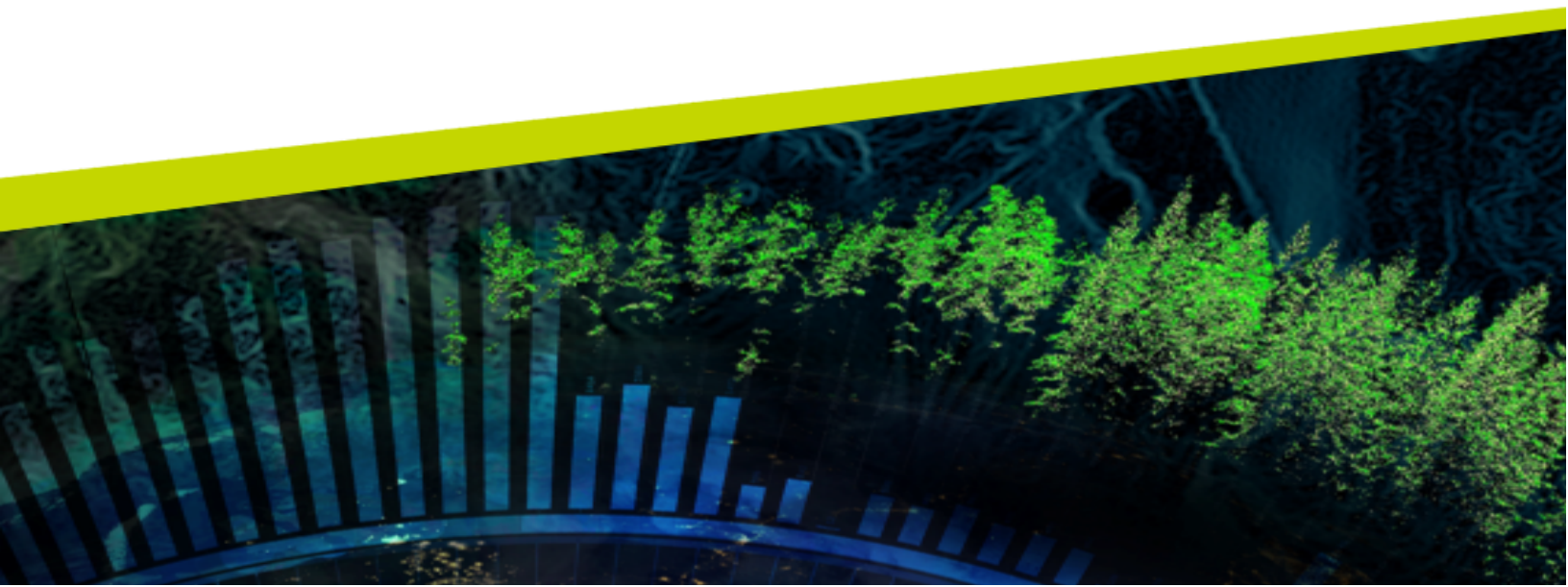
INFERENCIA Y MODELOS ESTADÍSTICOS

Jacqueline Köhler C. y José Luis Jara V.



CAPÍTULO 5. INFERENCIA CON MEDIAS MUESTRALES

En el capítulo 4 conocimos los principios de la inferencia y definimos los principales conceptos involucrados. En dicho capítulo conocimos el modelo normal, es decir, que la distribución muestral de la media sigue aproximadamente una distribución normal, supuesto que en general se cumple si la muestra tiene a lo menos 30 observaciones.

Veremos que diversas pruebas estadísticas consideran el modelo normal, aunque otras consideran estadísticos (estimaciones puntuales) diferentes que siguen otras distribuciones que ya conocimos en el capítulo 3.

En este capítulo veremos nuestras primeras pruebas estadísticas, las cuales nos permitirán inferir acerca de una o dos medias muestrales. Para ello nos basaremos principalmente en las explicaciones que ofrecen Diez y col. (2017, pp. 219-239) y Meena (2020).

5.1 PRUEBA Z

Como ya adelantamos, la prueba Z es adecuada para inferir acerca de las medias con una o dos muestras, aunque aquí solo veremos el primer caso. Para poder usarla, debemos **verificar el cumplimiento** de algunas condiciones, muchas de las cuales están asociadas al modelo normal que conocimos en el capítulo anterior:

- Las observaciones deben ser independientes, es decir que la elección de una observación para la muestra no influye en la selección de las otras.
- La población de donde se obtuvo la muestra sigue aproximadamente una distribución normal.
- La muestra debe tener al menos 30 observaciones (y asumir que la varianza observada corresponde a la varianza de la población). Si la muestra tiene menos de 30 observaciones, se debe conocer la varianza de la población.

Esta prueba resulta adecuada si queremos **asegurar** o **descartar** que la media de la población tiene un cierto **valor hipotético**. Esteban Quito es gerente de un grupo de inversiones que actualmente brinda apoyo financiero a más de 300 pequeñas empresas. El Sr. Quito desea saber para su campaña de marketing si, en promedio, las utilidades obtenidas el mes pasado por las empresas a las que brinda apoyo fueron de 20 millones de pesos. Para ello, nos ha informado que la desviación estándar para las utilidades de las empresas durante el mes pasado es de 2,32 millones de pesos y nos ha proporcionado una muestra, obtenida mediante muestreo aleatorio simple, con las utilidades (en millones de pesos) reportadas por 20 de las empresas durante dicho periodo, que se muestra en la tabla 5.1. La media observada en esta muestra es $\overline{x} = 26.066$.

Empresa	Utilidad [M$]	Empresa	Utilidad [M$]	Empresa	Utilidad [M$]	Empresa	Utilidad [M$]
1	19,33	6	22,22	11	22,55	16	29,68
2	29,37	7	31,26	12	20,69	17	29,27
3	29,14	8	26,92	13	24,68	18	26,72
4	32,10	9	31,40	14	28,74	19	27,08
5	25,04	10	17,66	15	26,85	20	20,62

Tabla 5.1: muestra para el ejemplo de prueba Z con una muestra.

El Sr. Quito nos ha dicho que debemos ser muy exigentes con respecto a nuestras conclusiones, por lo que se decide usar un nivel de significación $\alpha = 0,01$ (es decir, un nivel de confianza de 99 %).

Comencemos por formular nuestras hipótesis:

H_0: la media de las utilidades obtenidas por las empresas el mes pasado (μ) es de 20 millones de pesos, es decir: $\mu = 20$ [M$].

H_A: las utilidades obtenidas el mes pasado por las empresas son, en promedio, distintas de 20 millones de pesos, es decir: $\mu \neq 20$ [M$].

Ahora debemos verificar el cumplimiento de las condiciones para poder usar la prueba Z. En cuanto a la primera condición, el enunciado nos indica que, si bien la muestra tiene solo 20 observaciones, la desviación estándar de la población es conocida, por lo que se verifica su cumplimiento.

También podemos comprobar en el enunciado que las observaciones son independientes entre sí, pues fueron obtenidas mediante muestreo aleatorio simple y corresponden a menos del 10 % de la población.

En cuanto a la distribución de la muestra, el gráfico Q-Q de la figura 5.1 (obtenido mediante el script 5.1) nos muestra que no se observan valores atípicos.

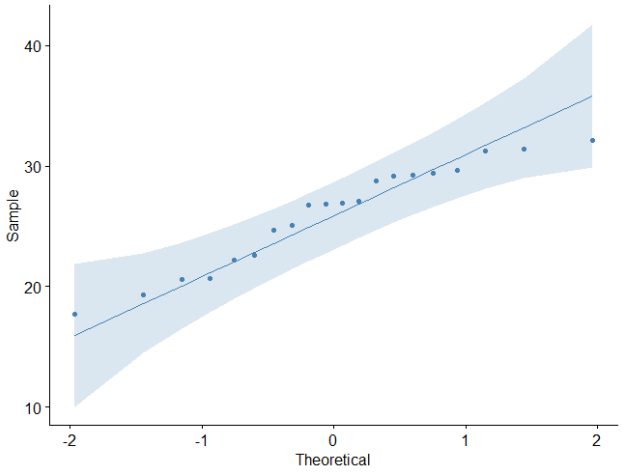


Figura 5.1: gráfico Q-Q para la muestra de la tabla 5.1.

Otra forma de comprobar esta condición es mediante la prueba de Shapiro-Wilk (Parada, 2019), que podemos realizar en R mediante la función `shapiro.test(x)`, donde `x` es un vector con las observaciones de la muestra. La hipótesis nula de esta prueba es que la muestra fue extraída desde una distribución normal (por ende, la hipótesis alternativa es que la distribución detrás de la muestra es diferente a la normal). Al ejecutar el script, podemos ver que el valor p obtenido es $p = 0,244$, muy superior a nuestro nivel de significación, por lo que podemos suponer con relativa confianza que la población de donde proviene la muestra sigue una distribución muestral.

