



# 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:

- La muestra debe tener al menos 30 observaciones. Si la muestra tiene menos de 30 observaciones, se debe conocer la varianza de la población.
- 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.

Esta prueba resulta adecuada si queremos **asegurar** o **descartar** que la media de la población tiene un cierto **valor hipotético**. Supongamos que queremos saber si, en promedio, las utilidades mensuales de una pequeña empresa son de 20 millones de pesos y que el gerente general, Esteban Quito, nos ha informado que la desviación estándar para las utilidades es de 2,32 millones de pesos. El Sr. Quito nos ha proporcionado una muestra, obtenida mediante muestreo aleatorio simple, con las utilidades (en millones de pesos) reportadas para 20 meses, que se muestra en la tabla 5.1.

| Obs. | Utilidad [M\$] | Obs. | Utilidad [M\$] | Obs. | Utilidad [M\$] | Obs. | 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 mensuales de la empresa ( $\mu$ ) es de 20 millones de pesos, es decir:  $\mu = 20$  [M\$].

$H_A$ : las utilidades mensuales de la empresa 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. Si bien no estamos seguros que esta muestra considere menos del 10 % de las observaciones, podemos suponerlo razonablemente.

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. 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.

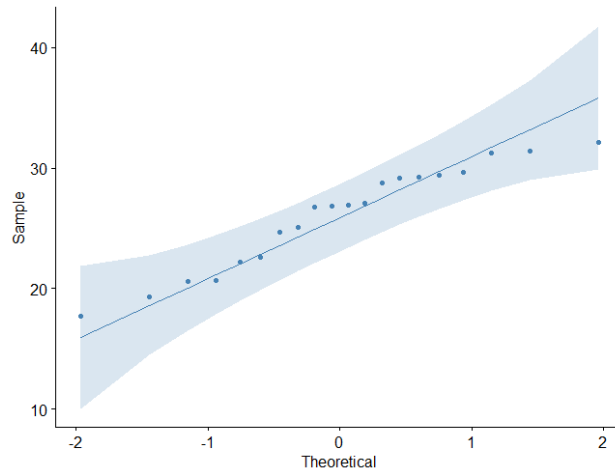


Figura 5.1: gráfico Q-Q para la muestra de la tabla 5.1.

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 3.7:

$$Z = \frac{\bar{x} - \mu}{\sigma} = \frac{26,066 - 20}{2,32} = 2.6147$$

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(2.6147, lower.tail = FALSE)`, obtenemos que  $p = 0,009 < 0.01^1$ , con lo que se rechaza la hipótesis nula en favor de la hipótesis alternativa. Sin embargo, debemos ser cuidadosos puesto que el valor p es bastante cercano al nivel de significación establecido, por lo que sería prudente evaluar los resultados con una muestra más grande. Así, concluimos que los datos **sugieren** que, en promedio, las utilidades mensuales de la empresa difieren de los 20 millones de pesos establecidos.

<sup>1</sup>Más precisamente,  $p = 0.008931758$ , pero, por convención, p-valores suelen reportarse con tres decimales.

Desde luego, gracias a R podemos realizar esta prueba simplemente con una llamada a la función `z.test(x, mu, stdev, alternative, conf.level)`, disponible en el paquete `TeachingDemos`, 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`” (hipótesis bilateral), “`less`” (hipótesis unilateral que la media de la población es menor que el valor nulo) o “`greater`” (hipótesis unilateral que la media de la población es mayor que el valor nulo).
- `conf.level`: nivel de confianza.

El script 5.1 muestra el desarrollo de este ejemplo en forma manual y luego, en la línea 42, 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 = 2.6147, n = 1.00, Std. Dev. = 2.32, Std. Dev. of the sample mean = 2.32, p-value = 0.008932
alternative hypothesis: true mean is not equal to 20
99 percent confidence interval:
 20.09008 32.04192
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
25 # Fijar un nivel de significación.
26 alfa <- 0.01
27
28 # Calcular la media de la muestra.
```

