un diagrama de caja que nos permite ver que la distribución no parece sesgada, no tiene datos atípicos ni datos aberrantes (outliers), por lo que pudiera pasar una prueba de normalidad de los datos.

Veamos un resumen de los datos:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  54.80   59.23   61.60   61.21   62.85   67.60
```

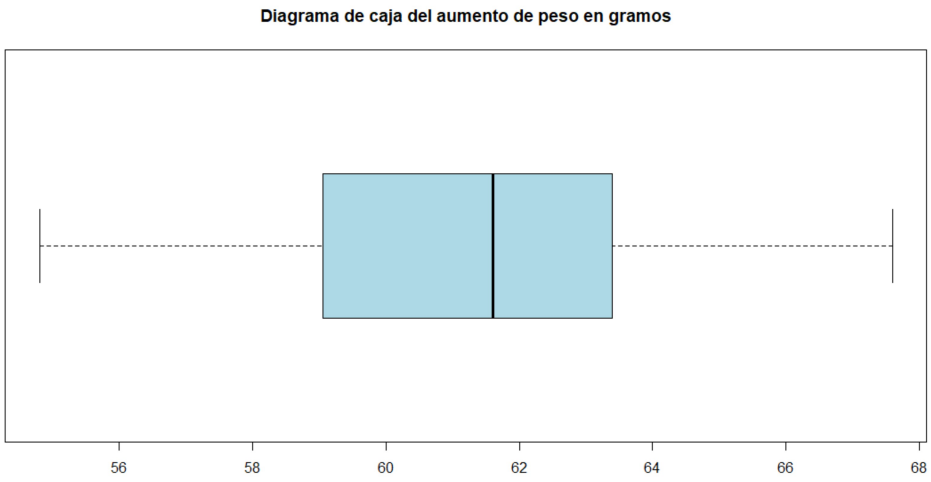

Figura 3.1: Diagrama de caja del aumento de peso en gramos.

A continuación se realiza la prueba de normalidad de Shapiro-Wilk. En ésta, de acuerdo a lo expuesto en el Apéndice A.2.2, en la hipótesis nula se plantea que la variable X, que representa el aumento de peso en gramos, desde el nacimiento hasta los tres meses de edad, se distribuye normalmente. De esta manera se tiene:

$$H_0 : \mathbf{X} \sim N(\mu, \sigma^2),$$
$$H_1 : \mathbf{X} \nsim N(\mu, \sigma^2).$$

Utilizando la instrucción *shapiro.test*() del software R, se obtiene:

```
  Shapiro-Wilk normality test

data:  x
W = 0.96496, p-value = 0.8516
```

Puede observarse que de acuerdo al p-valor obtenido, el cual es mayor que un nivel de significancia digamos de 0.05, no tenemos evidencia para rechazar la hipótesis nula; podemos suponer entonces que se cumple el supuesto de normalidad necesario para utilizar el siguiente intervalo de confianza

$$\bar{X} \pm t_{1-\alpha/2,n-1}\frac{S}{\sqrt{n}},$$

donde $t_{1-\alpha/2,n-1}$ denota el cuantil en la distribución t con $n-1$ grados de libertad, que deja un área de $1-\alpha/2$ a su izquierda.

Considerando un nivel de confianza del $1-\alpha = 95\,\%$, el cuantil que se requiere es un valor $t_{0.975,11} = 2.200985$. Sustituyendo éste, el valor que toman la media y desviación estándar

muestrales, se tiene que:

$$61.21 \pm 2.200985 \left(\frac{3.838906}{\sqrt{12}} \right),$$

$$61.21 \pm 2.44,$$

es decir,

$$(58.77, 63.65),$$

es un intervalo al $1 - \alpha = 95\,\%$ de confianza para la media de real de los aumentos de peso de este tipo de ratas, en el periodo considerado.

A continuación se utilizará el procedimiento bootstrap que se presentó en la sección 2.1 para obtener un intervalo de confianza bootstrap para la media real de los aumentos de peso en ratas. Para ello se utilizará la cantidad pivotal presentada en el Capítulo 2, $T^* = \frac{\bar{X}^* - \bar{X}}{S^*/\sqrt{n}}$ el cual se calcula remuestreando de los datos originales, simulando T^* tantas veces como se quiera, en este caso se utiliza el software R para generar 10,000 muestras bootstrap, con reemplazamiento, cada una de tamaño $n = 12$.

Veamos un resumen de T^* y $\bar{X}^*$, respectivamente:

```
      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-5.41094 -0.67315  0.00000  0.02125  0.70948  5.04795
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  57.15   60.49   61.21   61.20   61.92   65.22
```

Nótese que la media de la muestra original (61.60) está muy cerca de la media muestral bootstrap (61.21).

Ahora si podemos calcular el intervalo de bootstrap al 95 % confianza,

$$(\bar{X} + t^*_{0.025} \frac{S}{\sqrt{n}}, \bar{X} + t^*_{0.975} \frac{S}{\sqrt{n}}),$$

donde $t^*_{0.025}$ denota el cuantil en la distribución de T^*, que deja un área de 0.025 a su izquierda.

Reemplazando los valores correspondientes, el intervalo de confianza resulta:

$$\left(61.21 - 2.18 \left(\frac{3.83}{\sqrt{12}} \right), 61.21 + 2.28 \left(\frac{3.83}{\sqrt{12}} \right) \right),$$

$$(61.21 - 2.42, 61.21 + 2.52),$$

$$(58.79, 63.73).$$

Aunque el intervalo bootstrap de percentiles es muy intuitivo y de acuerdo a [5] hay mejores opciones, se calculó el intervalo de percentiles basado en la distribución de medias muestrales bootstrap, como puede observarse en la Figura 3.9, donde puede apreciarse el intervalo obtenido usando la distribución $t-$Student y el obtenido por medio de percentiles.

Por otra parte, en la Figura 3.15, puede observarse que los cuantiles obtenidos con la distribución de T^*, (-2.182729, 2.278027), son ligeramente diferentes de los cuantiles en la distribución $t-$Student (-2.200985 2.200985); sin embargo, el intervalo bootstrap-t, que utiliza cuantiles de T^* en lugar de los obtenidos a partir de una distribución $t-$Student, resulta (58.79, 63.73) y el calculado usando la Ecuación (3.2) es $(58.77, 63.65)$. Lo importante es notar que ambos intervalos son muy parecidos y que el generado por medio de bootstrap no requiere el supuesto de normalidad que requiere la construcción del otro intervalo.

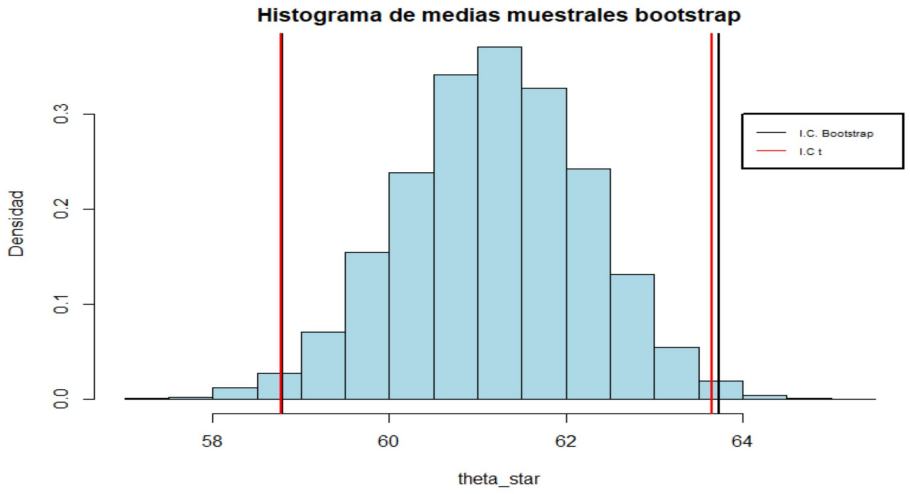

Figura 3.2: Histograma de medias muestrales bootstrap.

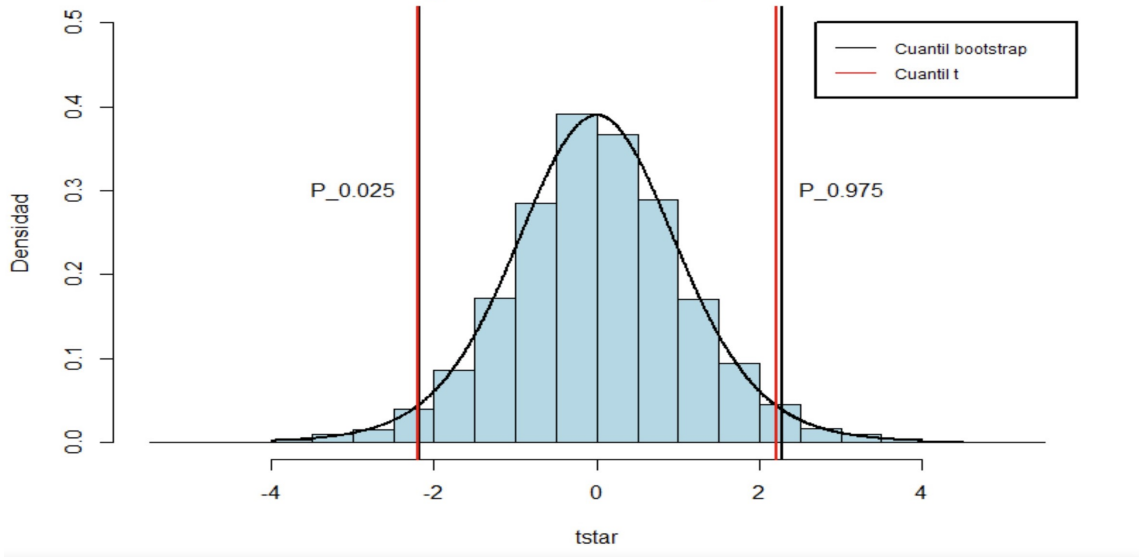

Figura 3.3: Histograma de T^*.

Ahora se utilizará el bootstrap paramétrico que se presentó anteriormente en la sección 2.2, para obtener el intervalo de confianza bootstrap paramétrico.

La muestra bootstrap es extraída de $F(\hat{\theta})$. Para cada muestra bootstrap

$$x^{*1}, x^{*2}, \ldots, x^{*B};$$

cada uno consiste de $n = 12$ valores de datos extraídos de una distribución parametrizada.

Veamos un resumen de T^* y $\bar{X}^*$:

Min	1st Qu	Median	Mean	3rd Qu	Max
-4.932949	-0.693111	0.007109	0.000162	0.692798	5.296340

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
57.10	60.46	61.22	61.21	61.96	65.29

Ahora si podemos calcular el intervalo de confianza bootstrap

$$(\bar{X} + t^*_{0.025}\frac{S}{\sqrt{n}}, \bar{X} + t^*_{0.975}\frac{S}{\sqrt{n}}).$$

Reemplazando valores, el intervalo de confianza queda de la siguiente manera:

$$\left(61.22 - 2.21\left(\frac{3.83}{\sqrt{12}}\right), 61.22 + 2.17\left(\frac{3.83}{\sqrt{12}}\right)\right),$$

$$\left(61.22 - 2.456, 61.22 + 2.41\right),$$

$(58.75, 63.62)$.

Podemos notar que ambos intervalos de confianza son muy parecidos.

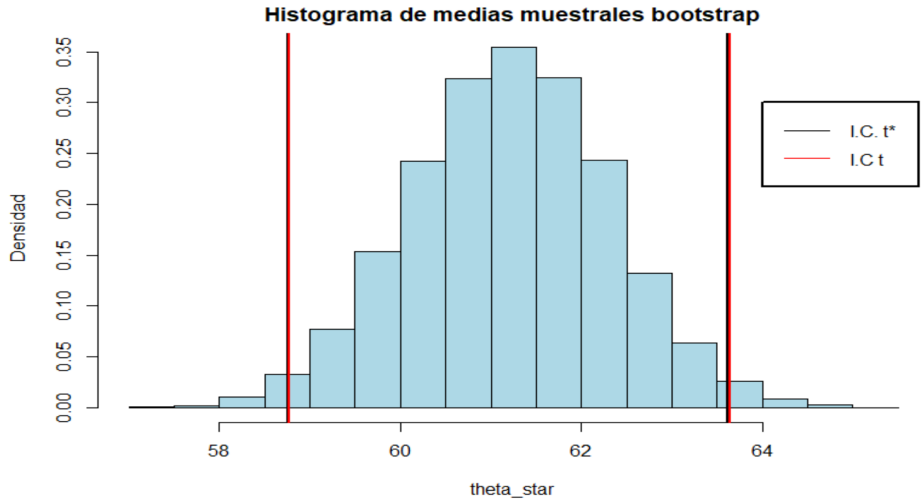

Figura 3.4: Histograma de medias muestrales usando bootstrap paramétrico.

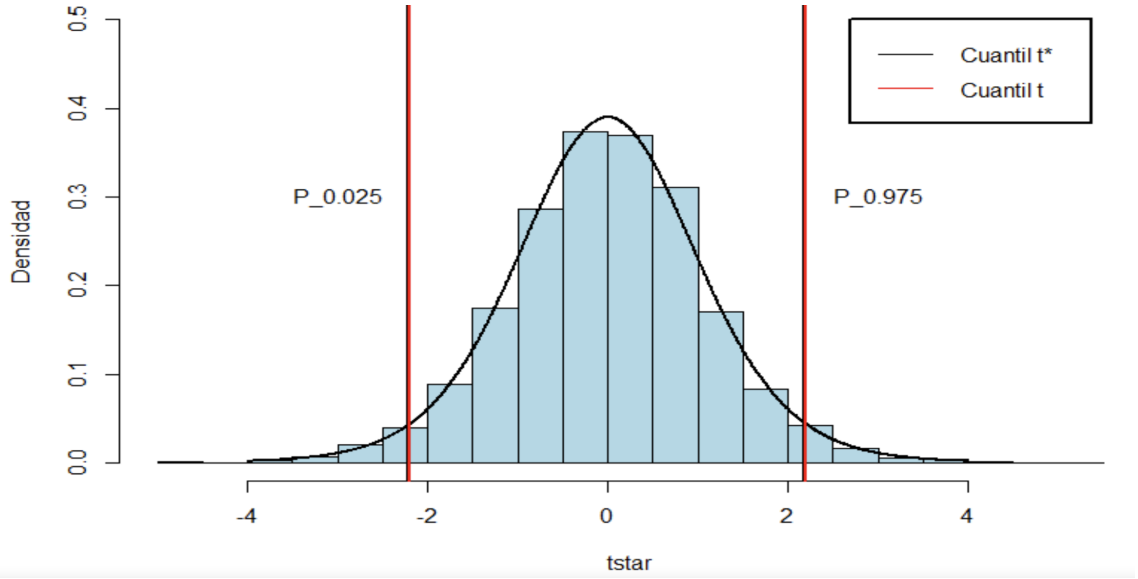

Figura 3.5: Histograma de T^* usando bootstrap paramétrico.

Observemos que la media de la muestra original (61.60) está muy cerca de la media muestral bootstrap (61.2). Los errores estándar para la muestra original y los obtenidos en el remuestreo bootstrap son 1.108197 y 1.099712, respectivamente. Los cuantiles de T^* (-2.216325, 2.175367) son ligeramente diferentes de los cuantiles t-Student (-2.200985 2.200985). Esto da como resultado una estimación de los intervalo de confianza: $(58.75, 63.62)$

cuando se utiliza T^* y $(58.77, 63.65)$ usando $t-$Student y suponiendo normalidad.

El código donde se hacen todos los cálculos se encuentra en el apéndice B.4.

3.1.2. Prueba de hipótesis para una media, varianza desconocida

Cuando se tiene una muestra aleatoria de n observaciones $x_1, x_2, \ldots x_n$ tomadas de una población normal con media μ y varianza σ^2 desconocida, es común estar interesados en probar una hipótesis del tipo $H_0 : \mu = \mu_0$. En general, podría ser de interés cualquiera de las siguientes opciones de hipótesis nulas y alternativas que se muestran en la Tabla 3.1:

Caso 1	Caso 2	caso 3
$H_0 : \mu \leq \mu_0$	$H_0 : \mu \geq \mu_0$	$H_0 : \mu = \mu_0$
$H_1 : \mu > \mu_0$	$H_1 : \mu < \mu_0$	$H_1 : \mu \neq \mu_0$

Tabla 3.1: Hipótesis nulas y alternativas para una media.

Dadas las características del problema, puede utilizarse el siguiente estadístico de prueba:

$$T_c = \frac{\bar{X} - \mu_o}{S/\sqrt{n}}, \tag{3.3}$$

el cual sigue una distribución $t-$Student con $n-1$ grados de libertad.

Con base en este estadístico, y para cada una de las opciones mostradas en la Tabla 3.1, se rechazará H_0 cuando, al nivel de significancia α el valor t_c que tome el estadístico T_c sea tal que:

- Caso 1: $t_c > t_{1-\alpha,n-1}$.
- Caso 2: $t_c < -t_{1-\alpha,n-1}$.
- Caso 3: $t_c < -t_{1-\alpha/2,n-1} \quad o \quad t_c > t_{1-\alpha/2,n-1}$.

Para ilustrar esta prueba de hipótesis se utilizará el siguiente ejemplo, el cual fue tomado de [2, p.26] y donde se analizan datos referentes a las aflatoxinas, que son un tipo de toxinas producidas por ciertos hongos en cultivos agrícolas como el maíz, los cacahuates o maní, los frutos secos, etc.

Ejemplo 3.2 ***Residuos de aflatoxinas en la mantequilla de maní.*** *En pruebas efectuadas a 12 lotes de mantequilla de maní se midieron los residuos de aflatoxinas, en partes por billón, resultando las siguientes mediciones: 4.94, 5.06, 4.53, 5.07, 4.99, 5.16, 4.38, 4.43, 4.93, 4.72, 4.92 y 4.96.*

Supóngase que en los sembradíos de maní se desea saber sobre posibles cambios en los niveles de aflatoxinas, y que por información histórica se venía trabajando con un promedio de 5.7 partes por billón, para estos residuos de aflatoxinas. Entonces, puede interesar probar:

$$H_0 : \mu = 5.7, \qquad H_1 : \mu \neq 5.7, \tag{3.4}$$

con $\alpha = 0.05$ Para abordar este tipo de problemas con las herramientas que generalmente se imparten en cursos de estadística, primeramente se analizan descriptivamente los datos y se verifica la normalidad de éstos, lo cual en este caso se efectuará por medio de una prueba de Shapiro–Wilk, para lo cual se plantea:

$$H_0 : \mathbf{X} \sim N(\mu, \sigma^2),$$
$$H_1 : \mathbf{X} \nsim N(\mu, \sigma^2),$$

donde $\mathbf{X}$ representa las mediciones de aflatoxinas en sembradíos de maní.

Antes de realizar esta prueba se muestra un resumen de los datos y un diagrama de caja que nos permitirá resumir las características principales de los mismos.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.380   4.673   4.935   4.841   5.008   5.160
```