Puesto que hemos comprobado que se cumplen todas las condiciones, podemos hacer una prueba Z para una muestra. Comencemos por calcular ahora el **estadístico de prueba** como ya hemos estudiado, usando para ello la ecuación 3.8. No debemos olvidat que estamos trabajando con la distribución muestral de la media en muestras de tamaño 20, por lo que debemos usar su error estándar en esta ecuación.

$$Z = \frac{\overline{x} - \mu}{SE} = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{26,066 - 20}{\frac{2,32}{\sqrt{20}}} = 11.69309$$

Con este resultado calculamos el valor p. Debemos recordar que las funciones de R (al igual que las antiguas tablas de probabilidades) nos entregan la probabilidad asociada al área correspondiente a una sola cola de la distribución, por lo que debemos multiplicar el resultado por 2 para considerar ambas colas si, como en este caso, se trata de una prueba bilateral. Al hacer la llamada `2 * pnorm(11.693, lower.tail = FALSE)`, obtenemos que $p = 1,382574 \times 10^{-31} < 0.01$[1] En este caso, el valor p obtenido es mucho menor que el nivel de significación establecido, con lo que se rechaza la hipótesis nula en favor de la hipótesis alternativa. Así, concluimos que los datos **sugieren** con 99 % de confianza que, en promedio, las utilidades obtenidas por las empresas durante el mes pasado difieren de los 20 millones de pesos establecidos como hipótesis.

Un comentario importante es que, si hubiésemos obtenido, por ejemplo, $p = 0,009$, deberíamos ser cuidadosos puesto que dicho valor es bastante cercano al nivel de significación establecido, por lo que sería prudente

[1] Por convención, los valores p suelen reportarse con tres decimales, pero en este caso presentamos el resultado con detalle por claridad.

evaluar los resultados con una muestra más grande.

R no tiene una función nativa para realizar esta prueba, pero sí está disponible en el paquete `TeachingDemos` por lo que puede obtenerse fácilmente con una llamada a la función `z.test(x, mu, stdev, alternative, conf.level)`, donde:

- `x`: vector con las observaciones de la muestra.
- `mu`: valor nulo.
- `stdev`: desviación estándar de la población.
- `alternative`: tipo de hipótesis alternativa. Puede tomar los valores `"two.sided"` para indicar que corresponde a que la media de la población es menor que el valor nulo (bilateral), `"less"` si es que la media de la población es menor que el valor nulo (unilateral) o `"greater"` si es que la media de la población es mayor que el valor nulo (también unilateral).
- `conf.level`: nivel de confianza.

El script 5.1 muestra el desarrollo de este ejemplo en forma manual y luego, en las líneas 42 y 49, dos alternativas equivalentes usando la función `z.test()`. El resultado que se obtiene al usar esta función es el que se muestra en la figura 5.2, idéntico al obtenido en nuestro desarrollo previo.

```
One Sample z-test

data:  media
z = 11.693, n = 20.00000, Std. Dev. = 2.32000, Std. Dev. of the sample mean = 0.51877,
p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 20
99 percent confidence interval:
 24.72974 27.40226
sample estimates:
mean of media
       26.066
```

Figura 5.2: resultado de la prueba Z para una muestra.

Script 5.1: prueba Z para una muestra.

```
1  library(TeachingDemos)
2  library(ggpubr)
3
4  # Ingresar los datos.
5  muestra <- c(19.33, 29.37, 29.14, 32.10, 25.04, 22.22, 31.26, 26.92,
6               31.40, 17.66, 22.55, 20.69, 24.68, 28.74, 26.85, 29.68,
7               29.27, 26.72, 27.08, 20.62)
8
9  # Establecer los datos conocidos.
10 desv_est <- 2.32
11 n <- length(muestra)
12 valor_nulo <- 20
13
14 # Crear gráfico Q-Q para verificar la distribución de la muestra,
15 datos <- data.frame(muestra)
16
17 g <- ggqqplot(datos, x = "muestra", color = "SteelBlue")
18 print(g)
19
20 # Verificar distribución muestral usando la prueba de normalidad
21 # de Shapiro-Wilk.
22 normalidad <- shapiro.test(muestra)
23 print(normalidad)
24
```

```r
# Fijar un nivel de significación.
alfa <- 0.01

# Calcular la media de la muestra.
cat("\tPrueba Z para una muestra\n\n")
media <- mean(muestra)
cat("Media =", media, "M$\n")

# Calcular el estadístico de prueba.
Z <- (media - valor_nulo) / (desv_est / sqrt(n))
cat("Z =", Z, "\n")

# Calcular el valor p.
p <- 2 * pnorm(Z, lower.tail = FALSE)
cat("p =", p, "\n")

# Hacer la prueba Z con R.
# Una alternativa es usando la media muestral y el tamaño de la muestra.
prueba1 <- z.test(media, mu = valor_nulo, n = 20, alternative = "two.sided",
                  stdev = desv_est, conf.level = 1-alfa)

print(prueba1)

# Otra opción es usando la muestra directamente.
prueba2 <- z.test(muestra, mu = valor_nulo, alternative = "two.sided",
                  stdev = desv_est, conf.level = 1-alfa)

print(prueba2)
```

5.2 PRUEBA T DE STUDENT

En la práctica, rara vez podemos conocer la desviación estándar de la población y a menudo nos encontraremos con muestras pequeñas, por lo que la prueba Z no es muy utilizada.

En el caso de la media, el teorema del límite central se cumple para datos normales, es decir, independientemente del tamaño de la muestra, la media muestral tendrá una distribución cercana a la normal siempre que las observaciones sean independientes y provengan de una distribución cercana a la normal. Sin embargo, cuando el conjunto de datos es pequeño, resulta muy difícil comprobar el cumplimiento de estas condiciones.

En el capítulo 3 conocimos la distribución t de Student, o simplemente distribución t. Vimos que un aspecto destacado de esta distribución, siempre centrada en 0 y definida únicamente por los grados de libertad (ν) como parámetro, es su semejanza con la distribución normal pese a que sus colas son algo más gruesas. Este grosor adicional de las colas tiene como consecuencia que, para la distribución t, es más probable que una observación esté a más de dos desviaciones estándares de la media que en el caso de la distribución normal. Este fenómeno permite que la estimación del error estándar sea más certera que al usar la distribución normal cuando el conjunto de datos es pequeño.

La prueba t de Student, basada en la distribución t, es en consecuencia la alternativa más ampliamente empleada para inferir acerca de una o dos medias muestrales.

5.2.1 Prueba t para una muestra

La prueba t opera bajo los siguientes supuestos:

1. Las observaciones son independientes entre sí.
2. Las observaciones provienen de una distribución cercana a la normal.

Podemos ver que estas condiciones son casi las mismas que para la prueba Z, excepto por el hecho de que no exige que el tamaño de la muestra sea mayor a 30. La ventaja evidente de eliminar esta restricción es que la distribución t permite su uso para muestras pequeñas, pero es igualmente adecuada cuando la muestra es grande. Esto se debe a que la forma de la distribución t es regulada por los grados de libertad y, a medida que aumentan, más se parece a una distribución normal. Este parámetro, al trabajar con medias de muestras de tamaño n, siempre estará dado por $\nu = n - 1$.

Tomemos el siguiente problema para ilustrar la prueba de hipótesis para la media de una muestra usando el modelo t: un ingeniero en Informática necesita determinar si el tiempo promedio que tarda una implementación dada de un algoritmo en resolver un problema, sabiendo que el algoritmo siempre se ejecuta en las mismas condiciones (misma máquina, igual disponibilidad de recursos de hardware y tamaño constante de las instancias), es **inferior a 500 milisegundos**. Para ello, ha seleccionado aleatoriamente 15 instancias del problema y registrado el tiempo de ejecución del algoritmo (en milisegundos) para cada una de ellas, como muestra la tabla 5.2.

Obs.	t [ms]	Obs.	t [ms]	Obs.	t [ms]
1	411,5538	6	388,6731	11	418,1169
2	393,2753	7	430,0382	12	408,4110
3	445,8905	8	469,4734	13	463,3733
4	411,4022	9	409,5844	14	407,0908
5	498,8969	10	442,0800	15	516,5222

Tabla 5.2: tiempo de ejecución para las instancias de la muestra.

El primer paso es formular las hipótesis:

H_0: el tiempo promedio que tarda el algoritmo en resolver una instancia del problema es igual a 500 milisegundos.
H_A: el tiempo promedio que tarda el algoritmo en resolver una instancia del problema es inferior a 500 milisegundos.

Recordemos que μ_0 es el valor nulo, por lo que en este caso $\mu_0 = 500$ [ms]. Matemáticamente, las hipótesis anteriores pueden formularse como:

Denotando como μ al tiempo medio que tarda la implementación del algoritmo en resolver una instancia cualquiera del problema:
H_0: $\mu = \mu_0$, esto es $\mu = 500$
H_A: $\mu < \mu_0$, es decir $\mu < 500$

Ahora debemos verificar que se cumplen las condiciones necesarias para usar la distribución t:

- Como las muestras fueron elegidas al azar, se puede asumir que son independientes.
- El gráfico de la figura 5.3 muestra que es válido suponer una distribución cercana a la normal. Si bien los puntos de la muestra no forman una recta, no se observan valores atípicos que se alejen de la región aceptable.

La media de la muestra es de $\overline{x} = 434,2921$, y la desviación estándar es $s = 38,0963$.