```

29 cat("\tPrueba Z para una muestra\n\n")
30 media <- mean(muestra)
31 cat("Media =", media, "M$\n")
32
33 # Calcular el estadístico de prueba.
34 Z <- (media - valor_nulo) / desv_est
35 cat("Z =", Z, "\n")
36
37 # Calcular el valor p.
38 p <- 2 * pnorm(Z, lower.tail = FALSE)
39 cat("p =", p, "\n")
40
41 # Hacer la prueba Z con R.
42 prueba <- z.test(media, mu = valor_nulo, alternative = "two.sided",
43                  stdev = desv_est, conf.level = 1-alfa)
44
45 print(prueba)

```

## 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 ( $\nu$ ) 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

Aunque la prueba t no opera bajo el supuesto de normalidad, aún así requiere verificar algunas condiciones para poder usarla:

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 limitan el tamaño de la muestra para que 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  $\mu_0$  es el valor nulo, por lo que en este caso  $\mu_0 = 500$  [ms]. Matemáticamente, las hipótesis anteriores pueden formularse como:

*Denotando como  $\mu$  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  $\bar{x} = 434,2921$ , y la desviación estándar,  $s = 38,0963$ .

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,  $\sigma$ ).

$$T = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \quad (5.1)$$

Así, para el ejemplo tenemos que:

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

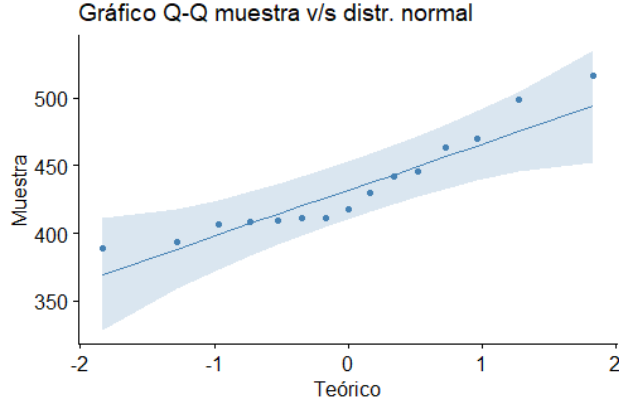


Figura 5.3: gráfico para comprobar el supuesto de normalidad.

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  $\alpha = 0,025$ ).

$$\bar{x} \pm t_{\nu}^* \cdot SE \quad (5.2)$$

Fijémonos en que en la ecuación 5.2 aparece el nuevo valor  $t_{\nu}^*$ , el cual se obtiene a partir del nivel de confianza y la distribución  $t$  con  $\nu$  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:

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

Así, el intervalo de confianza está dado por:

$$(-\infty, t_{nu}^* \cdot SE_{\bar{x}}] = (-\infty, 2,1448 \cdot 9,8364] = (-\infty, 455,3892]$$

Una vez más, R permite realizar esta prueba de manera rápida y sencilla, gracias a la función `t.test(x, alternative, mu, conf.level)`, donde:

- **x**: vector no vacío de valores numéricos (la muestra).
- **alternative**: tipo de prueba de hipótesis. Los posibles valores son “**two.sided**” (prueba bilateral), “**greater**” (hipótesis unilateral que la media de la población es mayor que el valor nulo) o “**less**” (hipótesis unilateral que la media de la población es menor que el valor nulo).
- **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 los resultados podemos observar que el valor  $p$  obtenido es muy pequeño, dando a entender que, si se cumple el supuesto de que la verdadera media es  $\mu = 500$  [ms] (hipótesis nula), sería muy improbable obtener una media muestral de  $\bar{x} = 434,2921$ . Además, el valor  $p$  es muchísimo menor que el nivel de significación, por lo que la evidencia a favor de  $H_A$  es muy fuerte. En consecuencia, se rechaza  $H_0$  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.

```
1 library(ggpubr)
2
3 # Cargar los datos.
4 tiempo <- c(411.5538, 393.2753, 445.8905, 411.4022, 498.8969,
5            388.6731, 430.0382, 469.4734, 409.5844, 442.0800,
6            418.1169, 408.4110, 463.3733, 407.0908, 516.5222)
7
8 # Establecer los datos conocidos.
9 n <- length(tiempo)
10 grados_libertad <- n - 1
11 valor_nulo <- 500
12
13
14 # Verificar si la distribución se acerca a la normal.
15 g <- ggqqplot(data = data.frame(tiempo),
16              x = "tiempo",
17              color = "steelblue",
18              xlab = "Teórico",
19              ylab = "Muestra",
20              title = "Gráfico Q-Q muestra v/s distr. normal")
21
22 print(g)
23
24 # Fijar un nivel de significación.
25 alfa <- 0.025
26
27 # Calcular el estadístico de prueba.
28 cat("\tPrueba t para una muestra\n\n")
29 media <- mean(tiempo)
30 cat("Media =", media, "M$\n")
31 desv_est <- sd(tiempo)
32 error <- desv_est / sqrt(n)
33 t <- (media - valor_nulo) / error
34 cat("t =", t, "\n")
35
36 # Calcular el valor p.
37 p <- pt(t, df = grados_libertad, lower.tail = TRUE)
38 cat("p =", p, "\n")
39
40 # Construir el intervalo de confianza.
41 t_critico <- qt(alfa, df = grados_libertad, lower.tail = FALSE)
42 superior <- media + t_critico * error
43 cat("Intervalo de confianza = (-Inf, ", superior, "]\n", sep = "")
44
45 # Aplicar la prueba t de Student con la función de R.
46 prueba <- t.test(tiempo,
47                  alternative = "less",
48                  mu = valor_nulo,
49                  conf.level = 1 - alfa)
50
51 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 **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 la diferencia entre cada par de observaciones, para lo cual podemos usar la técnica de la distribución t (también llamada prueba t de Student) vista en la sección anterior.