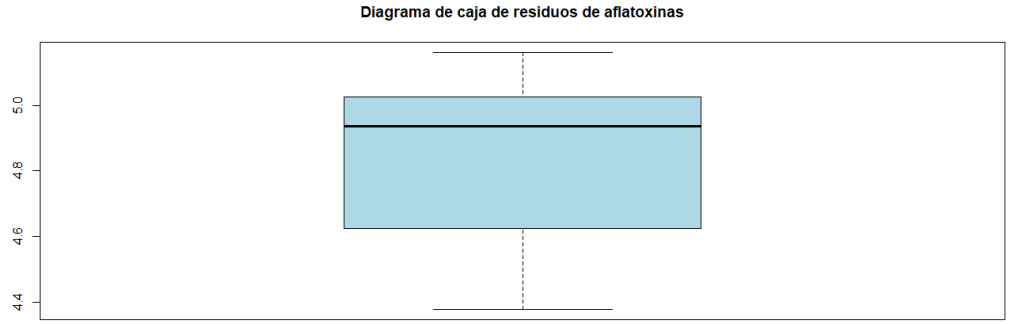

Figura 3.6: Diagrama de caja residuos de aflatoxinas en partes por billón.

En el diagrama de caja, aunque son pocas observaciones, puede verse que la mediana se encuentra más cercana al tercer cuartil, aunque no hay datos atípicos ni aberrantes, por lo cual podría cumplirse el supuesto de normalidad, que se verifica a continuación, utilizando R y una prueba de Shapiro-Wilk al nivel $\alpha = 0.05$.

```
  Shapiro-Wilk normality test

data:  x
W = 0.87092, p-value = 0.06713.
```

Podemos observar que el $p-$valor es mayor que nuestro nivel de significancia elegido (0.05), por lo que no tenemos evidencia para rechazar la hipótesis nula. Entonces, puede efectuarse la prueba de hipótesis planteada en (3.4), para la cual se obtienen los siguientes resultados al utilizar la opción *t.test* del software R.

```
  One Sample t-test
```

```
 data:  x
 t = -11.358, df = 11, p-value = 2.043e-07
 alternative hypothesis: true mean is not equal to 5.7
 95 percent confidence interval:
    4.674344 5.007323
 sample estimates:
mean of x
  4.840833
```

Los resultados que arroja R muestran un p-valor muy pequeño, de $1.021e^{-07}$ y menor que el nivel de significancia 0.05, por lo que se tiene evidencia para rechazar la hipótesis nula y podría pensarse que la concentración promedio de aflatoxinas ha disminuido.

Ahora lo haremos utilizando Bootstrap no paramétrico.

Veamos un resumen de T^* y $\bar{X}^*$ respectivamente.

```
     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-4.82451 -0.62647  0.03183  0.17487  0.76307 11.37250
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.556   4.791   4.843   4.840   4.891   5.070
```

Los cuantiles 0.025 y 0.975 de T^* resultan: -1.916293 y 3.146077 respectivamente, el intervalo de confianza bootstrap es $(4.695879, 5.078813)$, por lo que puede verse que el valor 5.7 no está incluido en el intervalo de confianza, y entonces, se tiene evidencia en contra del H_0.

Utilizando un bootstrap paramétrico obtenemos el siguiente resumen de resultados:

Veamos un resumen de T^* y $\bar{X}^*$ respectivamente.

```
      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-4.932949 -0.693111  0.007109  0.000162  0.692798  5.296340
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.561   4.790   4.841   4.841   4.892   5.119
```

Cuando se trabaja T^* de forma paramétrica, los cuantiles de probabilidad 0.025 y 0.975 resultan: -2.199762 y 2.155468 respectivamente, por lo que puede observarse que cambian bastante con respecto a los obtenidos de manera no-paramétrica, y el intervalo de confianza bootstrap es $(4.695879, 5.078813)$. Nuevamente vemos que el valor 5.7 no se encuentra dentro del intervalo de confianza, por que hay evidencia en la muestra para rechazar la hipótesis nula H_0.

El Código en R del ejemplo se encuentra en el apéndice B.5

3.2. Intervalos de confianza y prueba de hipótesis para dos medias

Existen ocasiones en que se desea comparar las medias de dos poblaciones. El problema que se abordará aquí es cuando las poblaciones se distribuyen normalmente, sus varianzas son desconocidas y se toman muestras independientes en estas poblaciones. La comparación de medias se ilustrará a través de intervalos de confianza y también mediante pruebas de hipótesis.

3.2.1. Intervalo de confianza y prueba de hipótesis para dos medias, varianzas desconocidas supuestas no homogéneas

En el caso de tener muestras aleatorias independientes $X_{11}, X_{12}, \ldots X_{1n_1}$ y $X_{21}, X_{22}, \ldots X_{2n_2}$, de tamaños n_1 y n_2, donde $X_1 \sim N(\mu_1, \sigma_1^2)$ y $X_2 \sim N(\mu_2, \sigma_2^2)$, con σ_1^2 y σ_2^2 desconocidas, una cantidad pivotal que permite construir un intervalo de confianza para la diferencia de medias es la siguiente

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}}, \tag{3.5}$$

la cual sigue una distribución t-Student con

$$v = \frac{[(S_1^2/n_1) + (S_2^2/n_2]^2}{\left[\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}\right]},$$

grados de libertad [10].

Para encontrar el intervalo del $(1-\alpha)100\,\%$ de confianza para $\mu_1 - \mu_2$ partimos de:

$$P\left(-t_{1-\alpha/2,v} \leq \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}} \leq t_{1-\alpha/2,v}\right) = 1 - \alpha, \tag{3.6}$$

donde S_1^2 y S_2^2 son las varianzas muestrales. Con un poco de álgebra podemos llegar al intervalo del $(1-\alpha)100\,\%$ de confianza para $\mu_1 - \mu_2$:

$$P\left((\bar{X}_1 - \bar{X}_2) - t_{1-\alpha/2,v}\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)} \leq \mu_1 - \mu_2 \leq (\bar{X}_1 - \bar{X}_2) + t_{1-\alpha/2,v}\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}\right) = 1 - \alpha,$$

cuyos extremos podemos escribir como:

$$(\bar{X}_1 - \bar{X}_2) \pm t_{1-\alpha/2,v}\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}. \tag{3.7}$$

El intervalo generado puede utilizarse para realizar una prueba de hipótesis y comparar las medias de estas dos poblaciones normales. La hipótesis nula a probar puede plantearse como $H_0 : \mu_1 - \mu_2 = \delta_0$, o bien especificarla como en la siguiente tabla, dependiendo del caso a resolver:

Caso 1	Caso 2	caso 3
$H_0 : \mu_1 - \mu_2 \leq \delta_0$	$H_0 : \mu_1 - \mu_2 \geq \delta_0$	$H_0 : \mu_1 - \mu_2 = \delta_0$
$H_1 : \mu_1 - \mu_2 > \delta_0$	$H_1 : \mu_1 - \mu_2 < \delta_0$	$H_1 : \mu_1 - \mu_2 \neq \delta_0$

Se rechazará la hipótesis nula cuando

- Caso 1: $t_c > t_{1-\alpha,v}$.
- Caso 2: $t_c < -t_{1-\alpha,v}$.
- Caso 3: $t_c < -t_{1-\alpha/2,v} \leq o \leq t_{1-\alpha/2,v}$.

Si no conocemos las varianzas poblacionales σ_1^2 y σ_2^2, y éstas se suponen no homogéneas, el estadístico de prueba será el siguiente:

$$T_c = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}, \tag{3.8}$$

que se distribuye como t-Student con v grados de libertad.

Para construir un intervalo bootstrap-t, ya sea de forma paramétrica, o no-paramétrica, se generarán valores equivalentes a los de la distribución $t-$Student, mediante la siguiente fórmula:

$$T^* = \frac{(\bar{X}_1^* - \bar{X}_2^*) - (\bar{X}_1 - \bar{X}_2)}{\sqrt{\left(\frac{S_1^{*2}}{n_1} + \frac{S_2^{*2}}{n_2}\right)}}. \tag{3.9}$$

También existe la propuesta de generar los valores de una distribución $t-$Student mediante la siguiente fórmula,

$$T^* = \frac{(\bar{X}_1^* - \bar{X}_2^*) - \delta_0}{\sqrt{\frac{S_1^{*2}}{n_1} + \frac{S_2^{*2}}{n_2}}}. \tag{3.10}$$

En cualquiera de los dos casos, el intervalo de confianza bootstrap para $\mu_1 - \mu_2$ se calcularía como:

$$P\left((\bar{X}_1-\bar{X}_2)+t^*_{\alpha/2}\sqrt{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)}\leq \mu_1-\mu_2\leq (\bar{X}_1-\bar{X}_2)+t^*_{1-\alpha/2}\sqrt{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)}\right)=1-\alpha, \tag{3.11}$$

donde $t^*_{\alpha/2}$ y $t^*_{1-\alpha/2}$ son los cuantiles de probabilidad $\alpha/2$ y $1-\alpha/2$ de la distribución de valores t^*, que tomó T^*.

Por medio del siguiente ejemplo que fue tomado de [12, p. 358-359] se ejemplificará lo planteado anteriormente.

Ejemplo 3.3 *El Departamento de Zoología de Virginia Tech llevó a cabo un estudio para determinar si existe una diferencia significativa en la densidad de organismos en dos estaciones diferentes ubicadas en Cedar Run, una corriente secundaria que se localiza en la cuenca del río Roanoke. El drenaje de una planta de tratamiento de aguas negras y el sobreflujo del estanque de sedimentación de la Federal Mogul Corporation entran al flujo cerca del nacimiento del río. Los siguientes datos proporcionan las medidas de densidad, en número de organismos por metro cuadrado, en las dos estaciones colectoras*
Estación 1(E_1): 5030, 13700, 10730, 11400, 860, 2200, 4250, 15040, 4980, 11910, 8130, 26850, 17660, 22800, 1130, 1690.
Estación 2(E_2): 2800, 4670, 6890, 7720, 7030, 7330, 2810, 1330, 3320, 1230, 2130, 2190

Veamos un resumen de y un diagrama de caja de E_1 y E_2, respectivamente.

```
   Min.  1st Qu.  Median  Mean   3rd Qu.   Max.
   860    3738    9430    9898   14035    26850
```

```
   Min.  1st Qu.  Median   Mean   3rd Qu.   Max.
  1230    2175    3065     4121    6925    7720
```

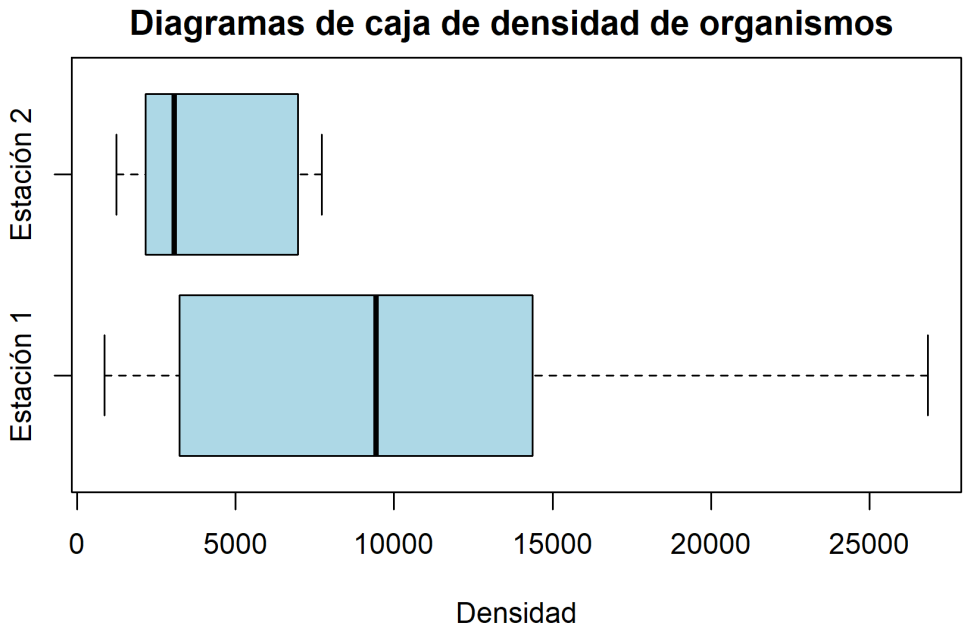

Figura 3.7: Diagrama de caja de la densidad de organismos de E_1 *y* E_2.

Como puede observarse en el resumen de datos, las medias muestrales son muy diferentes, lo mismo pasa con las medianas. Es decir, al parecer las distribuciones difieren en localización. Con respecto a los diagramas de caja, se observa mayor variabilidad en los datos obtenidos en la Estación 1. Sería conveniente entonces, verificar el supuesto de normalidad de los datos y hacer una prueba de hipótesis para comparar las varianzas poblacionales. En caso de no rechazar el supuesto de normalidad en ambas muestras y rechazar la hipótesis de homogeneidad de varianzas, es posible utilizar el intervalo de confianza mostrado en (3.7).

Utilizando el procedimiento descrito en el Apéndice A, se efectúa una prueba de normalidad Shapiro-Wilk, al nivel de significancia del 0.05, para los datos de las estaciones E1 y E2, respectivamente. Así, para la muestra $\mathbf{X_1}$ correspondiente a los datos de la Estación 1, tendríamos:

$$H_0 : \mathbf{X_1} \sim N(\mu_1, \sigma_1^2), \tag{3.12}$$
$$H_1 : \mathbf{X_1} \nsim N(\mu_1, \sigma_1^2). \tag{3.13}$$

De igual manera se plantea para los datos correspondientes a la Estación 2.

$$H_0 : \mathbf{X_2} \sim N(\mu_2, \sigma_2^2), \tag{3.14}$$
$$H_1 : \mathbf{X_2} \nsim N(\mu_2, \sigma_2^2). \tag{3.15}$$

Los resultados de la prueba de Shapiro-Wilk, para ambos sitios, se muestran a continuación:

```
Shapiro-Wilk normality test
data:  E1
```

```
W = 0.9218, p-value = 0.1803

Shapiro-Wilk normality test
data:  E2
W = 0.86141, p-value = 0.05092
```

De acuerdo a los $p-$valores obtenidos, no hay evidencia en las muestras para rechazar la hipótesis nula, pues ambos $p-$valores son superiores al nivel de significancia elegido (0.05)

Para la prueba que a continuación se describe, se sugiere ver el Apéndice A.2.3 donde se describe a detalle la prueba que se efectúa enseguida, la cual permite verificar si es posible suponer varianzas poblacionales homogéneas. Dependiendo del resultado de esta prueba se selecciona el intervalo de confianza apropiado para comparar las medias poblacionales.