En este caso, el estadístico de prueba es el estadístico T, el cual sigue una distribución t con $\nu = n-1$ grados de libertad y está dado por la ecuación 5.1, donde la subexpresión $(s/\sqrt{n})$ corresponde al error estándar de la media (cuando no se conoce la desviación estándar de la población σ).

$$T = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}} \tag{5.1}$$

Así, para el ejemplo tenemos que:

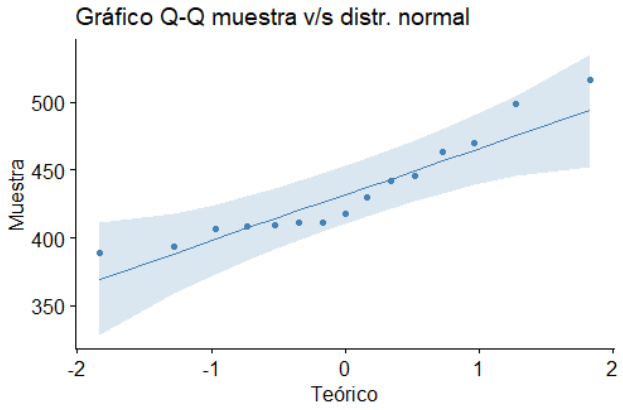


Figura 5.3: gráfico para comprobar el supuesto de normalidad.

$$T = \frac{434,2921 - 500}{\frac{38,0963}{\sqrt{15}}} = -6,6801$$

A partir de este resultado, obtenemos el valor p con ayuda de la función `pt()`, obteniéndose $p = 5,219 \cdot 10^{-6}$, o simplemente, como dicta la convención, $p < 0,001$.

La fórmula para construir el intervalo de confianza usando la distribución t es ligeramente diferente al caso normal, como muestra la ecuación 5.2. Para este ejemplo consideraremos un nivel de confianza de 97,5 % (es decir, un nivel de significación $\alpha = 0,025$).

$$\overline{x} \pm t^*_\nu \cdot SE_{\overline{x}} \tag{5.2}$$

Fijémonos en que en la ecuación 5.2 aparece el nuevo valor t^*_ν, el cual se obtiene a partir del nivel de confianza y la distribución t con ν grados de libertad (en este caso, $\nu = 14$), usando para ello una tabla de distribución t o la función `qt()` en R. Como puede verse al ejecutar el script 5.2, en este caso $t^*_\nu = 2,1448$.

Para el cálculo del error estándar, nuevamente se emplea la ecuación 4.2:

$$SE_{\overline{x}} = \frac{s}{\sqrt{n}} = \frac{38,0963}{\sqrt{15}} = 9,8364$$

Así, el intervalo de confianza está dado por:

$$(-\infty, \overline{x} + t^*_{nu} * SE_{\overline{x}}] = (-\infty, 434.2921 + 2,1448 * 9,8364] = (-\infty, 455,3892]$$

Una vez más, R proporciona una función nativa que nos permite realizar esta prueba de manera rápida y sencilla, con una llamada como `t.test(x, alternative, mu, conf.level)`, donde:

- `x`: vector no vacío de valores numéricos (la muestra).
- `alternative`: tipo de prueba de hipótesis, con la misma interpretación que para la función `z.test()`, es decir `"two.sided"` para una prueba bilateral, y `"greater"` o `"less"` para pruebas unilaterales.
- `mu`: valor nulo.
- `conf.level`: nivel de confianza.

El script 5.2 muestra el desarrollo en R para este ejemplo, incluyendo la construcción del gráfico de la figura 5.3, con iguales resultados al realizar la prueba paso a paso y con la función `t.test()`.

A partir de estos resultados, donde observamos que el valor p obtenido es muy pequeño, podemos entender que si se cumple el supuesto de que la media poblacional es $\mu = 500$ [ms] (hipótesis nula), sería muy improbable

obtener una media muestral de $\overline{x} = 434,2921$. Además, el valor p es menor que el nivel de significación definido, por lo que la evidencia nos sugiere rechazar H_0 en favor de H_A. Se puede afirmar, con 97,5 % de confianza, que el tiempo promedio que tarda el algoritmo en resolver una instancia del problema es inferior a 500 milisegundos.

Script 5.2: prueba t para una muestra.

```
library(ggpubr)

# Cargar los datos.
tiempo <- c(411.5538, 393.2753, 445.8905, 411.4022, 498.8969,
            388.6731, 430.0382, 469.4734, 409.5844, 442.0800,
            418.1169, 408.4110, 463.3733,407.0908, 516.5222)

# Establecer los datos conocidos.
n <- length(tiempo)
grados_libertad <- n - 1
valor_nulo <- 500


# Verificar si la distribución se acerca a la normal.
g <- ggqqplot(data = data.frame(tiempo),
              x = "tiempo",
              color = "steelblue",
              xlab = "Teórico",
              ylab = "Muestra",
              title = "Gráfico Q-Q muestra v/s distr. normal")

print(g)

# Fijar un nivel de significación.
alfa <- 0.025

# Calcular el estadístico de prueba.
cat("\tPrueba t para una muestra\n\n")
media <- mean(tiempo)
cat("Media =", media, "M$\n")
desv_est <- sd(tiempo)
error <- desv_est / sqrt(n)
t <- (media - valor_nulo) / error
cat("t =", t, "\n")

# Calcular el valor p.
p <- pt(t, df = grados_libertad, lower.tail = TRUE)
cat("p =", p, "\n")

# Construir el intervalo de confianza.
t_critico <- qt(alfa, df = grados_libertad, lower.tail = FALSE)
superior <- media + t_critico * error
cat("Intervalo de confianza = (-Inf, ", superior, "]\n", sep = "")

# Aplicar la prueba t de Student con la funcuón de R.
prueba <- t.test(tiempo,
                 alternative = "less",
                 mu = valor_nulo,
                 conf.level = 1 - alfa)

print(prueba)
```

5.2.2 Prueba t para dos muestras pareadas

Para esta prueba, supongamos ahora que el ingeniero en Informática del ejemplo anterior tiene dos algoritmos diferentes (A y B) que, en teoría, deberían tardar lo mismo en resolver un problema. Para ello, probó ambos algoritmos con 35 instancias del problema (elegidas al azar) de igual tamaño y registró los tiempos de ejecución (en milisegundos) de ambos algoritmos bajo iguales condiciones para cada una de ellas, además de calcular la diferencia en los tiempos de ejecución, como muestra la tabla 5.3. El ingeniero desea comprobar si efectivamente el rendimiento de ambos algoritmos es equivalente.

instancia	t_A [ms]	t_B [ms]	dif [ms]	instancia	t_A [ms]	t_B [ms]	dif [ms]
1	436,5736	408,5142	28,0594	19	438,5959	458,2536	-19,6577
2	470,7937	450,1075	20,6862	20	439,7409	474,9863	-35,2454
3	445,8354	490,2311	-44,3957	21	464,5916	496,0153	-31,4237
4	470,9810	513,6910	-42,7100	22	467,9926	485,8112	-17,8186
5	485,9394	467,6467	18,2927	23	415,3252	457,4253	-42,1001
6	464,6145	484,1897	-19,5752	24	495,4094	483,3700	12,0394
7	466,2139	465,9334	0,2805	25	493,7082	510,7131	-17,0049
8	468,9065	502,6670	-33,7605	26	433,1082	467,5739	-34,4657
9	473,8778	444,9693	28,9085	27	445,7433	482,5621	-36,8188
10	413,0639	456,3341	-43,2702	28	515,2049	453,5986	61,6063
11	496,8705	501,1443	-4,2738	29	441,9420	385,9391	56,0029
12	450,6578	471,7833	-21,1255	30	472,1396	548,7884	-76,6488
13	502,9759	441,1206	61,8553	31	451,2234	467,2533	-16,0299
14	465,6358	544,1575	-78,5217	32	476,5149	494,7049	-18,1900
15	437,6397	447,8844	-10,2447	33	440,7918	451,9716	-11,1798
16	458,8806	432,4108	26,4698	34	460,1070	522,3699	-62,2629
17	503,1435	477,1712	25,9723	35	450,1008	444,1270	5,9738
18	430,0524	482,4828	-52,4304				

Tabla 5.3: tiempos de ejecución de cada algoritmo para las instancias de la muestra.

Para este ejemplo, tenemos dos tiempos de ejecución diferentes para cada instancia del problema: uno con cada algoritmo. En consecuencia, los datos están **apareados** o **pareados**. Es decir, cada observación de un conjunto tiene una correspondencia o conexión especial con exactamente una observación del otro. Una forma de uso común para examinar datos pareados es usar **las diferencias entre cada par de observaciones**, para lo cual podemos usar la prueba t de Student vista en la sección anterior.

La media de las diferencias es $\overline{x}_{dif} = -12,08591$ y la desviación estándar es $s_{dif} = 36,08183$.

Una vez más, comenzamos por formular las hipótesis:

H_0: la media de las diferencias en los tiempos de ejecución es igual a 0.
H_A: la media de las diferencias en los tiempos de ejecución es distinta de 0.

Que matemáticamente se expresan como:

Denotando la media de las diferencias en los tiempos de ejecución necesitados por ambos algoritmos para cualquier instancia del problema como μ_{dif}:
H_0: $\mu_{dif} = 0$
H_A: $\mu_{dif} \neq 0$

Como siguiente paso, verificamos el cumplimiento de las condiciones. Como las instancias fueron escogidas al azar, se puede suponer razonablemente que las observaciones son independientes, pues además el conjunto de instancias posibles es muy grande (o infinito) y las 35 seleccionadas no superan el 10 % de la población. Además, al aplicar una prueba de normalidad de Shapiro-Wilk (ver script 5.3, línea 23) se obtiene $p = 0,357$, con lo que podemos concluir que la diferencia en los tiempos de ejecución se acerca razonablemente a una distribución normal. En consecuencia, podemos proceder con la prueba t de Student. El ingeniero no necesita ser especialmente riguroso, por lo que usaremos un nivel de confianza del 95 %.