La media de las diferencias es  $\bar{x}_{dif} = -12,08591$  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  $\mu_{dif}$ :*

$H_0$ :  $\mu_{dif} = 0$

$H_A$ :  $\mu_{dif} \neq 0$

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 30) 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 39) 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.

```
1 # Cargar los datos.
2 instancia <- seq(1, 35, 1)
3
4 t_A <- c(436.5736, 470.7937, 445.8354, 470.9810, 485.9394,
5         464.6145, 466.2139, 468.9065, 473.8778, 413.0639,
6         496.8705, 450.6578, 502.9759, 465.6358, 437.6397,
7         458.8806, 503.1435, 430.0524, 438.5959, 439.7409,
8         464.5916, 467.9926, 415.3252, 495.4094, 493.7082,
9         433.1082, 445.7433, 515.2049, 441.9420, 472.1396,
10        451.2234, 476.5149, 440.7918, 460.1070, 450.1008)
11
12 t_B <- c(408.5142, 450.1075, 490.2311, 513.6910, 467.6467,
13         484.1897, 465.9334, 502.6670, 444.9693, 456.3341,
14         501.1443, 471.7833, 441.1206, 544.1575, 447.8844,
15         432.4108, 477.1712, 482.4828, 458.2536, 474.9863,
16         496.0153, 485.8112, 457.4253, 483.3700, 510.7131,
17         467.5739, 482.5621, 453.5986, 385.9391, 548.7884,
18         467.2533, 494.7049, 451.9716, 522.3699, 444.1270)
19
20 diferencia <- t_A - t_B
21
22 # Verificar si la distribución se acerca a la normal.
23 normalidad <- shapiro.test(diferencia)
24 print(normalidad)
25
26 # Fijar un nivel de significación.
27 alfa <- 0.05
28
29 # Aplicar la prueba t de Student a la diferencia de medias.
30 prueba_1 <- t.test(diferencia,
31                   alternative = "two.sided",
32                   mu = valor_nulo,
33                   conf.level = 1 - alfa)
34
35 print(prueba_1)
36
37 # Otra alternativa puede ser aplicar la prueba t de Student
38 # para dos muestras pareadas.
39 prueba_2 <- t.test(x = t_A,
40                   y = t_B,
```

```

41         paired = TRUE,
42         alternative = "two.sided",
43         mu = valor_nulo,
44         conf.level = 1 - alfa)
45
46 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  $\bar{x} = -12,08591$ .

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, nuevamente, el resultado está cerca del borde de significación. En consecuencia, se puede afirmar con 95 % de confianza que pareciera no haber diferencia entre los tiempos de ejecución de ambos algoritmos, 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  $(\bar{x}_1 - \bar{x}_2)$ . Así, el estadístico T en este caso toma la forma de la ecuación 5.3.

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

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:

| 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.

Si  $\mu_A$  y  $\mu_B$  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 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, en 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
17 # Fijar un nivel de significación.
18 alfa <- 0.01
19
20 # Aplicar la prueba t para dos muestras independientes.
21 prueba <- t.test(x = vacuna_A,
22                 y = vacuna_B,
23                 paired = FALSE,
24                 alternative = "greater",
25                 mu = 0,
26                 conf.level = 1 - alfa)
27
28 print(prueba)
29
30 # Calcular la diferencia entre las medias.
31 media_A <- mean(vacuna_A)
32 media_B <- mean(vacuna_B)
33 diferencia <- media_A - media_B
34 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  $n_2 - 1$ .

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{SE_{\bar{x}_1}^2 + SE_{\bar{x}_2}^2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad (5.4)$$

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.

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

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.

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

### 5.3 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. 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.
  - a) Enuncia las hipótesis nula y alternativa a contrastar.
  - b) Analiza si es razonable en este caso considerar que los datos cumplen las condiciones para usar una prueba t de Student.
  - c) Independientemente del resultado anterior, aplica la prueba propuesta y obtenga un intervalo de confianza y un valor p.
  - d) Usando un nivel de significación adecuado, entrega una conclusión para la cuestión planteada.
3. 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.
  - a) Como primer paso, el departamento de control de calidad necesita saber si la concentración media de concentraciones de ingredientes activos depende del catalizador elegido.
  - b) Propón las hipótesis nula y alternativa que permitan responder el problema planteado con una prueba t de Student.
  - c) Muestra que es razonable considerar que estos datos cumplen las condiciones para usar la prueba propuesta y fija un nivel de significación apropiado.
  - d) Aplica la prueba propuesta y obtenga un intervalo de confianza y un valor p.
  - e) ¿Cuál sería tu respuesta al departamento de control mencionado?
4. 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.
  - a) ¿Qué hipótesis nula y alternativa se deberían docimar<sup>2</sup> con una prueba t de Student para responder a la inquietud planteada?
  - b) ¿Cumplen los datos obtenidos las condiciones para usar esta prueba t de Student?
  - c) Aplicando la prueba t de Student para este caso, obtén un intervalo de confianza y un valor p.
  - d) ¿Qué aconsejarías a los directivos de la fábrica?

---

<sup>2</sup>Término que suele ocuparse en estadística como sinónimo de “probar”.





## REFERENCIAS

- Diez, D., Barr, C. D. & Çetinkaya-Rundel, M. (2017). *OpenIntro Statistics* (3.<sup>a</sup> ed.).  
<https://www.openintro.org/book/os/>.
- Kassambara, A. (2019). *Practical Statistics in R II - Comparing Groups: 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](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>