Se plantean entonces las siguientes hipótesis:

$$H_0 : \sigma_1^2 = \sigma_2^2,$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2.$$

Sustituyendo la información del problema en la fórmula mostrada en (A.4), tenemos,

$$f_c = \frac{62005060}{6147936} = 10.08. \tag{3.16}$$

Si comparamos con el cuantil en la distribución F con 15 y 11 grados de libertad, que deja un área de 0.975 a su izquierda, éste resulta:

$$f_{n_2-1,1-\alpha/2}^{n_1-1} = f_{11,0.975}^{15} = 3.33.$$

Notemos que $f_c = 10.08 > f_{n_2-1,0.975}^{n_1-1} = 3.33$, por lo que se rechaza H_0 y concluimos que las muestras proporcionan evidencia de varianzas poblacionales no homogéneas.

Dado el resultado anterior, para la prueba de hipótesis sobre la diferencia de medias, donde se plantean las hipótesis siguientes:

$$H_0 : \mu_1 - \mu_2 = 0,$$
$$H_1 : \mu_1 - \mu_2 \neq 0,$$

se utilizará el estadístico de prueba presentado en (3.8), donde se sustituyen los cálculos respectivos de medias y varianzas muestrales, obteniendo:

$$T_c = \frac{(9897.5 - 4120.833 - 0)}{\sqrt{\frac{62005060}{16} + \frac{6147936}{12}}} = 2.75, \tag{3.17}$$

Necesitamos calcular los grados de libertad, para eso se utilizará (3.6).

$$v = \frac{[(3875316) + (512328)]^2}{938629752344 + 21873328828} = 20.04,$$

por lo que se toman 20 grados de libertad para el cálculo del cuantil $t_{1-\alpha/2,v} = t_{0.975,20} = 2.08$. Dado que $t_c > 2.08$, se tiene evidencia para rechazar la hipótesis nula $H_0 : \mu_1 - \mu_2 = 0$.

Se calculará el intervalo de confianza que por lo general se imparte en los cursos de estadística, esto es, el presentado en (3.7), donde sustituyendo las medias, varianzas y tamaños de las muestras se tiene

$$(9897.5 - 4120.833) \pm 2.08\sqrt{\left(\frac{62005060}{16} + \frac{6147936}{12}\right)},$$

$$5776.667 \pm 2.08\sqrt{(3875316 + 512328)},$$

$$(1407.26, 10146.07).$$

Puede observarse que el intervalo no incluye el valor cero, por lo cual se concluye que con un 95 % de confianza se tiene evidencia de que las medias parecen no ser iguales.

Ahora se utilizará el procedimiento bootstrap no paramétrico que se detalló anteriormente donde se estima cuantiles en una distribución $t-$Student, por medio de (3.9) para después conformar un intervalo de confianza bootstrap.

Utilizando un procedimiento similar al presentado en los ejemplos anteriores y por medio del software R, se generan $10,000$ muestras a partir de cada una de las dos muestras originales, para después obtener valores T^*.

Veamos un resumen de T^*:

```
     Min.      1st Qu.     Median       Mean      3rd Qu.      Max.
-7.460544  -0.711399  -0.005525  -0.071402   0.625981   3.760360
```

Y un resumen de $\theta^* = \bar{X}_1^* - \bar{X}_2^*$:

```
     Min.      1st Qu.     Median       Mean      3rd Qu.      Max.
  -884.2      4419.7      5764.3     5790.6      7087.9     13760.6
```

A simple vista y partiendo del resumen descriptivo del remuestreo, se observan diferencias entre las estaciones E_1 y E_2. Habiendo obtenido $t^*_{0.025} = -2.36$ y $t^*_{0.975} = 1.87$ en la distribución de T^*, se procede a calcular los extremos del intervalo del 95 % de confianza bootstrap,

$$\left((\bar{X}_1 - \bar{X}_2) + t^*_{0.025}\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}, (\bar{X}_1 - \bar{X}_2) + t^*_{0.975}\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}\right),$$

$$\left((9897.5-4120.8)-2.36\sqrt{\left[\frac{62005060}{16}+\frac{6147936}{12}\right]},(9897.5-4120.8)+1.87\sqrt{\left[\frac{62005060}{16}+\frac{6147936}{12}\right]}\right),$$

$$(834.59, 9694.39).$$

Es interesante observar la diferencia entre los cuantiles que se obtienen cuando se usa la distribución $t-$Student y los obtenidos a partir de los cuantiles 0.025 y 0.975 de T^*; esta diferencia puede apreciarse en la Figura 3.8.

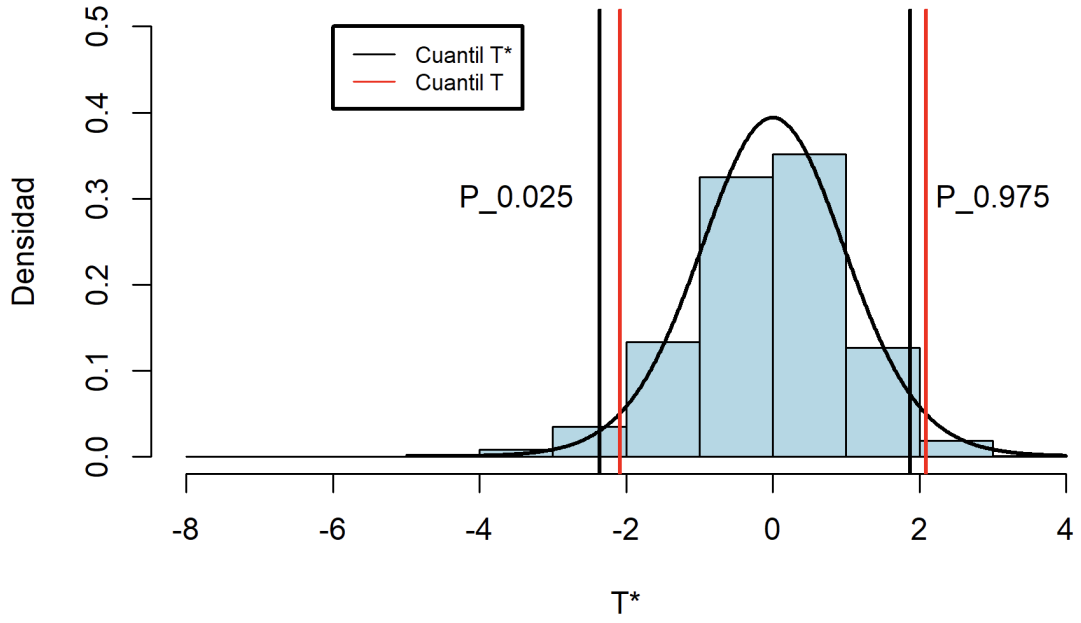

Figura 3.8: Histograma de T^*.@

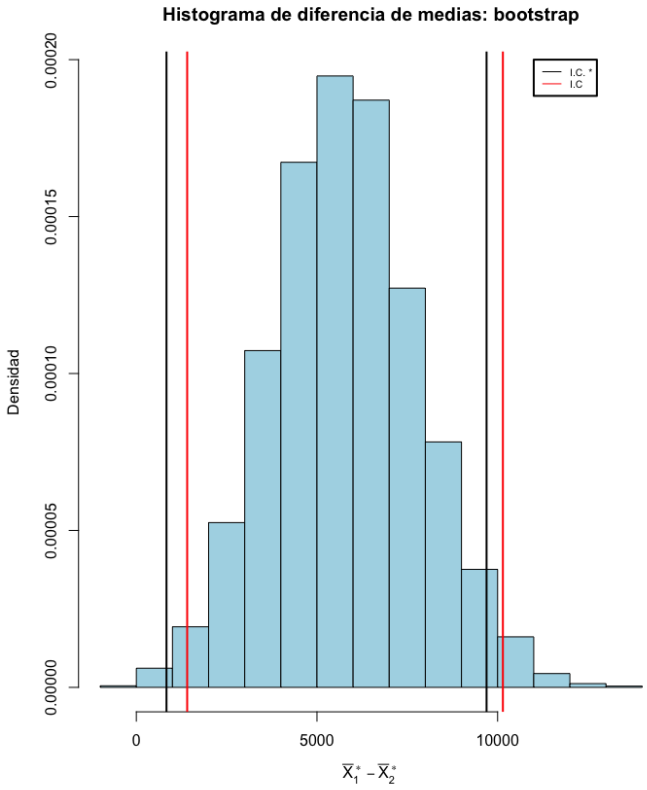

Figura 3.9: Histograma de diferencia de medias muestrales (bootstrap).

De igual manera puede observarse, en la Figura 3.9, las diferencias en los extremos del intervalo obtenido por medio de la formula en (3.7) y la obtenida por medio de bootstrap no-paramétrico.

Ahora utilizaremos el bootstrap paramétrico que se vió anteriormente en la Sección 2.2, donde a partir muestras generadas de distribuciones normales se obtendrá un intervalo de confianza bootstrap para la diferencia de medias $\mu_1 - \mu_2$. Esto es, para simular datos de la Estación 1, se generarán $10,000$ muestras de tamaño n_1 de una distribución normal con $\hat{\mu}_1 = \bar{\mathbf{x}}_1$ y $\sigma_1^2 = s_1^2$, donde $\bar{\mathbf{x}}_1$ y s_1^2 son la media y varianza muestrales que se obtienen a partir de la muestra original $\mathbf{x}$. De manera similar se realizará para la Estación 2. Para cada una de estas muestras se genera un valor t^* y de la distribución de T^* se obtienen los cuantiles que permitirán calcular el intervalo de confianza bootstrap.

A continuación se muestra un resumen de estadísticos para T^* y $\theta^* = \bar{X}_1^* - \bar{X}_2^*$ respectivamente:

```
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-4.508549 -0.666899  0.010926 -0.001017  0.680592  5.249408
```

```
    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  -3303    4406    5800    5776    7160   13697
```

Ahora si podemos calcular el intervalo de confianza bootstrap.

$$\left((\bar{X}_1 - \bar{X}_2) + t^*_{0.025}\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}, (\bar{X}_1 - \bar{X}_2) + t^*_{0.975}\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}\right),$$

$$\left((9897.5 - 4120.8) - 2.12\sqrt{\left(\frac{62005060}{16} + \frac{6147936}{12}\right)}, (9897.5 - 4120.8) + 2.06\sqrt{\left(\frac{62005060}{16} + \frac{6147936}{12}\right)}\right),$$

$$(1334.89, 10084.84).$$

Veamos el histograma de T^* y de θ^*

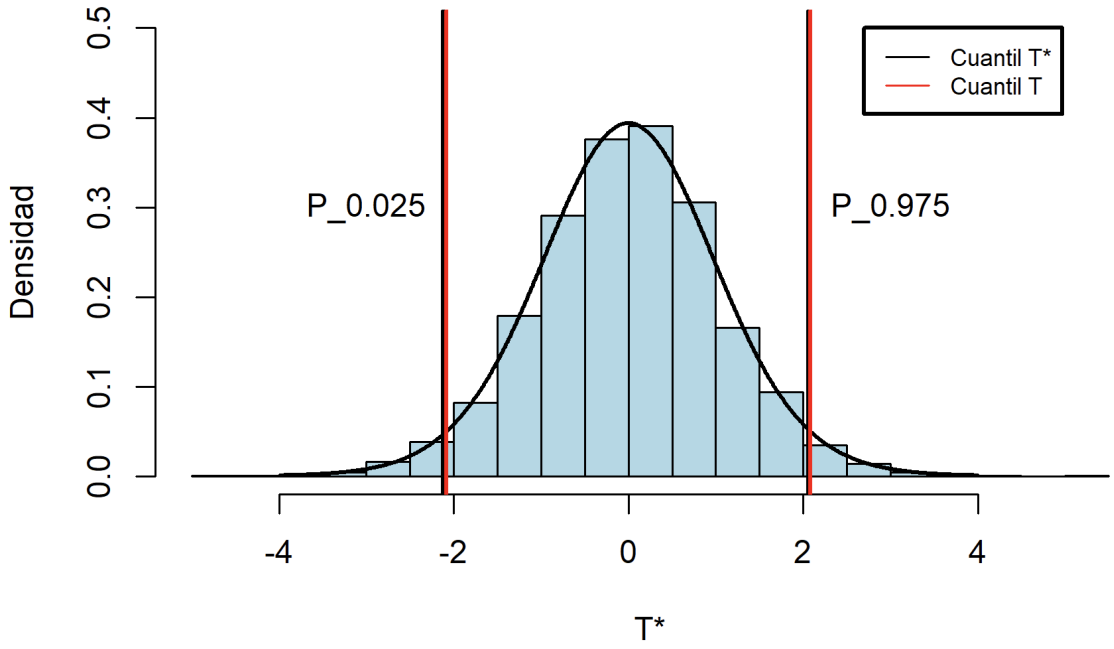

Figura 3.10: Histograma de T^*.

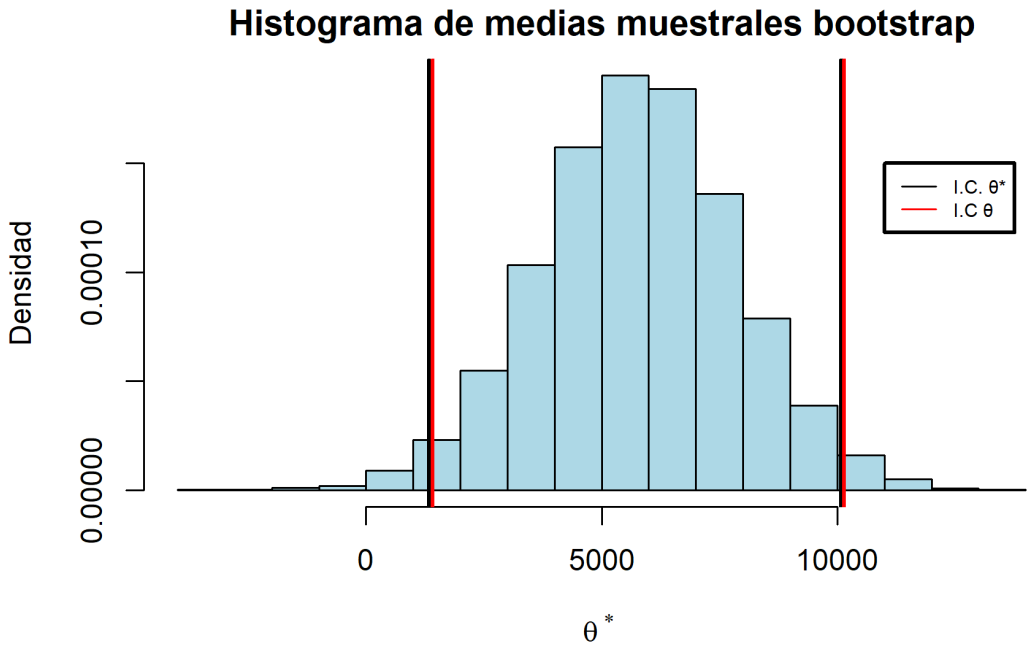

Figura 3.11: Histograma de θ^*.

Observemos que los cuantiles para la distribución $t-$Student $(-2.08, 2.08)$ son ligeramente diferentes de los cuantiles del bootstrap no paramétrico $(-2.36, 1.87)$, pero más parecidos al obtenido con el bootstrap paramétrico $(-2.12, 2.05)$; ésto último puede observarse en la Figura 3.10. Por otra parte, en la Figura 3.11 puede observarse que el intervalo bootstrap obtenido de forma paramétrica es muy parecido al obtenido mediante (3.7).