En este caso, la función `t.test()` de R permite efectuar la prueba de dos maneras diferentes (con idéntico resultado), como muestra el script 5.3. La primera de ellas (línea 32) es aplicar la prueba t directamente a las diferencias, tal como en la sección anterior (es decir, una prueba t para una muestra). La segunda (línea 41) consiste en entregar a la función ambas muestras por separado e indicarle que están pareadas. En este caso, la llamada tiene la forma `t.test(x, y, paired, alternative, mu, conf.level)`, donde los argumentos son:

- `x`: vector de valores numéricos para la primera muestra.
- `y`: vector de valores numéricos para la segunda muestra.
- `paired`: booleano (por defecto falso) que, cuando es verdadero, indica que ambas muestras están pareadas.
- `alternative`: tipo de prueba de hipótesis.
- `mu`: valor nulo.
- `conf.level`: nivel de confianza.

Script 5.3: inferencia con la media de las diferencias entre dos muestras pareadas usando la distribución t.

```
# Cargar los datos.
instancia <- seq(1, 35, 1)

t_A <- c(436.5736, 470.7937, 445.8354, 470.9810, 485.9394,
         464.6145, 466.2139, 468.9065, 473.8778, 413.0639,
         496.8705, 450.6578, 502.9759, 465.6358, 437.6397,
         458.8806, 503.1435, 430.0524, 438.5959, 439.7409,
         464.5916, 467.9926, 415.3252, 495.4094, 493.7082,
         433.1082, 445.7433, 515.2049, 441.9420, 472.1396,
         451.2234, 476.5149, 440.7918, 460.1070, 450.1008)

t_B <- c(408.5142, 450.1075, 490.2311, 513.6910, 467.6467,
         484.1897, 465.9334, 502.6670, 444.9693, 456.3341,
         501.1443, 471.7833, 441.1206, 544.1575, 447.8844,
         432.4108, 477.1712, 482.4828, 458.2536, 474.9863,
         496.0153, 485.8112, 457.4253, 483.3700, 510.7131,
         467.5739, 482.5621, 453.5986, 385.9391, 548.7884,
         467.2533, 494.7049, 451.9716, 522.3699, 444.1270)

diferencia <- t_A - t_B

# Verificar si la distribución se acerca a la normal.
normalidad <- shapiro.test(diferencia)
print(normalidad)

# Fijar un nivel de significación.
alfa <- 0.05

# Aplicar la prueba t de Student a la diferencia de medias.
valor_nulo <- 0

prueba_1 <- t.test(diferencia,
                   alternative = "two.sided",
                   mu = valor_nulo,
                   conf.level = 1 - alfa)

print(prueba_1)

# Otra alternativa puede ser aplicar la prueba t de Student
# para dos muestras pareadas.
prueba_2 <- t.test(x = t_A,
                   y = t_B,
                   paired = TRUE,
                   alternative = "two.sided",
                   mu = valor_nulo,
```

```
                    conf.level = 1 - alfa)

print(prueba_2)
```

Los resultados para esta prueba son:

- El valor para el estadístico de prueba T es $t = -1,9816$.
- Se consideran $df = 34$ grados de libertad para la distribución t.
- El valor p obtenido es $p = 0,05565$.
- El intervalo de confianza obtenido es $[-24,4804542; 0,3086313]$.
- La media de la muestra es $\overline{x} = -12,08591$.

En este caso, la media de las diferencias está dentro del intervalo de confianza, y además el valor p es mayor que el nivel de significación, por lo que se falla al rechazar la hipótesis nula. Pero el resultado está cerca del borde de significación. En consecuencia, se puede afirmar con 95 % de confianza que pareciera no haber suficiente evidencia para descartar que ambos algoritmos tardan, en promedio, lo mismo en procesar las instancias del problema, aunque sería necesario conseguir una muestra más grande para tener mayor certeza.

5.2.3 Prueba t para dos muestras independientes

En este caso, la prueba t se usa para comparar las medias de dos poblaciones en que las observaciones con que se cuenta no tienen relación con ninguna de las otras observaciones, ni influyen en su selección, ni en la misma ni en la otra muestra. En este caso la inferencia se hace sobre la diferencia de las medias: $\mu_1 - \mu_2 = d_0$, donde d_0 es un valor hipotético fijo para la diferencia. Usualmente se usa $d_0 = 0$, en cuyo caso las muestras podrían provenir de dos poblaciones distintas con igual media, o desde la misma población. Para ello, la prueba usa como estimador puntual **la diferencia de las medias muestrales** $(\overline{x}_1 - \overline{x}_2)$. Así, el estadístico T en este caso toma la forma de la ecuación 5.3.

$$T = \frac{(\overline{x}_1 - \overline{x}_2) - d_0}{SE_{(\overline{x}_1 - \overline{x}_2)}} \tag{5.3}$$

Al usar la distribución t de Student para la diferencia de medias, se deben cumplir los siguientes requisitos:

1. Cada muestra cumple las condiciones para usar la distribución t.
2. Las muestras son independientes entre sí.

Veamos el funcionamiento de esta prueba con un ejemplo. El doctor E. L. Matta Sanno desea determinar si una nueva vacuna A es más efectiva que otra vacuna B, a fin de inmunizar a la población mundial contra una terrible enfermedad. Para ello, ha reclutado a un grupo de 28 voluntarios en diferentes países, 15 de los cuales (seleccionados al azar) recibieron la vacuna A y los 13 restantes, la vacuna B. La tabla 5.4 muestra, para cada voluntario, la concentración de anticuerpos (en microgramos por cada mililitro de sangre) al cabo de un mes de recibir la vacuna.

Las hipótesis a formular en este caso son:

H_0: no hay diferencia entre la efectividad promedio de ambas vacunas.
H_A: la vacuna A es, en promedio, más efectiva que la B.

En lenguaje matemático:

Si μ_A y μ_B son la concentraciones medias de anticuerpos presentes en personas luego de un mes de recibir la vacuna A y B, respectivamente, entonces:
H_0: $\mu_A = \mu_B$
H_A: $\mu_A > \mu_B$

Como es habitual, debemos ahora verificar el cumplimiento de las condiciones. Ambas muestras son independientes entre sí, pues son diferentes voluntarios y fueron designados aleatoriamente a cada grupo. Además, se puede asumir que las observaciones son independientes, pues cada muestra es significativamente menor a

Anticuerpos [mg/ml]	
Vacuna A	Vacuna B
6,04	5,32
19,84	3,31
8,62	5,68
13,02	5,73
12,20	4,86
14,78	5,68
4,53	2,93
26,67	5,48
3,14	6,10
19,14	2,56
10,86	7,52
13,13	7,41
6,34	4,02
11,16	
7,62	

Tabla 5.4: Concentración de anticuerpos de los pacientes vacunados.

la población total a vacunar. En cuanto al supuesto de normalidad para cada muestra, al aplicar a cada una la prueba de Shapiro-Wilk (script 5.4, líneas 13 y 15) se obtiene, respectivamente, $p = 0,428$ y $p = 0,445$. En ambos casos el valor p es bastante alto, por lo que podemos concluir que ambas muestras provienen de poblaciones que se distribuyen de forma aproximadamente normal. Puesto que hemos verificado las condiciones, podemos llevar a cabo la prueba t para dos muestras independientes.

Ahora bien, como las muestras son algo pequeñas, sería prudente proceder con algo más de cautela. Además, este escenario, un error tipo I (rechazar H_0 cuando es verdadera) implicaría reducir innecesariamente la cantidad de vacunas disponibles y retrasar el proceso de vacunación, poniendo en riesgo a todos los habitantes del planeta. Un error tipo II, en cambio, podría causar que se continúe el uso indistinto de ambas vacunas retrasando ligeramente el efecto inmune en la población. En consecuencia, el error tipo I es más grave, por lo que el nivel de significación debiese ser aún más exigente. En consecuencia, optaremos por $\alpha = 0,01$.

Al aplicar la prueba t (script 5.4), obtenemos que la diferencia entre las medias es $6,683$ [mg/ml] y que el intervalo de confianza es $[2,2739; \infty)$. Además, el valor p es $p < 0.001$, muy inferior al nivel de significación $\alpha = 0,01$. Esto significa que la evidencia en favor de H_A es muy fuerte, por lo rechazamos la hipótesis nula. En consecuencia, podemos concluir con 99 % de confianza que la vacuna A es, en promedio, mejor que la vacuna B (produce una mayor concentración media de anticuerpos en las personas vacunadas con ella que la producida por la vacuna B).