Resumiendo, los extremos del intervalo de confianza obtenido bajo el supuesto de normalidad en las muestras, varianzas no-homogéneas y el uso de cuantiles en la distribución $t-$Student resulta parecidos: $(1407.26, 10146.07)$. Cuando se utiliza un bootstrap no-paramétrico, los extremos resultan: $(834.59, 9694.39)$, mientras que el bootstrap paramétrico y para bootstrap paramétrico $(1334.89, 10084.84)$.

En caso de querer realizar una prueba de hipótesis del tipo

$$H_0 : \mu_1 - \mu_2 = 0,$$
$$H_1 : \mu_1 - \mu_2 \neq 0$$

podemos observar simplemente los extremos de los intervalos obtenidos, bajo los diferentes procedimientos, y al constatar que ninguno de ellos contiene el valor cero, tenemos evidencia para rechazar la hipótesis nula, a un nivel de significancia $\alpha = 0.05$.

El código para construir estos intervalos, así como las gráficas presentadas en este ejemplo, se desarrolló en R y se encuentra disponible en el Apéndice B.6.

3.2.2. Intervalo de confianza y prueba de hipótesis para dos medias, varianzas desconocidas supuestas homogéneas

En el caso que se tengan muestras aleatorias independientes $\mathbf{X_1}$, $\mathbf{X_2}$, se cumpla el supuesto de normalidad en ambas muestras y también el supuesto de homogeneidad de varianzas, es posible construir un intervalo de confianza a partir de la siguiente cantidad pivotal

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2(\frac{1}{n_1} + \frac{1}{n_2})}}, \tag{3.18}$$

la cual sigue una distribución $t-$Student con $r = n_1 + n_2 - 2$ grados de libertad y donde

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}, \tag{3.19}$$

es un estimador conjunto de ambas varianzas poblacionales σ_1^2 y σ_2^2.

De esta forma, para encontrar un intervalo al $(1-\alpha)100\,\%$ de confianza para $\mu_1 - \mu_2$ partimos de:

$$P\left(-t_{1-\alpha/2,r} \leq \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2(\frac{1}{n_1} + \frac{1}{n_2})}} \leq t_{1-\alpha/2,r}\right) = 1 - \alpha, \tag{3.20}$$

y con un poco de álgebra llegamos al intervalo del $(1-\alpha)100\,\%$ de confianza para $\mu_1 - \mu_2$:

$$P\left((\bar{X}_1 - \bar{X}_2) - t_{1-\alpha/2,r}\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \leq \mu_1 - \mu_2 \leq (\bar{X}_1 - \bar{X}_2) + t_{1-\alpha/2,r}\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right) = 1 - \alpha, \tag{3.21}$$

cuyos extremos podemos escribir como:

$$(\bar{X}_1 - \bar{X}_2) \pm t_{1-\alpha/2,r}\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}. \tag{3.22}$$

Para construir el intervalo de confianza bootstrap para la diferencia de medias se calcula la distribución muestral de T^*:

$$T^* = \frac{(\bar{X}_1^* - \bar{X}_2^*) - (\bar{X}_1 - \bar{X}_2)}{\sqrt{S_p^{*2}(\frac{1}{n_1} + \frac{1}{n_2})}}, \tag{3.23}$$

donde un estimador bootstrap para las varianzas σ_1^2 y σ_2^2 sería el siguiente:

$$S_p^{*2} = \frac{(n_1 - 1)S_1^{*2} + (n_2 - 1)S_2^{*2}}{n_1 + n_2 - 2}. \tag{3.24}$$

Para encontrar el intervalo bootstrap del $100(1-\alpha)\,\%$ de confianza para $\mu_1 - \mu_2$ calculamos:

$$P\left((\bar{X}_1 - \bar{X}_2) + t^*_{\alpha/2}\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \leq \mu_1 - \mu_2 \leq (\bar{X}_1 - \bar{X}_2) + t^*_{1-\alpha/2}\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\right) = 1-\alpha, \tag{3.25}$$

donde $t^*_{\alpha/2,r}$ y $t^*_{1-\alpha/2,r}$ son cuantiles de probabilidad $\alpha/2$ y $1-\alpha/2$, respectivamente, obtenidos de la distribución de T^*.

Para el caso en que se desee realizar una prueba de hipótesis para probar $H_0 : \mu_1 - \mu_2 = \delta_0$, se seguirá el procedimiento descrito en la sección anterior, donde simplemente se verifica si el valor δ_0 se encuentra o no dentro del intervalo de confianza obtenido.

Con el siguiente ejemplo, tomado de [7, p. 252], se ejemplifica lo tratado en esta sección.

Ejemplo 3.4 *Un grupo de catorce hombres se dividió aleatoriamente para efectuar un experimento donde se desea determinar cuál de los dos fármacos produce un mayor aumento de la presión arterial, utilizando un intervalo del* $95\,\%$ *de confianza. El Fármaco 1 se administró a siete de los hombres elegidos al azar y el Fármaco 2 se administró a los siete restantes. Los aumentos observados en la presión arterial son:*
Fármaco: 1: 0.7, -0.2, 3.4, 3.7, 0.8, 0.0, 2.0.
Fármaco: 2: 1.9, 1.1, 4.4, 5.5, 1.6, 4.6, 3.4.

Si consideramos que la variable X_1 es una variable aleatoria que representa los aumentos de presión arterial cuando se utiliza el fármaco 1 y X_2 lo correspondiente para el fármaco 2, para poder calcular un intervalo al $(1-\alpha)100\,\%$ de confianza para $\mu_1 - \mu_2$, utilizando (3.21), debemos probar la normalidad para cada una de las muestras y efectuar una prueba para verificar también la homogeneidad de varianzas.

Primeramente se presenta un resumen con medidas descriptivas de las muestras $\mathbf{x_1}$ y $\mathbf{x_2}$ y los respectivos diagramas de caja para los datos obtenidos al usar los fármacos 1 y 2, respectivamente. En la Figura 3.12 puede observarse que no existen valores atípicos ni aberrantes, por lo que el supuesto de normalidad pudiera cumplirse.

```
   Min.   1st Qu.  Median    Mean   3rd Qu.    Max.
-0.200   0.350    0.800    1.486   2.700     3.700

   Min.   1st Qu.  Median    Mean    3rd Qu.    Max.
  1.100   1.750    3.400     3.214   4.500     5.500
```

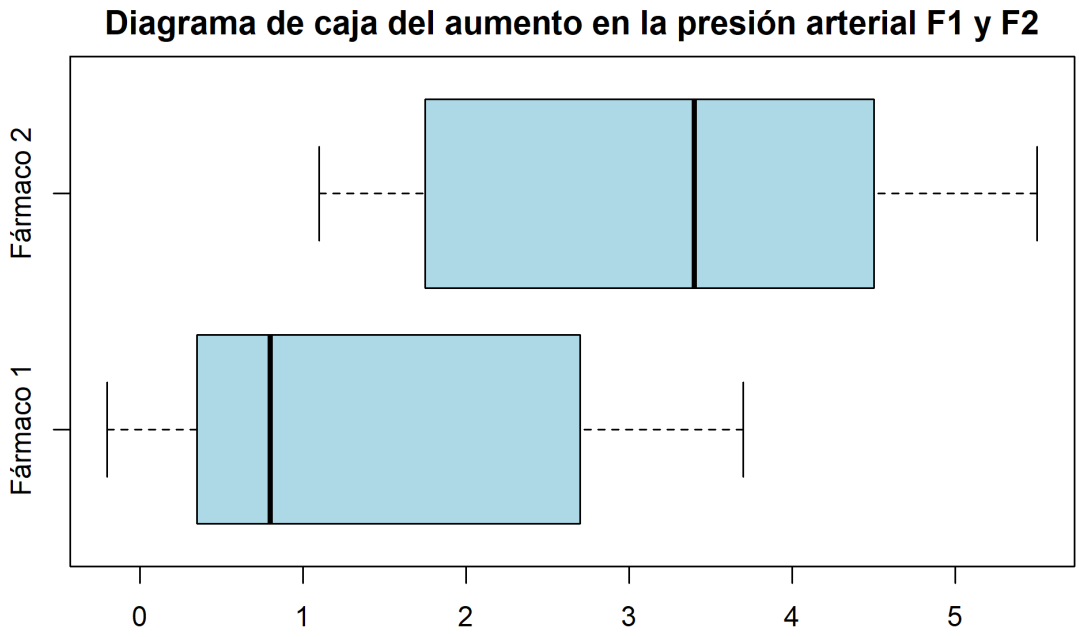

Figura 3.12: Diagrama de caja del aumento en la presión arterial de F_1 *y* F_2.

Con respecto al supuesto de normalidad, a continuación se muestran los resultados que arrojan la prueba de Shapiro-Wilk para los datos de mediciones con los fármacos 1 y 2, respectivamente, donde

$$H_0 : \mathbf{X_1} \sim N(\mu_1, \sigma_1^2),$$
$$H_1 : \mathbf{X_1} \nsim N(\mu_1, \sigma_1^2),$$

```
Shapiro-Wilk normality test
data:  F1
W = 0.88704, p-value = 0.2596
```

$$H_0 : \mathbf{X_2} \sim N(\mu_2, \sigma_2^2),$$
$$H_1 : \mathbf{X_2} \nsim N(\mu_2, \sigma_2^2),$$

```
Shapiro-Wilk normality test
data:  F2
W = 0.92178, p-value = 0.4833
```

Puede observarse que en ambos casos los p-valores son superiores a un nivel de significancia de 0.05, por lo que no tenemos evidencia para rechazar las hipótesis nulas de normalidad en ambos casos.

Nuevamente, para verificar el supuesto de homogeneidad de varianzas se utiliza lo expuesto en el Apéndice A.2.3, para lo cual se plantean las siguientes hipótesis:

$$H_0 : \sigma_1^2 = \sigma_2^2,$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2.$$

Sustituyendo las varianzas muestrales en (A.4), tenemos,

$$f_c = \frac{2.5}{2.9} = 0.86. \tag{3.26}$$

Si comparamos este resultado con el cuantil de probabilidad $\alpha/2 = 0.025$, en una distribución F_6^6 tenemos que:

$$f_{n_2-1,\alpha/2}^{n_1-1} = f_{6,0.025}^{6} = 0.17,$$

donde podemos observar que $f_c = 0.86 > f_{n_2-1,\alpha/2}^{n_1-1} = 0.17$, por lo que no tenemos evidencia para rechazar H_0 y entonces podemos suponer que las varianzas poblacionales son homogéneas.

Dado que se satisfacen los supuestos que requiere el utilizar la fórmula para un intervalo de confianza de diferencia de medias presentado en (3.22), se calcula ahora la estimación puntual de las varianzas utilizando (3.19), como se muestra a continuación:

$$S_p^2 = \frac{(6)2.5 + (6)2.9}{12} = 2.7. \tag{3.27}$$

Como el valor del cuantil de probabilidad $1 - \alpha/2 = 0.975$ en una distribución $t-$Student con 12 grados de libertad es 2.17, se procede a calcular un intervalo de confianza 3.21,

$$(1.48 - 3.21) \pm 2.17\sqrt{(2.7)(0.14 + 0.14)}. \tag{3.28}$$

De esta manera, un intervalo al 95 % de confianza para $\mu_1 - \mu_2$ resulta:

$$(-3.64, 0.18).$$

Como puede observarse, el valor cero se encuentra dentro del intervalo de confianza, lo que da evidencia a favor de que las medias poblacionales son homogéneas.

Si se utiliza ahora el procedimiento bootstrap no-paramétrico para obtener un intervalo de confianza bootstrap para $\mu_1 - \mu_2$, se calcularán primeramente los cuantiles de probabilidad $\alpha/2 = 0.025$ y $1 - \alpha/2 = 0.975$ de la distribución de T^*, mostrada en (3.23). Se utiliza el mismo procedimiento que se describe en la sección anterior, donde se simulan $10,000$ muestras de tamaños n_1 y n_2 de las muestras originales, todo ello utilizando el software R con el fin de construir la distribución de T^*, sobre el cual se muestran algunas medidas descriptivas, al igual que para $\theta^* = \bar{X_1}^* - \bar{X_2}^*$:

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-8.81969 -0.69972 -0.03151 -0.03772  0.64956  6.67909
```

y de $\theta^* = \bar{X_1}^* - \bar{X_2}^*$.

```
    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -4.643  -2.286  -1.757  -1.736  -1.186   1.286
```

Ahora si podemos calcular el intervalo bootstrap al 95 % de confianza:

$$\left((\bar{X}_1 - \bar{X}_2) + t^*_{0.025}\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}, (\bar{X}_1 - \bar{X}_2) + t^*_{0.975}\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\right), \tag{3.29}$$

sustituyendo las estimaciones puntuales respectivas y dado que $t^*_0.025 = -2.27$ y $t^*_0.975 = 2.11$,

$$\left((1.48 - 3.21) - 2.27\sqrt{(2.7)(0.14+0.14)}, (1.48 - 3.21) + 2.11\sqrt{(2.7)(0.14+0.14)}\right),$$

de esta manera el intervalo bootstrap resulta:

$$(-3.72, 0.12).$$

Los extremos de este intervalo bootstrap contienen al cero, por lo que se concluye que hay evidencia a favor de que las medias poblacionales son homogéneas.

Por otra parte, nótese que los valores de los cuantiles de probabilidad 0.025 y 0.975 que se obtienen tanto de una distribución $t-$Student, como los obtenidos a partir de la distribución muestral de T^*, no son tan diferentes, como se muestra en la figura 3.13.

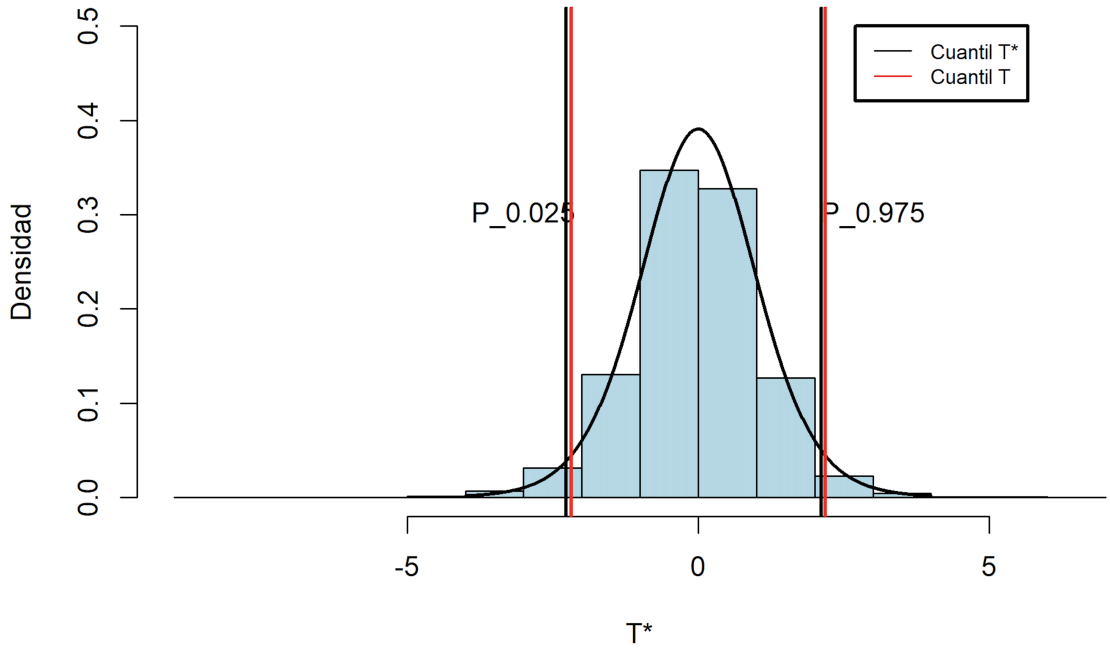

Figura 3.13: Histograma de T^*, bootstrap no-paramétrico.

En la Figura 3.16 donde se muestra la distribución de $\theta^* = \bar{x}_1^* - \bar{x}_2^*$ pueden notarse pequeñas diferencias en los intervalos obtenidos mediante estos dos procedimientos.

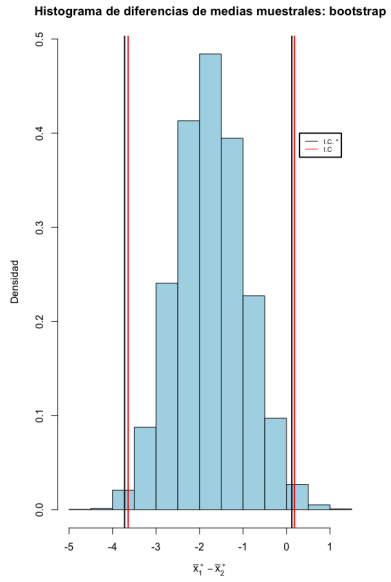

Figura 3.14: Histograma de diferencia de medias muestrales: bootstrap no-paramétrico.

Tal como se explicó en la sección anterior, se utilizará un bootstrap paramétrico para construir un intervalo para la diferencia de medias $\mu_1 - \mu_2$. La construcción de la distribución de T^* se realiza a través de la generación de muestras de distribuciones parametrizadas.

A continuación se muestra un resumen de algunas medidas para T^* y $\bar{X}^*$ respectivamente:

```
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-4.499795  -0.674882  0.008561   0.007709  0.688460   6.083004
```

```
   Min.  1st Qu.  Median    Mean  3rd Qu.    Max.
 -4.863  -2.304  -1.721   -1.723  -1.136   1.571
```

Ahora si podemos calcular el intervalo de confianza bootstrap, partiendo de (3.29) y considerando $t^*_{0.025} = -2.09$ y $t^*_{0.975} = -2.11$, valores ligeramente diferentes a los obtenidos con la distribución $t-$Student donde $t_{0.975,12} = 2.17$. Calculamos entonces el intervalo bootstrap al 95 % de confianza para $\mu_1 - \mu_2$,

$$\left((1.48 - 3.21) - 2.09\sqrt{(2.7)(0.14 + 0.14)}, (1.48 - 3.21) + 2.11\sqrt{(2.7)(0.14 + 0.14)}\right),$$

y éste resulta:

$$(-3.56, 0.12).$$

En la Figura 3.16 puede observarse la distribución de $\theta^* = \bar{x}_1^* - \bar{x}_2^*$, donde es posible comparar el intervalo que se obtiene utilizando la distribución $t-$Student y el obtenido con un bootstrap paramétrico.

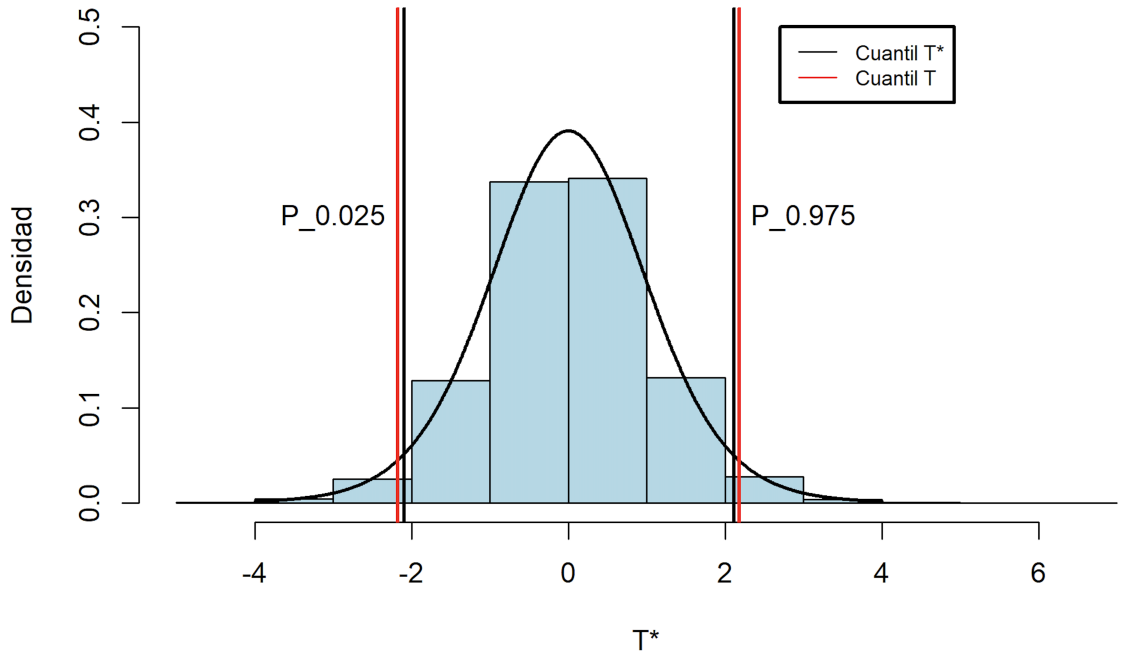

Figura 3.15: Histograma de T^*.

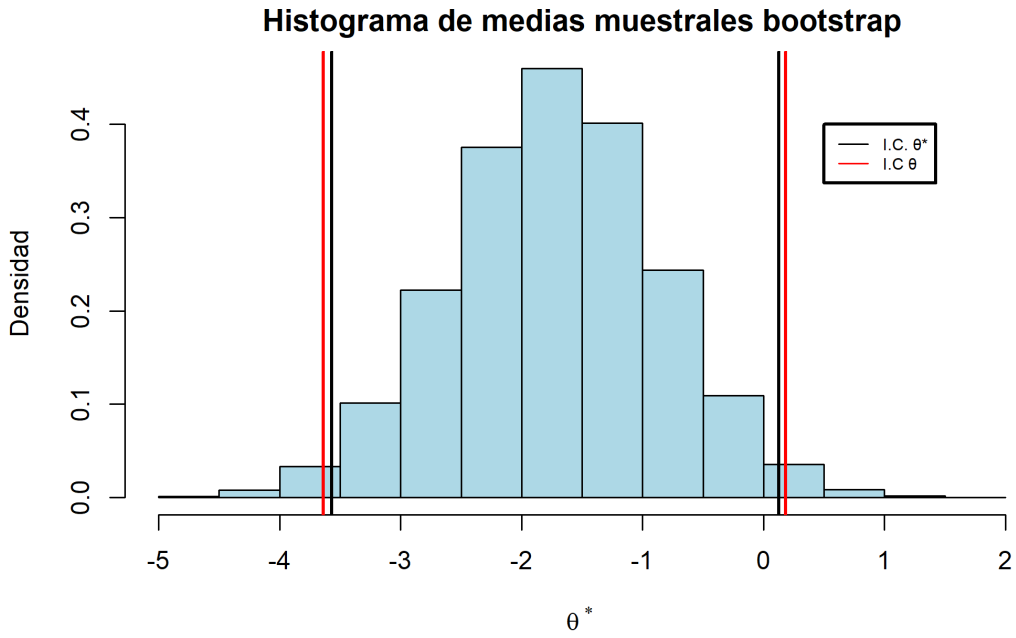

Figura 3.16: Histograma de diferencia de medias muestrales: bootstrap paramétrico.

De igual manera a lo planteado en la sección anterior, si se desea probar una hipótesis de la

forma:

$$H_0 : \mu_1 - \mu_2 = 0,$$
$$H_1 : \mu_1 - \mu_2 \neq 0$$

basta con verificar que el valor de cero está incluido en todos los intervalos construidos en esta comparación de fármacos, por lo que no habría evidencia para rechazar la hipótesis nula.

El código en R para calcular los diferentes intervalos de confianza se encuentra en el Apéndice B.7.

3.3. Otro caso de estudio

Aunque lo que se pretende en esta tesis es mostrar una introducción al uso de bootstrap en los cursos de estadística básica y se han retomado los temas relativos a parámetros de localización, que se cubren en este curso, esta sección tiene por objetivo mostrar que podemos construir intervalos de confianza para algún otro parámetro de interés, aunque no corresponda a material que se cubra en dicho curso.

A continuación se muestra una aplicación con la distribución exponencial, donde la variable aleatoria mide el tiempo que transcurre hasta que ocurra un determinado evento. La distribución exponencial es un caso especial de la distribución Gamma, como se presenta a continuación.

3.3.1. Intervalo de confianza para el parámetro λ: caso exponencial

Aunque no es un material que se cubra en los cursos de estadística, el intervalo de confianza que se calculará en este ejemplo es posible obtenerlo analíticamente, como se muestra a continuación. Presentar resultados analíticos para este caso y algunos otros de interés, en cursos de estadística, no es posible por el tiempo de duración del curso y los temas incluidos en éste. Sin embargo, aunque aquí se desarrolla de manera analítica como obtener un intervalo de confianza para el parámetro λ de una distribución exponencial, en un curso de estadística pudieran presentarse diversos ejemplos que se aborden simplemente a través de bootstrap.

A continuación se recuerda la función de densidad de una variable aleatoria que sigue una distribución Gamma, así como la de una con distribución exponencial.

Definición 3.1 (Distribución Gamma) *Si una variable aleatoria X tiene densidad dada por*

$$f_X(x; r, \lambda) = \frac{\lambda}{\Gamma(r)}(\lambda x)^{r-1} e^{-\lambda x} I_{[0,\infty]}(x) \tag{3.30}$$

donde $r > 0$ y $\lambda > 0$, entonces se dice que X tiene una distribución gamma. (Véase en [9, P.112].

Definición 3.2 (Función gamma) *La función gamma, denotada por* $\Gamma(\cdot)$*, esta definida como*

$$\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}dx \quad para \quad t > 0. \tag{3.31}$$

(Véase en [9, P.534]

Los siguientes teoremas y definiciones fueron tomados de [9, P.112]

Teorema 3.1 *Si* X *tiene distribución gamma con parámetros* r *y* λ*, entonces*

$$E[X] = \frac{r}{\lambda}, \quad Var[X] = \frac{r}{\lambda^2}, \quad M_X(t) = \left(\frac{\lambda}{\lambda - t}\right)^r, \quad para \quad t < \lambda. \tag{3.32}$$

Cuando $r = 1$ en la densidad gamma, se obtiene la densidad exponencial.

Definición 3.3 (Distribución Exponencial) *Si* X *es una variable aleatoria que tiene una densidad dada por*

$$f_X(x, \lambda) = \lambda e^{-\lambda x} I_{[0,\infty]}(x), \tag{3.33}$$

donde $\lambda > 0$*, entonces se dice que* X *tiene una distribución exponencial.*

Observación Si $\{X_1, X_2, \ldots, X_n\}$ es una muestra aleatoria de una distribución exponencial con parámetro λ, su suma $T = \sum_{i=1}^n X_i$ sigue una distribución *gamma* $G(\lambda, n)$

Teorema 3.2 *Si* X *tiene distribución exponencial, con parámetros* $r = 1$ *y* λ*, entonces*

$$E[X] = \frac{1}{\lambda}, \quad Var[X] = \frac{1}{\lambda^2}, \quad M_X(t) = \frac{\lambda}{\lambda - t}, \quad para \quad t < \lambda. \tag{3.34}$$

Es posible construir un intervalo al $(1 - \alpha)100\,\%$ de confianza para el parámetro λ de una exponencial, basado en la relación entre la distribución exponencial, la distribución Gamma y la distribución Ji-cuadrada. Consideremos la variable $V = 2\lambda T = 2\lambda \sum_{i=1}^n X_i = 2\lambda n\bar{X}$, como se ha obtenido a partir de T a través de un simple cambio de escala, V seguirá también una distribución gamma donde:

$$E[V] = 2\lambda E[T] = 2\lambda r\frac{1}{\lambda} = 2r, \tag{3.35}$$

$$Var[V] = 4\lambda^2 Var[T] = 4\lambda^2 r \cdot \frac{1}{\lambda^2} = 4r. \tag{3.36}$$

Para construir el intervalo de confianza se utiliza la siguiente cantidad pivotal

$$V = 2\lambda n\bar{x}, \tag{3.37}$$

que sigue una distribución $G(2,n) = \chi^2_{2n}$, con $2n$ grados de libertad. A partir de la distribución χ^2_{2n} se calculan los percentiles $\chi^2_{2n,1-\alpha/2}$ y $\chi^2_{2n,\alpha/2}$ de forma que:

$$P\left(\chi^2_{2n,\alpha/2} \leq V \leq \chi^2_{2n,1-\alpha/2}\right) = 1-\alpha. \tag{3.38}$$

Por tanto:

$$P\left(\chi^2_{2n,\alpha/2} \leq 2n\lambda\bar{x} \leq \chi^2_{2n,1-\alpha/2}\right) = 1-\alpha, \tag{3.39}$$

dividiendo todos los términos del interior del intervalo por $2n\bar{X}$ y luego tomando el recíproco de λ se obtiene:

$$P\left(\frac{2n\bar{x}}{\chi^2_{2n,1-\alpha/2}} \leq \frac{1}{\lambda} \leq \frac{2n\bar{x}}{\chi^2_{2n,\alpha/2}}\right) = 1-\alpha. \tag{3.40}$$

De esta forma el intervalo de confianza a nivel $1-\alpha$ para el parámetro λ de una distribución exponencial es:

$$\left(\frac{2n\bar{x}}{\chi^2_{2n,1-\alpha/2}}, \frac{2n\bar{x}}{\chi^2_{2n,\alpha/2}}\right). \tag{3.41}$$

Como el estimador de λ es $\hat{\lambda} = \frac{1}{\bar{x}}$. Para construir el intervalo de confianza bootstrap se construirá la distribución de:

$$V^* = 2\frac{1}{\bar{x}}n\bar{x}^*, \tag{3.42}$$

con el fin de obtener cuantiles aproximados a los de una distribución χ^2_{2n} y calcular un intervalo de confianza bootstrap para λ, de la siguiente manera:

$$\left(\frac{2n\bar{x}}{\chi^{2*}_{1-\alpha/2}}, \frac{2n\bar{x}}{\chi^{2*}_{\alpha/2}}\right). \tag{3.43}$$

Ahora veamos un ejemplo donde se trabaja una variable que se afirma sigue una distribución exponencial; este ejemplo se ha tomado de [7, p.174].

Ejemplo 3.5 *Los tiempos de supervivencia de los pacientes tratados por una determinada enfermedad se distribuyen exponencialmente. Con el tratamiento estándar, la supervivencia esperada es de 37.4 meses. Diez pacientes que recibieron un nuevo tratamiento sobrevivieron durante los siguientes tiempos (en meses):*

99, 60, 8, 44, 30, 6, 12, 105, 53, 17.

Se desea estimar un intervalo de confianza al 95 % para λ.

Calculamos $\bar{x} = 43.4$ y $\hat{\lambda} = 1/\bar{x} = 0.023$.
Utilizando (3.41) calculamos el intervalo al 95 % de confianza para λ, como

$$\left(\frac{\chi^2_{20,0.025}}{20(43.4)}, \frac{\chi^2_{20,0.975}}{20(43.4)}\right) = \left(\frac{9.59868}{868}, \frac{34.16}{868}\right) = (0.011, 0.039)\,. \tag{3.44}$$

Ahora, generaremos en el software R, 1000 muestras bootstrap usando la expresión (3.42), para así obtener los cuantiles de probabilidad $\alpha/2$ y $1-\alpha/2$, considerando $1-\alpha = 0.95$ y construir el intervalo de confianza bootstrap no-paramétrico al 95 % de confianza para λ.

Veamos un resumen de V^* y $1/\hat{\lambda}^* = \bar{x}^*$ respectivamente.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  6.774  16.221  19.585  19.930  23.341  36.129

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  14.70   35.20   42.50   43.25   50.65   78.40
```

El intervalo bootstrap no-paramétrico al 95 % de confianza resulta:

$$\left(\frac{\chi^{2*}_{0.975}}{20(43.4)}, \frac{\chi^{2*}_{0.025}}{20(43.4)}\right) = \left(\frac{10.73}{868}, \frac{30.92}{868}\right) = (0.012, 0.036)\,. \tag{3.45}$$

Ahora encontraremos el intervalo bootstrap al 95 % de confianza utilizando bootstrap paramétrico. Las muestras se simulan de una distribución exponencial con parámetro estimado por $\hat{\lambda} = 1/\bar{x}$. Una vez obtenidas las muestras se trabaja nuevamente con la expresión (3.42) para obtener los cuantiles de la distribución de V^*.