Script 5.4: prueba t para dos muestras independientes.

```
1  library(ggpubr)
2
3  # Cargar los datos.
4  vacuna_A <- c(6.04, 19.84, 8.62, 13.02, 12.20, 14.78, 4.53, 26.67,
5                3.14, 19.14, 10.86, 13.13, 6.34, 11.16, 7.62)
6
7  vacuna_B <- c(5.32, 3.31, 5.68, 5.73, 4.86, 5.68, 2.93, 5.48, 6.10,
8                2.56, 7.52, 7.41, 4.02)
9
10 # Verificar si las muestras se distribuyen de manera cercana
11 # a la normal.
12 normalidad_A <- shapiro.test(vacuna_A)
13 print(normalidad_A)
14 normalidad_B <- shapiro.test(vacuna_B)
15 print(normalidad_B)
16
```

```r
# Fijar un nivel de significación.
alfa <- 0.01

# Aplicar la prueba t para dos muestras independientes.
prueba <- t.test(x = vacuna_A,
                 y = vacuna_B,
                 paired = FALSE,
                 alternative = "greater",
                 mu = 0,
                 conf.level = 1 - alfa)

print(prueba)

# Calcular la diferencia entre las medias.
media_A <- mean(vacuna_A)
media_B <- mean(vacuna_B)
diferencia <- media_A - media_B
cat("Diferencia de las medias =", diferencia, "[mg/ml]\n")
```

Si estás leyendo atentamente, te habrás dado cuenta que ¡no hemos definido el error estándar para cuando tenemos dos muestras! En este caso, SE se construye a partir del error estándar de cada muestra, como se aprecia en la ecuación 5.4. En este escenario, la determinación de los grados de libertad es más compleja, por lo que se recomienda usar programas estadísticos o, en su defecto, escoger el menor valor entre $n_1 - 1$ y $n_2 - 1$.

$$SE_{\overline{x}_1 - \overline{x}_2} = \sqrt{SE^2_{\overline{x}_1} + SE^2_{\overline{x}_2}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{5.4}$$

Se puede lograr un mejor ajuste de la distribución t si se sabe con certeza que las desviaciones estándares de ambas poblaciones son casi iguales. En este caso, se puede usar una **varianza agrupada** (s_p^2, del inglés *pooled variance*) que reemplaza tanto a s_1^2 como a s_2^2 en la ecuación 5.4. Esta varianza agrupada se calcula como muestra la ecuación 5.5 y, en este caso, se consideran $n_1 + n_2 - 2$ grados de libertad.

$$s_p^2 = \frac{s_1^2 \cdot (n_1 - 1) + s_2^2 \cdot (n_2 - 1)}{n_1 + n_2 - 2} \tag{5.5}$$

Por defecto, R utiliza la corrección de Welch para la prueba t de Student de la diferencia de dos medias, variante considerada más segura, que en general entrega resultados muy similares a la versión original de la prueba cuando las muestras tienen varianzas similares. No obstante, los resultados son bastante mejores cuando los tamaños de las muestras y sus desviaciones estándares son muy diferentes (Kassambara, 2019). La corrección de Welch calcula el error estándar como muestra la ecuación 5.4, pero ajusta los grados de libertad de acuerdo a la ecuación 5.6.

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1 - 1)} + \frac{s_2^4}{n_2^2(n_2 - 1)}} \tag{5.6}$$

5. EJERCICIOS PROPUESTOS

1. Investiga acerca de la prueba de Kolmogorov-Smirnov y explica cómo puede usarse para verificar si una distribución se asemeja a la normal. Compara esta prueba con la de Shapiro-Wilk.

2. Da un ejemplo de una pregunta de investigación sobre las asignaturas comunes en ingeniería que requiera utilizar una prueba t de Student para una muestra. Explica bien qué variable se mide y enuncia las hipótesis a docimar, justificando el valor nulo escogido y el tipo de hipótesis alternativa (bilateral o unilateral).
3. Da un ejemplo de una pregunta de investigación sobre los conciertos realizados en Santiago que requiera una prueba t de Student para dos muestras apareadas. Explica bien qué variable se mide y enuncia las hipótesis a docimar, justificando el valor nulo escogido y el tipo de hipótesis alternativa (bilateral o unilateral).
4. Da un ejemplo de una pregunta de investigación sobre el estado de la salud mental de estudiantes universitarios que requiera una prueba t de Student para dos muestras independientes. Explica bien qué variable se mide y enuncia las hipótesis a docimar, justificando el valor nulo escogido y el tipo de hipótesis alternativa (bilateral o unilateral).

Considera el siguiente enunciado:

> Para confirmar que el tiempo que requieren los estudiantes de ingeniería para desarrollar una guía de ejercicios de Cálculo I es de dos horas, se eligió aleatoriamente a 16 estudiantes de esta asignatura y se les pidió anotar el tiempo [min.] invertido en la tarea. Los resultados fueron los siguientes: 140,6; 133,3; 142,4; 86,4; 129,9; 110,8; 133,2; 129,1; 142,5; 150,2; 141,6; 111,0; 127,2; 137,9; 131,9; 121,9.

y responde las siguientes preguntas:

5. Enuncia las hipótesis nula y alternativa a contrastar. Analiza si es razonable en este caso considerar que los datos cumplen las condiciones para usar una prueba t de Student.
6. Independientemente del último resultado, aplica la prueba propuesta y obtén un intervalo de confianza y un valor p. Usando un nivel de significación adecuado, entrega una conclusión para la cuestión planteada.

Considera el siguiente enunciado:

> El departamento de control de calidad de un importante laboratorio requiere analizar la concentración de ingredientes activos presente en una muestra de 10 botellas diferentes de detergente líquido que ellos seleccionaron aleatoriamente en el último mes. Como se sospecha que esta concentración depende del catalizador que se use, la mitad del contenido de cada botella fue sometida a un catalizador, y la otra mitad a otro catalizador. En orden por botella seleccionada, los resultados fueron:
>
> > Catalizador 1: 62,9; 67,2; 67,4; 67,4; 67,2; 64,6; 69,6; 65,7; 68,2; 72,0.
> >
> > Catalizador 2: 66,8; 69,3; 69,6; 67,3; 68,8; 68,4; 68,6; 70,3; 69,6; 71,7.
>
> Como primer paso, el departamento de control de calidad necesita saber si la concentración media de concentraciones de ingredientes activos depende del catalizador elegido.

y responde las siguientes preguntas:

7. Propón las hipótesis nula y alternativa que permitan responder el problema planteado con una prueba t de Student. Muestra que es razonable considerar que estos datos cumplen las condiciones para usar la prueba propuesta y fija un nivel de significación apropiado.
8. Aplica la prueba propuesta y obtenga un intervalo de confianza y un valor p. ¿Cuál sería tu respuesta al departamento de control mencionado?

Considera el siguiente enunciado:

> Una fábrica de detectores de radón recibió consultas de sus clientes sobre si era conveniente comprar su nuevo modelo de detectores Radolmes+®para reemplazar los antiguos aparatos Radolmes®en su poder. Si bien los técnicos están seguros que la inversión es conveniente, la gerencia decidió hacer un estudio previo a la recomendación. Para esto, se introdujeron en una tómbola oscura los números de serie de los aparatos producidos en los últimos meses de ambos modelos y se seleccionaron 26 números sin mirar y girando la tómbola cinco veces entre cada selección, resultando escogidos 12 aparatos Radolmes y 14 aparatos Radolmes+. Luego, cada detector seleccionado se expuso a 100 pCi/l de radón. Las lecturas resultantes fueron las siguientes:

Radolmes: 105,6; 100,1; 90,9; 105,0; 91,2; 99,6; 96,9; 107,7; 96,5; 103,3; 91,3; 92,4.

Radolmes+: 98,9; 94,3; 95,9; 107,7; 102,0; 94,2; 100,6; 98,5; 99,1; 101,3; 94,4; 103,6; 95,3; 106,7.

y responde las siguientes preguntas:

9. ¿Qué hipótesis nula y alternativa se deberían docimar[2] con una prueba t de Student para responder a la inquietud planteada? ¿Cumplen los datos obtenidos las condiciones para usar esta prueba t de Student?
10. Aplicando la prueba t de Student para este caso, obtén un intervalo de confianza y un valor p. Qué aconsejarías a los directivos de la fábrica?

5.4 BIBLIOGRAFÍA DEL CAPÍTULO

Diez, D., Barr, C. D., & Çetinkaya-Rundel, M. (2017). *OpenIntro Statistics* (3.ª ed.). https://www.openintro.org/book/os/.

Kassambara, A. (2019). *Practical Statistics in R II - Comparing Goups: Numerical Variables*. Datanovia.

Meena, S. (2020). *Statistics for Analytics and Data Science: Hypothesis Testing and Z-Test vs. T-Test*. Consultado el 22 de septiembre de 2021, desde https://www.analyticsvidhya.com/blog/2020/06/statistics-analytics-hypothesis-testing-z-test-t-test/#h2_1

Parada, L. F. (2019). *Prueba de normalidad de Shapiro-Wilk*. Consultado el 22 de septiembre de 2021, desde https://rpubs.com/F3rnando/507482

[2] Término que suele ocuparse en estadística como sinónimo de "probar".

CAPÍTULO 6. INFERENCIA CON PROPORCIONES MUESTRALES

En el capítulo 5 conocimos las pruebas Z y t de Student para contrastar hipótesis con una y dos medias. Ahora estudiaremos los métodos de Wald y de Wilson para inferir acerca de una y dos proporciones, basándonos para ello en los textos de Diez y col. (2017, pp. 274-286), NIST/SEMATECH (2013, pp. 7.2.4, 7.2.4.1), Pértega y Pita (2004), Champely, Ekstrom, Dalgaard, Gill, Weibelzahl, Anandkumar, Ford, Volcic y de Rosario (2020) y Kabacoff (2017).

6.1 MÉTODO DE WALD

En el capítulo 3 vimos que, cuando queremos responder preguntas del tipo "¿qué proporción de la ciudadanía apoya al gobierno actual?", estamos hablando de una variable aleatoria que sigue una distribución binomial. En general, no conocemos la **probabilidad de éxito** p de la población, por lo que tenemos que usar el estimador puntual (correspondiente a la proporción de éxito de la muestra), denotado por $\hat{p}$. Este estimador se distribuye de manera cercana a la normal cuando se cumplen las siguientes condiciones:

1. Las observaciones de la muestra son independientes.
2. Se cumple la **condición de éxito-fracaso**, que establece que se espera observar al menos 10 observaciones correspondientes a éxito y al menos 10, correspondientes a fracasos. Matemáticamente, $np \geq 10$ y $n(1-p) \geq 10$.

Así, si la distribución muestral de $\hat{p}$ cumple con las condiciones anteriores, se dice que es cercana a la normalidad con media $\mu = p$, desviación estándar $\sigma = \sqrt{p(1-p)}$ y error estándar dado por la ecuación 6.1.

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \tag{6.1}$$

Obviamente el valor de p es desconocido, por lo que se debe sustituir por una aproximación en las ecuaciones anteriores, como veremos en las siguientes secciones.

6.1.1 Método de Wald para una proporción

El **método de Wald** permite construir intervalos de confianza y contrastar hipótesis bajo el supuesto de normalidad para una proporción. Consideremos el siguiente ejemplo: Aquiles Baeza, ingeniero en informática, desea conocer qué proporción de las ejecuciones de un algoritmo de ordenamiento para instancias con 100.000 elementos (bajo iguales condiciones de hardware y sistema) tardan menos de 25 segundos. Para ello, registró los tiempos de ejecución para 150 instancias generadas de manera aleatoria, encontrando que 64 % de dichas instancias fueron resueltas en un tiempo menor al señalado.

Notemos dos datos importantes del enunciado: se tiene una muestra de tamaño $n = 150$ donde se observa una proporción de éxitos $\hat{p} = 0,64$. Para responde a Baeza, podemos calcular el intervalo de confianza para la verdadera proporción a partir de esta muestra, si es que se cumplen las condiciones para que la distribución sea cercana a la normal.

En el enunciado del ejemplo nos indican que las instancias del problema fueron escogidas de manera aleatoria y sabemos que éstas representan menos del 10 % del total de instancias posibles, con lo que se verifica la independencia de las observaciones. Para verificar la condición de éxito-fracaso usamos la proporción de éxito **muestral** $\hat{p}$ como aproximación del parámetro p, por lo que en la muestra esperamos encontrar $0,64 \cdot 150 = 96$ instancias que tardan menos de 25 segundos y $(1-0,64) \cdot 150 = 54$ fracasos (instancias que tardan 25 segundos

o más). Luego, se cumple la condición de éxito-fracaso y podemos asumir que la distribución muestral de $\hat{p}$ sigue aproximadamente a la normal.

Podemos estimar el error estándar usando la ecuación 6.1, reemplazando una vez más p por $\hat{p}$:

$$SE_{\hat{p}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0,64(1-0,64)}{150}} = 0,0392$$

Con ello, construimos el intervalo de confianza para un nivel de significación $\alpha = 0,05$ usando la ecuación general (4.6) con $\hat{p}$ como estimador puntual:

$$\hat{p} \pm z^*_{0,975} \cdot SE_{\hat{p}} \approx 0,64 \pm 1,96 \cdot 0,0392 = [0,5632; 0.7168]$$

Este intervalo significa que tenemos 95 % de confianza que la proporción de instancias del problema (de 100.000 elementos) que el algoritmo ordena en menos de 25 segundos se encuentra entre 56,32 % y 71,6 %.

Desde luego, también podemos usar el modelo normal en el contexto de la prueba de hipótesis para una proporción. Para ello, se deben cumplir las condiciones de independencia y éxito-fracaso, que ya verificamos para construir el intervalo de confianza, pero que en este caso debemos utilizar el **valor nulo**, denotado p_0 en este contexto, como aproximación de p. Una vez verificadas ambas condiciones, el error estándar y el estadístico Z que permiten determinar el p-valor se calculan usando las ecuaciones 6.2 y 6.3, respectivamente.

$$SE_{\hat{p}} \approx \sqrt{\frac{p_0(1-p_0)}{n}} \tag{6.2}$$

$$Z = \frac{\hat{p} - p_o}{SE_{\hat{p}}} \tag{6.3}$$

Supongamos ahora, volviendo a nuestro ejemplo, que Baeza, entusiasmado por el intervalo de confianza obtenido anteriormente, le asegura a su jefe[1] que más del 70 % de las instancias de tamaño 100.000 se ejecutan en menos de 25 segundos. Sin embargo, su jefe no está convencido por lo que decide comprobarlo mediante una prueba de hipótesis con un nivel de significación $\alpha = 0,05$:

H_0: el 70 % de las instancias se ejecutan en menos de 25 segundos.
H_A: más del 70 % de las instancias se ejecutan en menos de 25 segundos.

De acuerdo a estas hipótesis del jefe de Baeza, el valor nulo es $p_0 = 0,7$, con lo que estas pueden formularse matemáticamente como:

Denotando como p a la proporción de todas las instancias de tamaño 100.000 que se ejecutan en menos de 25 segundos y considerando el valor hipotético $p_0 = 0,7$ *para este parámetro:*
H_0: p = p_0
H_A: p >p_0

Ya antes habíamos comprobado que se verifica la independencia de las observaciones. Además, considerando que el valor nulo fuese verdadero esperaríamos encontrar $0,7 \cdot 150 = 105$ éxitos y $(1-0,7) \cdot 150 = 45$ fracasos, ambos valores mayores que 10, por lo que la condición de éxito-fracaso se verifica.

Con ello, podemos calcular el estadístico de prueba:

$$SE_{\hat{p}} \approx \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0,7(1-0,7)}{150}} = 0,0374$$

$$Z = \frac{\hat{p} - p_0}{SE_{\hat{p}}} \approx \frac{0,64 - 0,7}{0,0374} = -1,6043$$

[1] Equivocadamente. Ver pregunta 1 al final del capítulo.

El valor p asociado, calculado en R mediante la llamada a la función `2 * pnorm(-1.6042)`, es $p = 0,109$. En consecuencia, la evidencia no es suficiente para rechazar la hipótesis nula, por lo que se concluye, con 95 % de confianza, que no es posible aceptar que el algoritmo se ejecute en menos de 25 segundos para más del 70 % de las instancias de tamaño 100.000.

R no ofrece esta prueba, como función. Sin embargo, podemos hacerla como muestra el script 6.1 para nuestro ejemplo.

Script 6.1: método de Wald para una proporción.

```r
# Fijar valores conocidos
n <- 150
p_exito <- 0.64
alfa <- 0.05
valor_nulo <- 0.7

# Construcción del intervalo de confianza.
error_est <- sqrt((p_exito * (1 - p_exito)) / n)
Z_critico <- qnorm(alfa / 2, lower.tail = FALSE)
inferior <- p_exito - Z_critico * error_est
superior <- p_exito + Z_critico * error_est
cat("Intervalo de confianza = [", inferior, ", ", superior, "]\n", sep = "")

# Prueba de hipótesis.
error_est_hip <- sqrt((valor_nulo * (1 - valor_nulo)) / n)
Z <- (p_exito - valor_nulo) / error_est_hip
p <- pnorm(Z, lower.tail = FALSE)
cat("Hipótesis alternativa unilateral\n")
cat("Z =", Z, "\n")
cat("p =", p)
```

6.1.2 Método de Wald para dos proporciones

También podemos usar el método de Wald para estudiar la **diferencia entre las proporciones** de dos poblaciones, considerando para ello como estimador puntual la diferencia $\hat{p}_1 - \hat{p}_2$.

De manera similar a lo que ya vimos para una única proporción, también en este caso debemos verificar ciertas condiciones antes de poder aplicar el modelo normal:

1. Cada proporción, por separado, sigue el modelo normal.
2. Las dos muestras son independientes una de la otra.

El error estándar para la diferencia entre dos proporciones muestrales está dado por la ecuación 6.4, donde p_1 y p_2 corresponden a las proporciones de las poblaciones, y n_1 y n_2, a los tamaños de las muestras. La construcción del intervalo de confianza se realiza, una vez más, con la ecuación general 4.6.

$$SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{SE_{\hat{p}_1}^2 + SE_{\hat{p}_2}^2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \tag{6.4}$$

A modo de ejemplo, supongamos que la Facultad de Ingeniería de una prestigiosa universidad desea determinar si la tasa de reprobación de estudiantes que rinden la asignatura de programación por primera vez es igual para hombres y mujeres. Para ello, se examina la situación final de los estudiantes que rindieron la asignatura durante el segundo semestre de 2017. Para una muestra de 48 hombres (de un total de 632), se encontró que 26 de ellos reprobaron la asignatura. De manera similar, para una muestra de 42 mujeres (de un total de 507), se encontraron 20 reprobaciones[2], con ambas muestras tomadas de manera aleatoria.

[2]Los datos aquí presentados son ficticios, creados únicamente con fines pedagógicos.

Como ya es habitual, comencemos por verificar las condiciones de normalidad para cada una de las muestras. En ambos casos, las observaciones son independientes entre sí, pues provienen de personas diferentes que representan a menos del 10 % de la población. Además, los datos entregados evidencian que en ambos casos se cumple la condición de éxito-fracaso. Adicionalmente, ambas muestras son independientes entre sí, pues ambas categorías se excluyen mutuamente. Con esto último se verifican entonces las condiciones de normalidad para la diferencia de proporciones.

Sean $\hat{p}_1$ y $\hat{p}_2$ las proporciones de éxito muestrales (considerando en este contexto la reprobación como éxito) para hombres y mujeres, respectivamente:

$$\hat{p}_1 = 26/48 = 0,5417$$

$$\hat{p}_2 = 20/42 = 0,4762$$

$$\hat{p}_1 - \hat{p}_2 = 0,5417 - 0,4762 = 0,0655$$

El error estándar puede estimarse como:

$$SE_{\hat{p}_1-\hat{p}_2} \approx \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = \sqrt{\frac{0,5417(1-0,5417)}{48} + \frac{0,4762(1-0,4762)}{42}} = 0,1054$$

Suponiendo un nivel de significación $\alpha = 0,05$, el intervalo de confianza corresponde a:

$$\hat{p}_1 - \hat{p}_2 \pm z^*_{0.975} SE_{\hat{p}_1-\hat{p}_2} \approx 0,0655 \pm 1,96 \cdot 0,1054 = [-0,1411; 0,2721]$$

En consecuencia, podemos afirmar con 95 % de confianza que la diferencia en la tasa de reprobación de la asignatura de programación para hombres y mujeres varía entre -14,11 % y 27,21 %.