Veamos un resumen de V^* y $1/\hat{\lambda}^* = 1/\bar{x}^*$ respectivamente.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.714  15.021  19.355  19.967  23.994  47.430

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.23   32.59   42.00   43.33   52.07  102.92
```

El intervalo de confianza bootstrap paramétrico al 95 % es:

$$\left(\frac{\chi^{2*}_{0.975}}{20(43.4)}, \frac{\chi^{2*}_{0.025}}{20(43.4)}\right) = \left(\frac{9.88}{868}, \frac{34.39}{868}\right) = (0.011, 0.040)\,. \tag{3.46}$$

Nótese que existe similitud en el intervalo obtenido cuando se usa la distribución ji-cuadrada, así como los que resultan cuando se simula la variable V^*, que permite calcular los cuantiles que se utilizarán en los intervalos de confianza bootstrap.
El código con el que se obtuvieron los intervalos anteriores se encuentra en el Apéndice B.8

Conclusiones

Entre las tareas primordiales de un investigador en estadística está la de resumir una muestra, originada a través de cierto estudio, y generalizar los hallazgos a la población de donde se tomó la muestra; ésto por medio de algún procedimiento científico.

Por lo general, cuando se trabaja con un estadístico o estimador de interés, se quisiera tener una idea de todos sus valores posibles, presentados éstos en forma de una distribución de probabilidad, conocida como distribución muestral del estadístico. El desarrollo matemático de las distribuciones muestrales de ciertos estadísticos, a veces no es claro para algunos estudiantes, y se complica cuando las herramientas necesarias para conocer dicha distribución no se han cubierto en cursos previos. De ahí la importancia de simular ciertas distribuciones muestrales. Y aunque a partir de éstas pueden calcularse percentiles y tener una idea de un intervalo de confianza, en este trabajo se utilizó básicamente el procedimiento conocido como bootstrap-t, para la construcción de estos intervalos, ya que de acuerdo con Hesterberg [5], el intervalo de percentiles es mucho más angosto cuando la distribución es simétrica; algo equivalente a utilizar $Z_{1-\alpha/2}\sigma/\sqrt{n}$, en lugar de usar $t_{1-\alpha/2}s/\sqrt{n}$. Recordemos que la distribución $t-$Student se acerca a la distribución normal estándar conforme aumentan los grados de libertad.

Como pudo verse a través de esta tesis, no es difícil construir los intervalos bootstrap-t, tanto de manera paramétrica como no-paramétrica. Los ejemplos desarrollados y la comparación que se hizo entre los diversos procedimientos para calcular intervalos de confianza, pueden brindar al estudiante herramientas de gran utilidad que puede aplicar en situaciones donde no se tenga una fórmula explícita para un intervalo de confianza o bien, cuando no se satisfagan los supuestos requeridos de un cierto procedimiento. Sin embargo, se debe considerar que existen diversas propuestas de bootstrap, dependiendo del parámetro de interés y de las características que se observen en la distribución del estadístico bajo estudio. En bibliografía como [4], [5] y [2] pueden consultarse alternativas que se adecuen al parámetro que se desea estimar.

Por otra parte, además de los ejemplos que muestran temáticas que se cubren en un curso de estadística básico, como motivación se incluyó un ejemplo de estimación del parámetro λ en una distribución exponencial, ello con la finalidad de que un estudiante de un curso básico de estadística pueda dimensionar el panorama de utilidad de un procedimiento como es bootstrap.

Finalmente, es importante mencionar que para correr los códigos que se elaboraron para la

construcción de los intervalos mostrados en los diversos ejemplos, mismos que se incluyen en el Apéndice B, no se requiere de computadoras con grandes capacidades para realizar las simulaciones efectuadas, éstas por lo general toman menos de un minuto y es posible utilizar una gran cantidad de software que hoy en día está disponible de manera gratuita. En particular, aquí se desarrolló todo en el software R, el cual también tiene librerías para bootstrap en caso que se quiera recurrir a éstas.

Apéndice A

Material adicional

A.1. Teorema de límite central

Definición A.1 *Sea $X_1, X_2, \ldots$ una sucesión de variables aleatorias con $F_1, F_2, \ldots$ la sucesión correspondiente de función de distribución acumulada. $\{X_n\}$ converge a X si*

$$\lim_{n\to\infty} F_n(x) = F_X(x),$$

para toda x donde F_X es continua. Lo denotaremos por $X_n \xrightarrow{d} X$.

Teorema A.1 (Teorema del límite central) *Sea $X_1, X_2, X_3, \ldots$ una sucesión de variables aleatorias iid de una distribución que tiene una función generadora de momentos definida en un intervalo que contiene al 0. Entonces*

$$\lim_{n\to\infty} P\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq z\right) = \Phi(z),$$

donde $\bar{X}n = \frac{X_1+\ldots+X_n}{n}$.En otras palabras,

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1).$$

Demostración.

Definamos algunas variables:

$$Z_i = \frac{X_i - \mu}{\sigma},$$
$$W_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{Z_1 + \ldots + Z_n}{\sqrt{n}} = \frac{Z_1}{\sqrt{n}} + \ldots + \frac{Z_n}{\sqrt{n}}.$$

Además

$$\begin{aligned} E(\bar{X}n) = \mu, \quad & Var(\bar{X}n) = \sigma/\sqrt{n}, \\ E(Z_i) = 0, \quad & Var(Z_i) = 1, \\ E(W_n = 0, \quad & Var(W_n = 1. \end{aligned}$$

Sea $M(t)$ una función generadora de momentos de las $Z_i's$. Entonces

$$M_{W_n}(t) = \left[M\left(\frac{t}{\sqrt{n}} \right) \right]^n.$$

Como $E(Z_i) = 0$ y $Var(Z_i) = 1$, sabemos que $M(0) = 1$, $M'(0) = 0$ y $M''(0) = 1$.
Por el teorema de Taylor, para cada valor de t existe un numero real $r \in (-t, t)$ tal que,

$$\begin{aligned} M(t) &= M(0) + M'(0)t + \frac{1}{2}M''(r)t^t \\ &= 1 + \frac{t^2}{2}M''(r). \end{aligned}$$

Si aplicamos esto a $\frac{t}{\sqrt{n}}$ en lugar de t, vemos que para $s \in (-\frac{t}{\sqrt{n}}, \frac{t}{\sqrt{n}})$

$$M\left(\frac{t}{\sqrt{n}} \right) = 1 + \frac{t^2}{2n}M''(s),$$

y

$$M_{W_n}(t) = \left[1 + \frac{t^2}{2n}M''(s) \right]^n = \left[1 + \frac{\frac{t^2}{2}M''(s)}{n} \right]^n.$$

Ahora como $n \to \infty$, $s \to 0$ ya que $s \in (-\frac{t}{\sqrt{n}}, \frac{t}{\sqrt{n}})$, y como $M''(s) \to M'(0) = 1$ entonces M'' es continua. M tiene derivadas de todos los órdenes, por lo que todos son continuos. Además,

$$\lim_{n\to\infty} \left[1 + \frac{a_n}{n} \right]^n = e^a,$$

si $a_n \to a$ como $n \to \infty$. Entonces

$$\lim_{n\to\infty} M_{W_n}(t) = e^{t^2/2}.$$

Pero esta es la fgm de una distribución normal estándar,

$$\lim_{n\to\infty} F_{W_n}(w) = \Phi(w).$$

■

En la mayoría de las técnicas inferenciales que se enseñan en los cursos de estadística, se requiere el supuesto de normalidad. Aunque existen diversas pruebas para verificar este supuesto, en este apéndice se muestra un procedimiento gráfico, que es la construcción de una gráfica cuantil-cuantil y la prueba estadística de Shapiro-Wilk, misma que se utiliza en los diversos ejemplos que se incluyen en esta tesis.

A.2. Verificación de supuesto de normalidad

A.2.1. Gráfico cuantil-cuantil

La gráfica cuantil-cuantil o Q-Q Plot nos permite comparar una distribución de un conjunto de datos con la distribución teórica. La construcción de estas gráficas se realizan utilizando los datos ordenados y los cuantiles de la distribución a comparar.
Supongamos que queremos probar si X_i, $(i = 1, 2, \ldots, n)$ podría haber provenido razonablemente de una distribución específica con $F(x)$ su función de distribución.
Se construye el gráfico de probabilidad de la siguiente manera.

1. Ordenar las observaciones de menor a mayor

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}.$$

2. Determinar los valores

$$p_i = \frac{i - 0.5}{n}, \quad i = 1, 2, \ldots, n.$$

   Sea $Q_x(p)$ el cuantil de orden $p(0 < p < 1)$ de las observaciones, entonces

$$x_i = Q_x(P_i) \quad i = 1, 2, \ldots, n.$$

3. Establecer los cuantiles de orden $p_i, i = 1, 2, \ldots, n$ de la distribución teórica F, es decir,

$$Q_t(p_i) = F^{-1}(p_i) \quad i = 1, 2, \ldots, n.$$

4. Representar el conjunto de puntos $(Q_t(p_i)), (Q_x(p_i)), \quad i = 1, 2, \ldots, n$, es decir, los puntos $(F^{-1}(p_i), x_{(i)}), \quad i = 1, 2, \ldots, n.$

Si la distribución teórica tiene una buena aproximación de la distribución empírica, entonces los cuantiles de los datos estarán muy próximos a los de la distribución teórica y, por tanto, los puntos de la gráfica estarán cercanos a la recta y = x.

A.2.2. Prueba de normalidad de Shapiro-Wilk

Aunque en la sección anterior se mostró un procedimiento para decidir si los datos podrían provenir de de una distribución normal, en ocasiones no es claro un procedimiento gráfico y se efectúan pruebas de hipótesis sobre normalidad. Una de ellas es la prueba de Shapiro-Wilk

[11] que es una prueba de bondad de ajuste.

Para realizar la prueba Shapiro-Wilk supone contar con una muestra compuesta de n observaciones independientes e idénticamente distribuidas $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, provenientes de una distribución normal con media μ y varianza σ^2. En la hipótesis nula se plantea que los datos provienen de una distribución normal y la hipótesis alternativa sostiene que la distribución no es normal, es decir:

$$H_0 : \mathbf{X} \sim N(\mu, \sigma^2),$$
$$H_1 : \mathbf{X} \nsim N(\mu, \sigma^2).$$

Para calcular el estadístico de prueba de Shapiro-Wilk:

- Primeramente se ordenan las n observaciones de menor a mayor, obteniendo: $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, donde $x_{(1)}$ representa el dato más pequeño y $x_{(n)}$ el dato más grande.
- Se calcula $b = a_1(x_{(n)} - x_{(1)} + a_2(x_{(n-1)} - x_{(2)}$,..., donde $a_1, a_2, ...$ son coeficientes que se obtienen de tablas como la mostrada en Figura A.1.
- Se calcula
$$W = \frac{b^2}{(n-1)S^2}, \tag{A.1}$$
donde S^2 es la varianza muestral.
- Finalmente se compara el valor que tomó el estadístico de prueba con un valor crítico de la tabla de valores propuesta por Shapiro y Wilk, valores que denotaremos por $Tab\ W$ y que se muestran la figura A.2. Si el valor del estadístico de prueba es mayor que $Tab\ W$, no se tendrá evidencia para rechazar la hipótesis nula.

La siguiente tabla fue obtenida de [8, p. 239].

$n =$	2	3	4	5	6	7	8	9	10	11	12	13
a_1	0.7071	0.7071	0.6872	0.6646	0.6431	0.6233	0.6052	0.5888	0.5739	0.5601	0.5475	0.5359
a_2		0.0	0.1677	0.2143	0.2806	0.3031	0.3164	0.3244	0.3291	0.3315	0.3325	0.3325
a_3				0.0	0.0875	0.1401	0.1743	0.1976	0.2141	0.226	0.2347	0.2412
a_4						0.0	0.0561	0.0947	0.1224	0.1429	0.1586	0.1707
a_5								0.0	0.0399	0.0695	0.0922	0.1099
a_6										0.0	0.0303	0.0539
a_7												0.0

$n =$	14	15	16	17	18	19	20	21	22	23	24	25
a_1	0.5251	0.515	0.5056	0.4968	0.4886	0.4808	0.4734	0.4643	0.459	0.4542	0.4493	0.445
a_2	0.3318	0.3306	0.329	0.3273	0.3253	0.3232	0.3211	0.3185	0.3156	0.3126	0.3098	0.3069
a_3	0.246	0.2495	0.2521	0.254	0.2553	0.2561	0.2565	0.2578	0.2571	0.2563	0.2554	0.2543
a_4	0.1802	0.1878	0.1939	0.1988	0.2027	0.2059	0.2085	0.2119	0.2131	0.2139	0.2145	0.2148
a_5	0.124	0.1353	0.1447	0.1524	0.1587	0.1641	0.1686	0.1736	0.1764	0.1787	0.1807	0.1822
a_6	0.0727	0.088	0.1005	0.1109	0.1197	0.1271	0.1334	0.1399	0.1443	0.148	0.1512	0.1539
a_7	0.024	0.0433	0.0593	0.0725	0.0837	0.0932	0.1013	0.1092	0.115	0.1201	0.1245	0.1283
a_8		0.0	0.0196	0.0359	0.0496	0.0612	0.0711	0.0804	0.0878	0.0941	0.0997	0.1046
a_9				0.0	0.0163	0.0303	0.0422	0.053	0.0618	0.0696	0.0764	0.0823
a_{10}						0.0	0.014	0.0263	0.0368	0.0459	0.0539	0.061
a_{11}								0.0	0.0122	0.0228	0.0321	0.0403
a_{12}										0.0	0.0107	0.02
a_{13}												0.0

Figura A.1: Coeficientes a_i para la prueba de normalidad de Shapiro-Wilk (para $n \leq 25$)

La siguiente tabla fue obtenida de [8, p. 240].

Level of significance				
n	0.01	0.02	0.05	0.1
3	0.753	0.756	0.767	0.789
4	0.687	0.707	0.748	0.792
5	0.686	0.715	0.762	0.806
6	0.713	0.743	0.788	0.826
7	0.73	0.76	0.803	0.838
8	0.749	0.778	0.818	0.851
9	0.764	0.791	0.829	0.859
10	0.781	0.806	0.842	0.869
11	0.792	0.817	0.85	0.876
12	0.805	0.828	0.859	0.883
13	0.814	0.837	0.866	0.889
14	0.825	0.846	0.874	0.895
15	0.835	0.855	0.881	0.901
16	0.844	0.863	0.887	0.906
17	0.851	0.869	0.892	0.91
18	0.858	0.874	0.897	0.914
19	0.863	0.879	0.901	0.917
20	0.868	0.884	0.905	0.92
21	0.873	0.888	0.908	0.923
22	0.878	0.892	0.911	0.926
23	0.881	0.895	0.914	0.928
24	0.884	0.898	0.916	0.93
25	0.888	0.901	0.918	0.931

Figura A.2: Valores críticos de prueba de Shapiro-Wilk de W.

Con el fin de mostrar cómo efectuar esta prueba, se retoma un ejemplo propuesto por King

and Eckersley [8, p. 156-167].

Ejemplo A.1 *Los siguientes datos son los niveles de contaminante del fármaco "Wonder 2": 20, 20, 21, 26, 43, 43, 54, 54, 55. Se desea efectuar una prueba de normalidad utilizando el estadístico de Shapiro-Wilk.*

Considerando que X son los niveles del nivel de contaminante en el fármaco "Wonder 2", se plantea:

$$H_0 : \mathbf{X} \sim N(\mu, \sigma^2)$$
$$H_1 : \mathbf{X} \nsim N(\mu, \sigma^2),$$

y para calcular el estadístico de prueba mostrado en A.1

- Se utilizan lo coeficientes mostrados en la Figura A.1, considerando $n = 9$. De esta manera calculamos $b = 0.5888(55-20)+0.3244(54-20)+0.1976(54-21)+0.0947(43-26) = 39.7683$
- Se calcula el estadístico de prueba: $W = \frac{39.7683^2}{(9-1)15.52^2} = 0.8203$.
- Se obtiene el valor critico $Tab\ W$ de los valores mostrados en la Figura A.2, para $n = 9$, resultando $Tab\ W = 0.829$.
- Como $W < Tab\ W$, es decir, $0.8203 < 0.829$, se rechaza la hipótesis nula para un nivel de signficancia $\alpha = 0.05$, por lo cual se tiene evidencia en contra del supuesto de normalidad para los niveles de contaminante en el fármaco “Wonder 2”.

En el desarrollo de los ejemplos incluidos en esta tesis, se efectúan diversas pruebas de normalidad utilizando la prueba Shapiro-Wilk, y ello se efectúa con el software R. Por ello, utilizando estos datos de niveles de contaminante, se muestra cómo se realiza esta prueba en este software.