Desde luego, también podemos realizar pruebas de hipótesis en este escenario. Para el ejemplo tenemos que:

H_0: no hay diferencia en la tasa de reprobación de hombres y mujeres.
H_A: las tasas de reprobación son diferentes para hombres y mujeres.

Matemáticamente:

Denotando como p_1 y p_2 a las proporciones de hombres y mujeres, respectivamente, que reprueban la asignatura de programación la primera vez que la cursan:
H_0: $p_1 - p_2 = 0$
H_A: $p_1 - p_2 \neq 0$

Ya verificamos las condiciones para operar bajo el supuesto de normalidad cuando construimos el intervalo de confianza. Sin embargo, **cuando la hipótesis nula supone que no hay diferencia entre las proporciones**, la verificación de la condición de éxito-fracaso y la estimación del error estándar se realizan usando para ello la **proporción agrupada**, dada por la ecuación 6.5, donde $\hat{p}_1 n_1$ y $\hat{p}_2 n_2$ representan la cantidad de éxitos en la primera y segunda muestra, respectivamente.

$$\hat{p} = \frac{\text{número de éxitos}}{\text{número de casos}} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2} \tag{6.5}$$

Así, en este caso tenemos:

$$\hat{p} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2} = \frac{0,5417 \cdot 48 + 0,4762 \cdot 42}{48 + 42} = 0,5111$$

En consecuencia, en el caso de los hombres esperamos encontrar $\hat{p}n_1 = 24,5328 > 10$ éxitos (reprobaciones) y $(1-\hat{p})n_1 = 23,4672 > 10$ fracasos. Del mismo modo, para las mujeres esperamos $\hat{p}n_2 = 21,4662 > 10$ éxitos y $(1-\hat{p})n_2 = 20,5338 > 10$ fracasos, con lo que se verifican las condiciones para emplear el modelo normal.

El error estándar se calcula, como ya mencionamos, usando la proporción agrupada:

$$SE_{\hat{p}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}} = \sqrt{\frac{0,5111 \cdot (1-0,5111)}{48} + \frac{0,5111 \cdot (1-0,5111)}{42}} = 0,1056$$

El estimador puntual para la diferencia es $\hat{p}_1 - \hat{p}_2 = 0,0655$, con lo cual el estadístico de prueba está dado por (recordando la f´):

$$Z = \frac{\hat{p} - p_0}{SE_{\hat{p}}} \approx \frac{0,0655 - 0}{0,1056} = 0,6203$$

En consecuencia, el valor p correspondiente es $p = 0,5351$. Puesto que el valor p es mayor que $\alpha = 0,05$, se falla en rechazar la hipótesis nula. Así, podemos decir con 95 % de confianza que no existe evidencia suficiente para concluir que hay diferencia en la tasa de reprobación de hombres y mujeres para el primer curso de programación.

El script 6.2 muestra el desarrollo de este ejemplo en R.

Script 6.2: método de Wald para la diferencia entre dos proporciones (ejemplo 1).

```r
# Fijar valores conocidos
n_hombres <- 48
n_mujeres <- 42
exitos_hombres <- 26
exitos_mujeres <- 20
alfa <- 0.05
valor_nulo <- 0

# Calcular probabilidades de éxito.
p_hombres <- exitos_hombres / n_hombres
p_mujeres <- exitos_mujeres / n_mujeres

# Estimar la diferencia.
diferencia <- p_hombres - p_mujeres

# Construcción del intervalo de confianza.
error_hombres <- (p_hombres * (1 - p_hombres)) / n_hombres
error_mujeres <- (p_mujeres * (1 - p_mujeres)) / n_mujeres
error_est <- sqrt(error_hombres + error_mujeres)
Z_critico <- qnorm(alfa / 2, lower.tail = FALSE)
inferior <- diferencia - Z_critico * error_est
superior <- diferencia + Z_critico * error_est
cat("Intervalo de confianza = [", inferior, ", ", superior, "]\n", sep = "")

# Prueba de hipótesis.
p_agrupada <- (exitos_hombres + exitos_mujeres) / (n_hombres + n_mujeres)
error_hombres <- (p_agrupada * (1 - p_agrupada)) / n_hombres
error_mujeres <- (p_agrupada * (1 - p_agrupada)) / n_mujeres
error_est_hip <- sqrt(error_hombres + error_mujeres)
Z <- (diferencia - valor_nulo) / error_est_hip
p <- 2 * pnorm(Z, lower.tail = FALSE)
cat("Hipótesis alternativa bilateral\n")
cat("Z =", Z, "\n")
cat("p =", p)
```

Cuando contrastamos hipótesis para la **diferencia entre dos proporciones con un valor nulo distinto de 0**, el procedimiento es ligeramente diferente. En este caso, la comprobación de la condición de éxito-fracaso se realiza de manera independiente para ambas muestras y el error estándar se calcula, como ya se estudió para los intervalos de confianza, mediante la ecuación 6.4.

Supongamos ahora que la Facultad de Ingeniería de la Universidad anterior ha decidido replicar el estudio realizado para el curso de programación, esta vez para una asignatura de física. No obstante, las autoridades están convencidas de que la tasa de reprobación es 10 % mayor para los hombres y que, incluso, la diferencia podría ser mayor. Desean comprobar con un nivel de confianza de 95 % y para ello, seleccionaron aleatoriamente a 89 de los 1.023 hombres y a 61 de las 620 mujeres de la cohorte correspondiente al primer semestre de 2019. En las muestras se encuentran, respectivamente, 45 y 21 reprobaciones.

Las hipótesis son, en este caso:

H_0: la tasa de reprobación de los hombres es exactamente 10 % más alta que la de las mujeres.
H_A: la tasa de reprobación de los hombres es más de 10 % más alta que la de las mujeres.

Matemáticamente:

Denotando como p_1 y p_2 a las proporciones de hombres y mujeres, respectivamente, que reprueban la asignatura de física estudiada la primera vez que la cursan:
H_0: $p_1 - p_2 = 0,1$
H_A: $p_1 - p_2 > 0,1$

Al igual que en los ejemplos previos, las observaciones de cada muestra son independientes entre sí pues corresponden a menos del 10 % de la población y fueron escogidos aleatoriamente. A su vez, los datos proporcionados indican que se cumple la condición de éxito-fracaso para cada muestra. Como ambas muestras pertenecen a grupos diferentes de estudiantes, son independientes entre sí. En consecuencia, se cumplen las condiciones para operar bajo el modelo normal.

En el caso de los hombres, la tasa de éxito se estima como:

$$\hat{p}_1 = \frac{45}{89} = 0,5056$$

Análogamente, para las mujeres tenemos:

$$\hat{p}_2 = \frac{21}{61} = 0,3443$$

Con lo que el estimador puntual para la diferencia es:

$$\hat{p}_1 - \hat{p}_2 = 0,5056 - 0,3443 = 0,1613$$

Ahora calculamos el error estándar:

$$SE_{\hat{p}_1 - \hat{p}_2} \approx \sqrt{\frac{0,5056(1-0,0,5056)}{89} + \frac{0,3443(1-0,3443)}{61}} = 0,0807$$

Con lo cual podemos calcular el estadístico de prueba:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - p_0}{SE_{\hat{p}_1 - \hat{p}_2}} \approx= \frac{0,1613 - 0,1}{0,0807} = 0,7596$$

Con lo que se puede obtener el valor p, correspondiente a $p = 0.2237 > \alpha = 0,05$.

En consecuencia, se falla en rechazar H_0 en favor de H_A, por lo que concluimos, con 95 % de confianza, que no es posible descartar que la tasa de reprobación de los hombres es 10 % superior a la de las mujeres para el curso de física.

En R, esta prueba puede realizarse como muestra el script 6.3.

Script 6.3: método de Wald para la diferencia entre dos proporciones (ejemplo 2).

```
# Fijar valores conocidos
n_hombres <- 89
n_mujeres <- 61
exitos_hombres <- 45
exitos_mujeres <- 21
alfa <- 0.05
valor_nulo <- 0.1

# Calcular probabilidades de éxito.
p_hombres <- exitos_hombres / n_hombres
p_mujeres <- exitos_mujeres / n_mujeres

# Estimar la diferencia.
diferencia <- p_hombres - p_mujeres

# Prueba de hipótesis.
p_agrupada <- (exitos_hombres + exitos_mujeres) / (n_hombres + n_mujeres)
error_hombres <- (p_hombres * (1 - p_hombres)) / n_hombres
error_mujeres <- (p_mujeres * (1 - p_mujeres)) / n_mujeres
error_est <- sqrt(error_hombres + error_mujeres)
Z <- (diferencia - valor_nulo) / error_est
p <- pnorm(Z, lower.tail = FALSE)
cat("Hipótesis alternativa bilateral\n")
cat("Z =", Z, "\n")
cat("p =", p)
```

6.2 MÉTODO DE WILSON

El método de Wald, tratado en la sección anterior, es el método que tradicionalmente se ha usado y el que aparece en la mayoría de los libros clásicos de inferencia estadística. Sin embargo, el método está siendo muy criticado hoy en día debido a que hace importantes simplificaciones matemáticas en su procedimiento y ya hay evidencia empírica que ha demostrado sus limitaciones (Agresti & Coull, 1998).