```
x <- c( 20, 20, 21, 26, 43, 43, 54, 54, 55)
shapiro.test(x)
```

```
  Shapiro-Wilk normality test

 data:  x
 W = 0.82098, p-value = 0.03535
```

Como puede verse, R proporciona el valor del estadístico y el $p - valor$ asociado a éste, el cual, considerando nuevamente un nivel de significancia de 0.05, se tiene evidencia para rechazar el supuesto de normalidad que se planteó en la hipótesis nula.

A.2.3. Prueba de hipótesis para dos varianzas

Sean dos muestras independientes,

$$X : x_1, x_2, \ldots, x_{n1} \quad tal \quad que \quad X \sim N(\mu_1, \sigma_1^2),$$
$$Y : y_1, y_2, \ldots, y_{n2} \quad tal \quad que \quad Y \sim N(\mu_2, \sigma_2^2)$$

Suponemos que μ_1 y μ_2 son desconocidas. La inferencia sobre σ_1^2 y σ_2^2 están basadas en las estadísticas $V_1 = \sum(X_i - \bar{X})^2$ y $V_2 = \sum(Y_i - \bar{Y})^2$.

Estas son variables independientes con

$$U_1 \equiv V_1/\sigma_1^2 \sim \chi^2_{(n_1-1)}; \quad U_2 \equiv V_2/\sigma_2^2 \sim \chi^2_{(n_2-1)}.$$

Las varianzas estimadas son $S_1^2 \equiv \frac{1}{n_1-1}V_1$ y $S_2^2 \equiv \frac{1}{n_2-1}V_2$. La razón U_1/U_2 o cualquier función de esta relación, tendrá una distribución que depende sólo de la relación $\lambda = \sigma_1^2/\sigma_2^2$. Formaremos el estadístico para un cociente de varianzas, al que llamaremos ahora F_c, donde,

$$F_c = \frac{U_1/(n-1)}{U_2/(n_2-2)} = \frac{\sigma_2^2}{\sigma_1^2} \cdot \frac{V_1/(n_1-1)}{V_2/(n_2-1)} = \frac{\sigma_2^2}{\sigma_1^2} \cdot \frac{S_1^2}{S_2^2}. \tag{A.2}$$

el cual tiene una distribución F con $n_1 - 1$ grados de libertad en el numerador y $n_2 - 1$ grados de libertad en el denominador, es decir,

$$f_c = \frac{\sigma_2^2}{\sigma_1^2} \cdot \frac{S_1^2}{S_2^2} \sim F_{n_1-1,n_2-1} \tag{A.3}$$

Si en la hipótesis nula consideramos varianzas homogéneas, es decir, $\sigma_1^2 = \sigma_2^2$, entonces el estadistico de prueba es:

$$f_c = \frac{S_1^2}{S_2^2} \tag{A.4}$$

En el cuadro siguiente se muestran las hipótesis nula y alternativa que podemos plantear para comparar las varianzas poblacionales.

Caso 1	Caso 2	Caso 3
$H_0 : \sigma_1^2 \leq \sigma_2^2$	$H_0 : \sigma_1^2 \geq \sigma_2^2$	$H_0 : \sigma_1^2 = \sigma_2^2$
$H_1 : \sigma_1^2 > \sigma_2^2$	$H_1 : \sigma_1^2 < \sigma_2^2$	$H_1 : \sigma_1^2 \neq \sigma_2^2$

Se rechazará H_0 cuando:

- Caso 1: $f_c > f^{n_1-1}_{n_2-1,1-\alpha}$
- Caso 2:$f_c < f^{n_1-1}_{n_2-1,\alpha}$
- Caso 3: $f_c > f^{n_1-1}_{n_2-1,\alpha/2} \quad o \quad f_c > f^{n_1-1}_{n_2-1,1-\alpha/2}$

Apéndice B

Códigos en R

En este apéndice se muestran los *R Markdown* que se hicieron para realizar figuras y desarrollar los ejemplos antes vistos.

B.1. Código en *R* para Figura 1.1

```
ye <- seq(-4,4,0.01)
denN <- dnorm(ye)
k <- c(10,20,30)
dent <- dt(ye,30)

M <- array(NA, dim = c(3, length(ye)))
colnames(M) = ye

for(i in 1:3){
  M[i, ] = dt(ye,k[i])
}

plot(ye, denN, main = "Ajuste t-student a una normal estandar",
     col = "red", type = "l")
lines(ye, M[1,], col = "blue")
lines(ye, M[2,], col = "green")
lines(ye, M[3,], col = "orange")

legend("right", c("Normal","n = 10","n = 20","n = 30"),
       col = c("red", "blue", "green", "orange"), lty = 1, bty = "n")
```

B.2. Código en *R* para Figura 1.2

.

```
n <- 10
p <- 0.5
Bin <- pbinom(6, n, p)
Bin
```

```
## [1] 0.828125
```

Sin corrección por continuidad

```
Norm <- pnorm(6, n*p, sqrt(n*p*(1-p)))
Norm
```

```
## [1] 0.7364554
```

Con corrección por continuidad

```
NormC <- pnorm(6.5, n*p, sqrt(n*p*(1-p)))
NormC
```

```
## [1] 0.8286091
```

```
XNorm <- seq(0, 10, 0.01)
XBinom <- seq(0, 8, 1)

Ynorm <- dnorm(XNorm, n*p, sqrt(n*p*(1-p)))
Ybinom<- dbinom(XBinom, n, p)

plot(XNorm, Ynorm, type = "l", main = "Normal-Binomial (n = 10, p = 0.5)",
     xlab = "X", ylab = "Densidad de X", col = "firebrick")
lines(XBinom, Ybinom, type = "h", col = "royalblue")
lines(XBinom, Ybinom, type = "s", col = "royalblue")
abline(v = 6, col = "darkgoldenrod1")
abline(v = 6.5, col = "darkgreen")



legend("topleft", c("Con corrección", "Sin corrección"),
       col = c("darkgreen", "darkgoldenrod1"), lty = 1, bty = "n")
```

B.3. Código en R para Figura 1.3

```
require(graphics)

k <- 1000
n <- c(10,50,100)
p <- c(0.1, 0.3, 0.5)
```

```r
par(mfrow = c(3,3))

for (i in 1:3) {
  for (j in 1:3) {

    y = rbinom(k,n[i], p[j])
    qqnorm(y)
    qqline(y, col = "red", lwd = 1)
  }
}
```

B.4. Código en R para Ejemplo 3.1

.

```r
#install.packages("latex2exp")
library("latex2exp")
```

```r
x <- c(55.3, 54.8, 65.9, 60.7, 59.4, 62.0, 62.1,
       58.7, 64.5, 62.3,67.6, 61.2)
shapiro.test(x) #Prueba de normalidad

summary(x) #Resumen de x

boxplot(x, horizontal = TRUE, col = "lightblue",
        main = "Diagrama de caja del aumento de peso en gramos")
```

```r
n <- length(x)
xbar <- mean(x)
Se <- sd(x)/sqrt(n)
```

```r
#Cuantiles para la distribucion t-student
cuantil_t <- qt(c(0.025,0.975), n-1)

#Intervalo de confianza t-student
IC_t <- xbar + qt(c(0.025,0.975), n-1) * Se
```

```r
#Bootstrap no paramétrico

theta_star <- NULL
tstar <- NULL

set.seed(1)
for (i in 1:10000){

  R_boots <- sample(x, n, replace = TRUE) #remuestra
```

```
  xbar_star <- mean(R_boots)
  s_star <- sd(R_boots)
  theta_star[i] <- xbar_star
  tstar[i] <- (xbar_star - xbar)/(s_star/sqrt(n)) #cantidad pivotal
}

summary(tstar)

summary(theta_star)
```

```
#Cuantiles del percentil de t-bootstrap
set.seed(1)
cuantil_np <- quantile(tstar, probs = c(0.025,0.975))

#Intervalo de confianza t-bootstrap
IC_np <- xbar +  quantile(tstar, probs = c(0.025,0.975)) * Se

#Histograma de T*

hist(tstar, col = "lightblue", freq = FALSE, ylim = c(0,0.5),
     main = "Histograma de la cantidad pivotal  T*",
     ylab = "Densidad", xlab = "T*")
text(-3, 0.3,"P_0.025", col = "black")
text(3, 0.3, "P_0.975", col = "black")
ejex <- seq(-4,4,0.01)
points(ejex, dt(ejex, n-1), lwd = 2, lty = 1, type = "l")
abline(v = cuantil_np[1],  lwd = 2, col = "black")
abline(v = cuantil_np[2],  lwd = 2, col = "black")
abline(v = cuantil_t[1],  lwd = 2, col = "red")
abline(v = cuantil_t[2],  lwd = 2, col = "red")

legend(x = 2.7, y = 0.5, c("Cuantil T*", "Cuantil T"),
       col = c("black", "red"), lty = 1, bty = "o",
       box.lty = 1, box.lwd = 2, cex = 0.75)
```

```
#Histograma de *

hist(theta_star, col = "lightblue", freq = FALSE,
     main = "Histograma de medias muestrales bootstrap",
     xlab = TeX(r'($\theta^{*}$)'),  ylab  = "Densidad")
text(-2.5, 0.2,"q1", col ="black")
text(2.5, 0.2,"q3", col ="black")
abline(v = IC_np[1],  lwd = 2, col = "black")
abline(v = IC_np[2],  lwd = 2, col = "black")
abline(v = IC_t[1],  lwd = 2, col = "red")
```

```r
abline(v = IC_t[2],  lwd = 2, col = "red")

legend(x = 64, y = 0.3, inset = 0.05, c("I.C. T*", "I.C T"),
       col = c("black", "red"), lty = 1, bty = "o",
       box.lty = 1, box.lwd = 2, cex = 0.65)
```

```r
#Bootstrap paramétrico

theta_starP <- NULL
tstarP <- NULL

set.seed(1)
for (i in 1:10000){

  R_bootsP <- rnorm(n, xbar, sd(x)) #nueva remuestra
  xbar_starP <- mean(R_bootsP)
  s_starP <- sd(R_bootsP)/sqrt(n)
  theta_starP[i] <- xbar_starP
  tstarP[i] <- (xbar_starP - xbar)/s_starP #cantidad pivotal
}

summary(tstarP)

summary(theta_starP)
```

```r
#Cuantiles del percentil de t-bootstrap
set.seed(1)
cuantil_p <- quantile(tstar, probs = c(0.025,0.975))
cuantil_p
```

```r
#Intervalo de confianza t-bootstrap
IC_p <- xbar +  quantile(tstar, probs = c(0.025,0.975)) * Se
IC_p
```

```r
#Histograma de T*

hist(tstarP, col = "lightblue", freq = FALSE, ylim = c(0,.5),
     main = "Histograma de la cantidad pivotal  T*",
     xlab = "T*", ylab = "Densidad")
text(-3, 0.3,"P_0.025", col = "black")
text(3, 0.3, "P_0.975", col = "black")
ejex <- seq(-4,4,0.01)
points(ejex, dt(ejex, n-1), lwd = 2, lty = 1, type = "l")
abline(v = cuantil_p[1],  lwd = 2, col = "black")
abline(v = cuantil_p[2],  lwd = 2, col = "black")
abline(v = cuantil_t[1],  lwd = 2, col = "red")
```

```
abline(v = cuantil_t[2],  lwd = 2, col = "red")

legend(x = 2.7, y = 0.5, c("Cuantil T*", "Cuantil T"),
       col = c("black", "red"), lty = 1, bty = "o",
       box.lty = 1, box.lwd = 2, cex = 0.75)
```

```
#Histograma de *

hist(theta_starP, col = "lightblue", freq = FALSE,
     main = "Histograma de medias muestrales bootstrap",
     xlab = TeX(r'($\theta^{*}$)'), ylab  = "Densidad")
text(-2.5, 0.2,"q1",col="black")
text(2.5, 0.2,"q3",col="black")
abline(v = IC_p[1],  lwd = 2, col = "black")
abline(v = IC_p[2],  lwd = 2, col = "black")
abline(v = IC_t[1],  lwd = 2, col = "red")
abline(v = IC_t[2],  lwd = 2, col = "red")

legend(x = 64, y = 0.3, inset = 0.05, c("I.C. T*", "I.C T"),
       col = c("black", "red"), lty = 1, bty = "o",
       box.lty = 1, box.lwd = 2, cex = 0.65)
```

B.5. Código en R para Ejemplo 3.2

```
x <- c(4.94, 5.06, 4.53, 5.07, 4.99, 5.16, 4.38, 4.43, 4.93, 4.72, 4.92, 4.96)
shapiro.test(x) #Prueba de normalidad
```

```
summary(x) #Resumen de x
```

```
boxplot(x, horizontal = TRUE, col = "lightblue",
        main = "Diagrama de caja de residuos de aflatoxinas")
```

```
xbar <- mean(x)
```

```
t.test(x, y = NULL,
       conf.level = 0.95)
```

```
n <- length(x)
xbar <- mean(x)
Se <- sd(x)/sqrt(n)
```

```
#Bootstrap no paramétrico

theta_star <- NULL
tstar <- NULL
```

```r
set.seed(1)
for (i in 1:10000){

  R_boots <- sample(x, n, replace = TRUE) #remuestra
  xbar_star <- mean(R_boots)
  s_star <- sd(R_boots)
  theta_star[i] <- xbar_star
  tstar[i] <- (xbar_star - xbar)/(s_star/sqrt(n)) #cantidad pivotal
}
```

```r
summary(tstar)
summary(theta_star)
```

```r
#Intervalo de confianza t-bootstrap
IC_p <- xbar +  quantile(tstar, probs = c(0.025,0.975)) * Se
IC_p
```

```r
#Bootstrap paramétrico

theta_starP <- NULL
tstarP <- NULL

set.seed(1)
for (i in 1:10000){

  R_bootsP <- rnorm(n, xbar, sd(x)) #nueva remuestra
  xbar_starP <- mean(R_bootsP)
  s_starP <- sd(R_bootsP)/sqrt(n)
  theta_starP[i] <- xbar_starP
  tstarP[i] <- (xbar_starP - xbar)/s_starP #cantidad pivotal
}
```

```r
summary(tstarP)
summary(theta_starP)
```

```r
#Cuantiles del percentil de t-bootstrap
set.seed(1)
cuantil_p <- quantile(tstarP, probs = c(0.025,0.975))
cuantil_p
```

```r
#Intervalo de confianza t-bootstrap
IC_p <- xbar +  quantile(tstarP, probs = c(0.025,0.975)) * Se
IC_p
```

B.6. Código en R para Ejemplo 3.3

```
library("latex2exp")
```

```
E1 <- c(5030, 13700, 10730, 11400, 860, 2200, 4250, 15040, 4980, 11910,
        8130, 26850, 17660, 22800, 1130, 1690)
E2 <- c(2800, 4670, 6890, 7720, 7030, 7330, 2810, 1330, 3320, 1230,
        2130, 2190)
```

```
#Prueba de normalidad
shapiro.test(E1)
shapiro.test(E2)
summary(E1) #Resumen de x
summary(E2)
boxplot(E1, E2, horizontal = T, col = "lightblue",
        main = "Diagramas de caja de densidad de organismos ",
        xlab = "Densidad",
        names = c("Estación 1", "Estación 2"))
```

```
n1 <- length(E1)
n2 <- length(E2)
xbar1 <- mean(E1)
xbar2 <- mean(E2)
```

```
#Prueba de Hipotesis para cociente de varianzas
fc<- sd(E1)^2/sd(E2)^2
q1 <- qf(0.025, n1-1, n2-1)
q2 <- qf(0.975, n1-1, n2-1)
pf <- pf(fc, n1-1, n2-1)
var.test(E1, E2, alternative = "two.sided",
         conf.level = 0.95)
```

```
  F test to compare two variances

 data:  E1 and E2
 F = 10.086, num df = 15, denom df = 11, p-value = 0.000452
 alternative hypothesis: true ratio of variances is not equal to 1
 95 percent confidence interval:
   3.028741 30.335474
 sample estimates:
 ratio of variances
           10.08551
```

```
Se <- sqrt((sd(E1)^2/n1 + sd(E2)^2/n2))
Tc <- (xbar1-xbar2)/Se #Estadistico de prueba
```

```r
#grados de libertad
v <- round((((sd(E1)^2/n1)+(sd(E2)^2/n2))^2)/(((sd(E1)^2/n1)^2/n1-1) +
((sd(E2)^2/n2)^2/n2-1)))
```

```r
#Cuantiles para la distribucion t-student
t <- qt(c(0.025,0.975), v)
```

```r
Se <- sqrt((sd(E1)^2/n1 + sd(E2)^2/n2))
```

```r
#Intervalo de confianza t-student
IC_t <- (xbar1 - xbar2) + t * Se
```

```r
#Bootstrap no paramétrico

theta_star <- NULL
Tstar <- NULL

set.seed(1)
for (i in 1:10000){

  E1_np <- sample(E1, n1, replace = TRUE) #remuestra de E_1
  E2_np <- sample(E2, n2, replace = TRUE) #remuestra de E_2

  xbar_star1 <- mean(E1_np)
  xbar_star2 <- mean(E2_np)

  theta_star[i] <- xbar_star1 - xbar_star2

  Tstar[i] <- ((xbar_star1 - xbar_star2) - (xbar1 - xbar2))/sqrt((sd(E1_np)^2/n1
              + sd(E2_np)^2/n2)) #cantidad pivotal
}
```