Gracias al aumento del poder de cómputo y la disponibilidad de software estadístico, han surgido diversas alternativas, entre las cuales destaca el **método de Wilson** (junto con algunas variaciones), considerado el más robusto por diversos autores (Agresti & Coull, 1998; Brown y col., 2001; Devore, 2008; Wallis, 2013). Este método opera del mismo modo que el de Wald, aunque introduce ajustes en la estimación de la proporción de éxito de la muestra y el error estándar de su distribución muestral cuando se estiman los intervalos de confianza. Específicamente, el intervalo de confianza para p con nivel de confianza aproximadamente de $100 \cdot (1 - \alpha)\,\%$ se calcula empleando las siguientes fórmulas:

$$p' = \frac{\hat{p} + \dfrac{z^2_{\alpha/2}}{2n}}{1 + \dfrac{z^2_{\alpha/2}}{n}} \tag{6.6}$$

$$SE'_{\hat{p}} = \frac{\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n} + \dfrac{z^2_{\alpha/2}}{4n^2}}}{1 + \dfrac{z^2_{\alpha/2}}{n}} \tag{6.7}$$

En R, podemos hacer esta prueba usando la función `prop.test(x, n, p, alternative, conf.level, ...)`, cuyos principales parámetros son:

- `x`: cantidad de éxitos en la muestra.
- `n`: tamaño de la muestra.
- `p`: valor nulo (por defecto `p=NULL`).
- `alternative`: tipo de hipótesis alternativa, por defecto bilateral (`alternative="two.sided"`), y valores `"less"` y `"greater"` para hipótesis unilaterales.
- `conf.level`: nivel de confianza (`conf.level=0.95` por defecto).

El script 6.4 muestra el uso de esta función con el mismo ejemplo que usamos para presentar la prueba de Wald para una proporción. Del mismo modo, el script 6.5 usa la función `prop.test()` para el primer ejemplo del método de Wald para la diferencia entre dos proporciones. Sin embargo, esta función tiene la limitante de que, al trabajar con dos proporciones, no permite establecer un valor nulo distinto de cero para la diferencia. Esto se debe a que no existe una prueba de hipótesis general que sea útil para todos los casos. Una lista de posibles métodos puede encontrarse en NCSS (2018).

Script 6.4: método de Wilson para una proporción.

```r
# Fijar valores conocidos
n <- 150
p_exito <- 0.64
alfa <- 0.05
valor_nulo <- 0.7

# Calcular cantidad de éxitos.
exitos <- p_exito * n

# Prueba de Wilson en R.
prueba <- prop.test(exitos, n = n, p = valor_nulo,
                    alternative = "greater", conf.level = 1 - alfa)

print(prueba)
```

Script 6.5: método de Wilson para la diferencia entre dos proporciones.

```r
# Fijar valores conocidos (hombres, mujeres)
n <-c(48, 42)
exitos <- c(26, 20)
alfa <- 0.05

# Prueba de Wilson en R.
prueba <- prop.test(exitos, n = n, alternative = "two.sided",
                    conf.level = 1 - alfa)

print(prueba)
```

6.3 EJERCICIOS PROPUESTOS

1. Considera el ejemplo usado en la sección 6.1.1. ¿Por qué es equivocado que el señor Baeza haya asegurado a su jefe que *más del 70 % de las instancias* de tamaño 100.000 se resuelven en menos de 25 segundos con su programa en base al intervalo de confianza obtenido: $[0,5632; 0.7168]$?
2. El patrón de un gran fundo de nogales está preocupado porque se ha detectado la presencia de una plaga en varios árboles. Si bien existe un pesticida para el parásito, este es bastante caro y su aplicación solo se justifica económicamente si más del 20 % de los árboles está infectado. En consecuencia, el patrón ha decidido estimar la extensión de la infestación revisando una muestra aleatoria de 200 nogales (una porción bastante pequeña de los más de 20.000 árboles en el fundo). En base a lo anterior, determina:
   a) ¿Cuál es la variable dicotómica (experimento Bernulli) en este caso?

   b) ¿Qué distribución sigue la variable en estudio? ¿Con qué parámetro(s) de interés?
   c) ¿Qué estimador existe(n) para este(os) parámetro(s)?
   d) ¿Qué hipótesis respondería las dudas del patrón del fundo?
3. En el experimento del ejercicio anterior se encontró que 45 árboles de la muestra estaban infectados:
   a) ¿Se puede asumir que esta proporción muestral sigue el modelo normal?
   b) Independientemente de la respuesta anterior, obtén un intervalo con 95 % confianza para la verdadera proporción de árboles infectados en el fundo.
   c) ¿Qué recomendarías al patrón del fundo?
4. Un laboratorio homeopático acaba de lanzar un tónico que asegura que ayuda a prevenir el resfrío durante el periodo invernal, con igual eficacia tanto en mujeres como en hombres. Para comprobar esta promesa, el laboratorio está realizando un estudio de la eficacia del producto en una muestra aleatoria de 100 mujeres y 200 hombres:
   a) ¿Cuál es la variable aleatoria en este caso y qué distribución sigue?
   b) ¿Qué hipótesis se deberían docimar para comprobar o refutar la homogeneidad de la eficacia del tónico para el resfrío?
5. El estudio anterior encontró que, durante las semanas de prueba, 38 mujeres y 102 hombres presentaron síntomas de resfrío. ¿Es homogénea la eficacia del producto con un nivel de significación de 0,05?
6. Da un ejemplo de una pregunta de investigación sobre las asignaturas comunes en ingeniería que requiera utilizar una prueba de hipótesis bilateral para una proporción. Explica bien qué variable se mide y enuncia las hipótesis a docimar, justificando el valor nulo escogido y el tipo de hipótesis alternativa (bilateral o unilateral).
7. Para tu ejemplo anterior, supón justificadamente un tamaño de la muestra obtenida y el número de éxitos observados. Usando la función `prop.test()`, obtén un intervalo con 99 % de confianza y el correspondiente valor p. ¿Cuál debería ser la conclusión?
8. Da un ejemplo de una pregunta de investigación sobre los conciertos realizados en Santiago que requiera una prueba de hipótesis bilateral para dos proporciones. Explica bien qué variable se mide y enuncia las hipótesis a docimar, justificando el valor nulo escogido.
9. Para tu ejemplo anterior, supón justificadamente los tamaños de las muestras obtenidas y el número de éxitos observados en cada una. Usando la función `prop.test()`, obtén un intervalo con 95 % de confianza y el correspondiente valor p. ¿Cuál debería ser la conclusión?
10. Da un ejemplo de una pregunta de investigación sobre el estado de la salud mental de estudiantes universitarios que requiera una prueba de hipótesis unilateral para dos proporciones. Explica bien qué variable se mide y enuncia las hipótesis a docimar, justificando el valor nulo escogido.
11. Para tu ejemplo anterior, supón justificadamente los tamaños de las muestras obtenidas y el número de éxitos observados en cada una. Usando la función `prop.test()`, obtén un intervalo con 95 % de confianza y el correspondiente valor p. ¿Cuál debería ser la conclusión?
12. Investiga para qué sirve y cómo funciona el parámetro `correct` de la función `prop.test()` de R.

6.4 BIBLIOGRAFÍA DEL CAPÍTULO

Agresti, A., & Coull, B. A. (1998).
Approximate is better than "exact" for interval estimation of binomial proportions.
The American Statistician, *52*(2), 119-126.

Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion.
Statistical science, *16*(2), 101-117.

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Anandkumar, A., Ford, C., Volcic, R., & de Rosario, H. (2020). *pwr: Basic Functions for Power Analysis.* Consultado el 1 de octubre de 2021, desde https://cran.r-project.org/web/packages/pwr/vignettes/pwr-vignette.html

Devore, J. L. (2008). *Probabilidad y Estadística para Ingeniería y Ciencias* (7.ª ed.). CENAGE Learning.

Diez, D., Barr, C. D., & Çetinkaya-Rundel, M. (2017). *OpenIntro Statistics* (3.ª ed.).
https://www.openintro.org/book/os/.

Kabacoff, R. I. (2017). *Power Analysis.*
Consultado el 1 de octubre de 2021, desde https://www.statmethods.net/stats/power.html

NCSS. (2018). *PASS Sample Size Software: Chapter 205 Non-Zero Null Tests for the Difference Between Two Proportions*. NCSS LLC. Consultado el 10 de septiembre de 2021, desde https://www.ncss.com/wp-content/themes/ncss/pdf/Procedures/PASS/Non-Zero_Null_Tests_for_the_Difference_Between_Two_Proportions.pdf

NIST/SEMATECH. (2013). *e-Handbook of Statistical Methods*. Consultado el 29 de abril de 2021, desde http://www.itl.nist.gov/div898/handbook/

Pértega, S., & Pita, S. (2004). *Asociación de variables cualitativas: El test exacto de Fisher y el test de Mcnemar*. Consultado el 29 de abril de 2021, desde https://www.fisterra.com/mbe/investiga/fisher/fisher.asp#Mcnemar

Wallis, S. (2013). Binomial confidence intervals and contingency tests: mathematical fundamentals and the evaluation of alternative methods. *Journal of Quantitative Linguistics*, *20*(3), 178-208.