```r
summary(Tstar)
summary(theta_star)
```

```r
#Cuantiles del percentil de t-bootstrap
set.seed(1)
tstar_np <- quantile(Tstar, probs = c(0.025, 0.975))
```

```r
#Intervalo de confianza t-bootstrap
IC_np <- (xbar1 - xbar2) +  tstar_np * Se
```

```r
#Histograma de T*

hist(Tstar, col = "lightblue", freq = FALSE, ylim = c(0,0.5),
     main = "Histograma de la cantidad pivotal  T*",
     ylab = "Densidad", xlab = "T*")
```

```
text(-3.5, 0.3,"P_0.025", col = "black")
text(3, 0.3, "P_0.975", col = "black")
ejex <- seq(-4,4,0.01)
points(ejex, dt(ejex, v), lwd = 2, lty = 1, type = "l")
abline(v = tstar_np[1],  lwd = 2, col = "black")
abline(v = tstar_np[2],  lwd = 2, col = "black")
abline(v = t[1],  lwd = 2, col = "red")
abline(v = t[2],  lwd = 2, col = "red")

legend(x = -6, y = 0.5, c("Cuantil T*", "Cuantil T"),
       col = c("black", "red"), lty = 1, bty = "o",
       box.lty = 1, box.lwd = 2, cex = 0.75)
```

```
#Histograma de *

hist(theta_star, col = "lightblue", freq = FALSE,
     main = "Histograma de diferencia de medias: bootstrap",
     xlab = TeX(r'($\heta^{*}$)'), ylab  = "Densidad")
text(-2.5, 0.2,"q1", col ="black")
text(2.5, 0.2,"q3", col ="black")
abline(v = IC_np[1],  lwd = 2, col = "black")
abline(v = IC_np[2],  lwd = 2, col = "black")
abline(v = IC_t[1],  lwd = 2, col = "red")
abline(v = IC_t[2],  lwd = 2, col = "red")

legend(x = 11000, y = 0.00020, inset = 0.05, c("I.C. *", "I.C "),
       col = c("black", "red"), lty = 1, bty = "o",
       box.lty = 1, box.lwd = 2, cex = 0.65)
```

```
#Bootstrap paramétrico

theta_starP <- NULL
TstarP <- NULL

set.seed(1)
for (i in 1:10000){

  E1_p <- rnorm(n1, xbar1, sd(E1)) #nueva remuestra
  E2_p <- rnorm(n2, xbar2, sd(E2)) #nueva remuestra

  xbar_starP1 <- mean(E1_p)
  xbar_starP2 <- mean(E2_p)


  Sp_star2P <- ((n1-1) * sd(E1_p)^2 + (n2-1) * sd(E2_p)^2)/v
```

```
  theta_starP[i] <- xbar_starP1 - xbar_starP2

  TstarP[i] <- ((xbar_starP1 - xbar_starP2) - (xbar1 - xbar2))/
               sqrt((sd(E1_p)^2/n1 + sd(E2_p)^2/n2)) #cantidad pivotal
}
```

```
summary(TstarP)
summary(theta_starP)
```

```
#Cuantiles del percentil de t-bootstrap
set.seed(1)
tstar_P <- quantile(TstarP, probs = c(0.025,0.975))
```

```
#Intervalo de confianza t-bootstrap
IC_p <- (xbar1 - xbar2) +  tstar_P * Se
```

```
#Histograma de T*

hist(TstarP, col = "lightblue", freq = FALSE, ylim = c(0,.5),
     main = "Histograma de la cantidad pivotal  T*",
     xlab = "T*", ylab = "Densidad")
text(-3, 0.3,"P_0.025", col = "black")
text(3, 0.3, "P_0.975", col = "black")
ejex <- seq(-4,4,0.01)
points(ejex, dt(ejex, v), lwd = 2, lty = 1, type = "l")
abline(v = tstar_P[1],  lwd = 2, col = "black")
abline(v = tstar_P[2],  lwd = 2, col = "black")
abline(v = t[1],  lwd = 2, col = "red")
abline(v = t[2],  lwd = 2, col = "red")

legend(x = 2.7, y = 0.5, c("Cuantil T*", "Cuantil T"),
       col = c("black", "red"), lty = 1, bty = "o",
       box.lty = 1, box.lwd = 2, cex = 0.75)
```

```
#Histograma de *

hist(theta_starP, col = "lightblue", freq = FALSE,
     main = "Histograma de medias muestrales bootstrap",
     xlab = TeX(r'($\theta^{*}$)'), ylab  = "Densidad")
text(-2.5, 0.2,"q1",col="black")
text(2.5, 0.2,"q3",col="black")
abline(v = IC_p[1],  lwd = 2, col = "black")
abline(v = IC_p[2],  lwd = 2, col = "black")
abline(v = IC_t[1],  lwd = 2, col = "red")
abline(v = IC_t[2],  lwd = 2, col = "red")
```

```
legend(x = 11000, y = 0.00015, inset = 0.05, c("I.C. *", "I.C "),
       col = c("black", "red"), lty = 1, bty = "o",
       box.lty = 1, box.lwd = 2, cex = 0.65)
```

B.7. Código en R para Ejemplo 3.4

```
library("latex2exp")
```

```
F1 <- c(0.7, -0.2, 3.4, 3.7, 0.8, 0.0, 2.0)
F2 <- c(1.9, 1.1, 4.4, 5.5, 1.6, 4.6, 3.4)
```

```
shapiro.test(F1) #Prueba de normalidad
shapiro.test(F2) #Prueba de normalidad
```

```
summary(F1) #Resumen de F1
summary(F2) #Resumen de F2
```

```
boxplot(F1,F2, horizontal = TRUE, col = "lightblue",
        main = "Diagrama de caja del aumento en la presión arterial F1 y F2",
        names = c("Fármaco 1","Fármaco 2"))
```

```
n1 <- length(F1) #Muestra del fármaco 1
n2 <- length(F2) #Muestra del fármaco 2
xbar1 <- mean(F1)
xbar2 <- mean(F2)
```

```
#Prueba de hipotesis
fc<- sd(F1)^2/sd(F2)^2
fc
pf(fc,6,6)
```

```
q1 <- qf(0.025,6,6)
q2 <- qf(0.975,6,6)
var.test(F1, F2, alternative="two.sided",
         conf.level=0.95)
```

```
  F test to compare two variances

 data:  F1 and F2
 F = 0.86083, num df = 6, denom df = 6, p-value = 0.8603
 alternative hypothesis: true ratio of variances is not equal to 1
 95 percent confidence interval:
  0.1479148 5.0098102
 sample estimates:
```

```
 ratio of variances
          0.8608281
```

```r
#grados de libertad
r <- n1 + n2 -2
```

```r
Sp2 <- ((n1-1)*sd(F1)^2 + (n2-1)*sd(F2)^2)/r
Se <- sqrt(Sp2 * (1/n1 + 1/n2))
Tc <- (xbar2-xbar1)/Se #Estadistico de prueba
q1 <- qt(0.025,r)
q2 <- qt(0.975,r)
```

```r
#Cuantiles para la distribucion t-student
t <- qt(c(0.025,0.975), r)

Se <- sqrt(Sp2 * (1/n1 + 1/n2))

#Intervalo de confianza t-student
IC_t <- (xbar1 - xbar2) + t * Se
```

```r
#Bootstrap no paramétrico

theta_star <- NULL
Tstar <- NULL

set.seed(1)
for (i in 1:10000){

  F1_np <- sample(F1, n1, replace = TRUE) #remuestra de F_1
  F2_np <- sample(F2, n2, replace = TRUE) #remuestra de F_2

  xbar_star1 <- mean(F1_np)
  xbar_star2 <- mean(F2_np)

  Sp_star2 <- ((n1-1) * sd(F1_np)^2 + (n2-1) * sd(F2_np)^2)/r

  theta_star[i] <- xbar_star1 - xbar_star2

  Tstar[i] <- ((xbar_star1 - xbar_star2) - (xbar1 - xbar2))/sqrt(Sp_star2 *
              (1/n1 + 1/n2)) #cantidad pivotal
}
```

```r
summary(Tstar)
summary(theta_star)
```

```r
#Cuantiles del percentil de t-bootstrap
set.seed(1)
```

```r
tstar_np <- quantile(Tstar, probs = c(0.025,0.975))
#Intervalo de confianza t-bootstrap
IC_np <- (xbar1 - xbar2) +  tstar_np * Se
```

```r
#Histograma de T*

hist(Tstar, col = "lightblue", freq = FALSE, ylim = c(0,0.5),
     main = "Histograma de la cantidad pivotal  T*",
     ylab = "Densidad", xlab = "T*")
text(-3, 0.3,"P_0.025", col = "black")
text(3, 0.3, "P_0.975", col = "black")
ejex <- seq(-4,4,0.01)
points(ejex, dt(ejex, r), lwd = 2, lty = 1, type = "l")
abline(v = tstar_np[1],  lwd = 2, col = "black")
abline(v = tstar_np[2],  lwd = 2, col = "black")
abline(v = t[1],  lwd = 2, col = "red")
abline(v = t[2],  lwd = 2, col = "red")

legend(x = 2.7, y = 0.5, c("Cuantil T*", "Cuantil T"),
       col = c("black", "red"), lty = 1, bty = "o",
       box.lty = 1, box.lwd = 2, cex = 0.75)
```

```r
#Histograma de *

hist(theta_star, col = "lightblue", freq = FALSE,
     main = "Histograma de medias muestrales bootstrap",
     xlab = TeX(r'($\theta^{*}$)'),  ylab  = "Densidad")
abline(v = IC_np[1],  lwd = 2, col = "black")
abline(v = IC_np[2],  lwd = 2, col = "black")
abline(v = IC_t[1],  lwd = 2, col = "red")
abline(v = IC_t[2],  lwd = 2, col = "red")

legend(x = 64, y = 0.3, inset = 0.05, c("I.C. *", "I.C "),
       col = c("black", "red"), lty = 1, bty = "o",
       box.lty = 1, box.lwd = 2, cex = 0.65)
```

```r
#Bootstrap paramétrico

theta_starP <- NULL
TstarP <- NULL

set.seed(1)
for (i in 1:10000){

  F1_p <- rnorm(n1, xbar1, sd(F1)) #nueva remuestra
```

```
  F2_p <- rnorm(n2, xbar2, sd(F2)) #nueva remuestra

  xbar_starP1 <- mean(F1_p)
  xbar_starP2 <- mean(F2_p)


  Sp_star2P <- ((n1-1) * sd(F1_p)^2 + (n2-1) * sd(F2_p)^2)/r

  theta_starP[i] <- xbar_starP1 - xbar_starP2

  TstarP[i] <- ((xbar_starP1 - xbar_starP2) - (xbar1 - xbar2))/sqrt(Sp_star2P
              * (1/n1 + 1/n2)) #cantidad pivotal
}
```

```
summary(TstarP)
summary(theta_starP)
```

```
#Cuantiles del percentil de t-bootstrap
set.seed(1)
tstar_P <- quantile(TstarP, probs = c(0.025,0.975))

#Intervalo de confianza t-bootstrap
IC_p <- (xbar1 - xbar2) +  tstar_P * Se
```

```
#Histograma de T*

hist(TstarP, col = "lightblue", freq = FALSE, ylim = c(0,.5),
     main = "Histograma de la cantidad pivotal  T*",
     xlab = "T*", ylab = "Densidad")
text(-3, 0.3,"P_0.025", col = "black")
text(3, 0.3, "P_0.975", col = "black")
ejex <- seq(-4,4,0.01)
points(ejex, dt(ejex, r), lwd = 2, lty = 1, type = "l")
abline(v = tstar_P[1],  lwd = 2, col = "black")
abline(v = tstar_P[2],  lwd = 2, col = "black")
abline(v = t[1],  lwd = 2, col = "red")
abline(v = t[2],  lwd = 2, col = "red")

legend(x = 2.7, y = 0.5, c("Cuantil T*", "Cuantil T"),
       col = c("black", "red"), lty = 1, bty = "o",
       box.lty = 1, box.lwd = 2, cex = 0.75)
```

```
#Histograma de *

hist(theta_starP, col = "lightblue", freq = FALSE,
```

```
     main = "Histograma de medias muestrales bootstrap",
     xlab = TeX(r'($\theta^{*}$)'), ylab  = "Densidad")
abline(v = IC_p[1],  lwd = 2, col = "black")
abline(v = IC_p[2],  lwd = 2, col = "black")
abline(v = IC_t[1],  lwd = 2, col = "red")
abline(v = IC_t[2],  lwd = 2, col = "red")

legend(x = 0.5, y = 0.4, inset = 0.05, c("I.C. *", "I.C "),
       col = c("black", "red"), lty = 1, bty = "o",
       box.lty = 1, box.lwd = 2, cex = 0.65)
```

B.8. Código en R para Ejemplo 3.5

```
x <- c(99, 60, 8, 44, 30, 6, 12, 105, 53, 17)
n <- length(x)
xbar <- mean(x)
summary(x)
```

```
#Intervalo de confianza
IC <-(qchisq(c(0.025, 0.975), 2*n))/ (xbar*2*n)
```

```
#Estimación de máxima verosimilitud lambda
lambdahat <- 1/xbar
```

```
#Generar muestras bootstrap
VstarNP<- NULL
recip_lambdahatstarNP <- NULL

set.seed(1)
for (i in 1:1000){

  bootstrapsampleNP <- sample(x,n, replace = TRUE) #remuestra de x
  xbarstarNP <- mean(bootstrapsampleNP)
  VstarNP[i] <- (2*n*xbarstarNP)/xbar

  recip_lambdahatstarNP[i] <- mean(bootstrapsampleNP)
}
```

```
summary(VstarNP)

summary(recip_lambdahatstarNP)
```

```
#Intervalo de confianza bootstrap no paramétrico.
IC_NP <-quantile(VstarNP, probs = c(0.025, 0.975))/ (xbar*2*n)
```

```r
#Generar muestras bootstrap paramétrico
VstarP<- NULL
recip_lambdahatstarP <- NULL

set.seed(1)
for (i in 1:1000){

 bootstrapsampleP <- rexp(n,lambdahat)
  xbarstarP <- mean(bootstrapsampleP)
  VstarP[i]=(2*n*xbarstarP)/xbar
  recip_lambdahatstarP[i]=mean(bootstrapsampleP)

}
```

```r
summary(VstarP)

summary(recip_lambdahatstarP)
```

```r
#Intervalo de confianza bootstrap no paramétrico.
IC_P <-quantile(VstarP, probs = c(0.025, 0.975))/ (xbar*2*n)
```

Bibliografía

[1] George Casella and Roger L. Berger. *Statistical inference*. Brooks/COle Cengage Learning, 2002.

[2] Michael R. Chernick and Robert A. LaBudde. *An introduction to bootstrap methods with applications to R*. Wiley, 2011.

[3] B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1), 1979.

[4] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Number 57 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, Boca Raton, Florida, USA, 1993.

[5] Tim C. Hesterberg. What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum. *The American Statistician*, 69(4):371–386, 2015.

[6] Robert V. Hogg, Joseph W. McKean, and Allen T. Craig. *Introduction to mathematical statistics*. Pearson Education International, 6 edition, 2005.

[7] J. G. Kalbfleisch. *Probability and Statistical Inference II*, volume 2.

[8] Andrew P. King and Robert J. Eckersley. *Statistics for Biomedical Engineers and scientists: How to visualize and Analyze Data*. Academic Press, an imprint of Elsevier, 2019.

[9] Alexander M. Mood, Franklin A. Graybill, and Duane C. Boes. *Introduction to the theory of statistics*. McGraw-Hill, 3 edition, 1974.

[10] W. D. Ray and A. E. N. T. Pitman. An exact distribution of the fisher-behrens-welch statistic for testing the difference between the means of two normal populations with unknown variance. *Journal of the Royal Statistical Society. Series B (Methodological)*, 23(2):377–384, 1961.

[11] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611, 1965.

[12] Ronald E. Walpole, Raymond H. Myers, and Sharon L. Myers. *Probabilidad y estadistica para Ingenieria y Ciencias (9a. Ed.)*. Pearson Educacion, 2012.