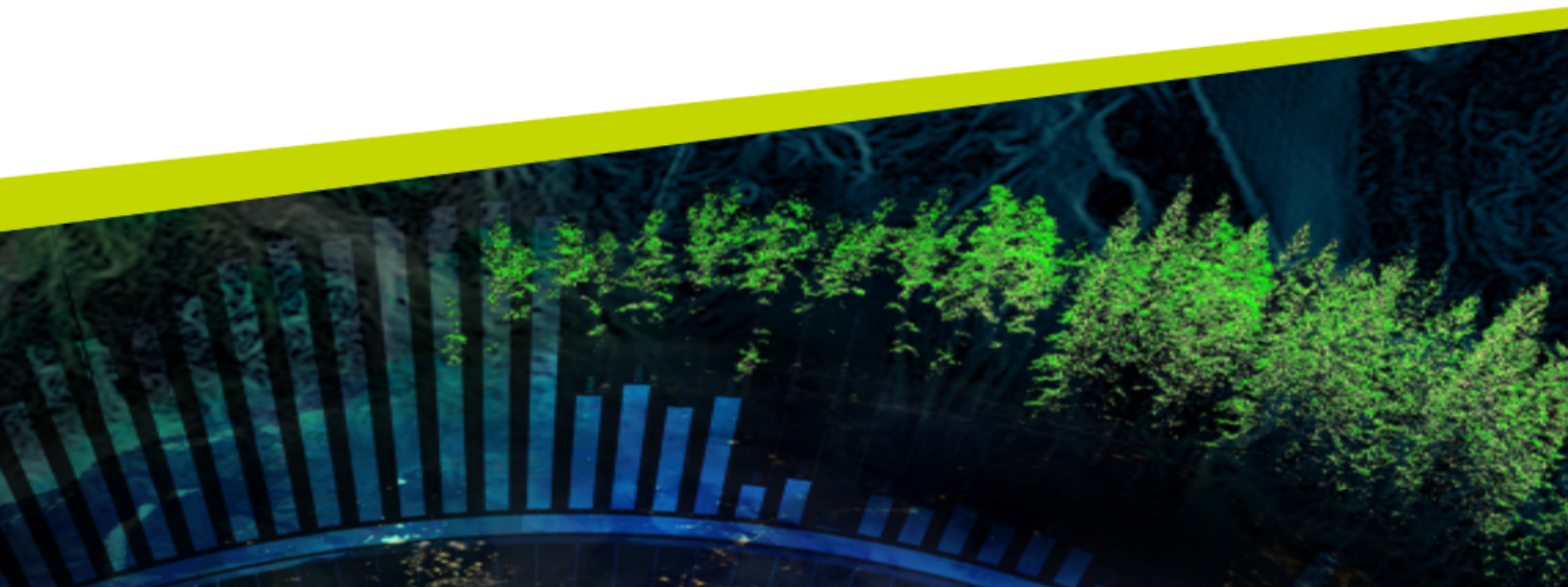
INFERENCIA Y MODELOS ESTADÍSTICOS

Jacqueline Köhler C. y José Luis Jara V.



CAPÍTULO 4. FUNDAMENTOS PARA LA INFERENCIA

En el capítulo 1 se definen los conceptos de población, entendido como todo el conjunto de interés, y muestra, que es un subconjunto de la población. También se introducen las nociones de parámetro, correspondiente a un valor que resume la población (por ejemplo la media de la población, μ), y de estadístico, como valor que resume una muestra (por ejemplo, la media muestral, $\overline{x}$). La **inferencia estadística** tiene por objeto entender cuán cerca está el estadístico del parámetro real de la población. En este capítulo conoceremos los principios necesarios para la inferencia estadística, con base en Diez y col. (2017, pp. 168-202) y Field y col. (2012, pp. 40-47).

4.1 ESTIMADORES PUNTUALES

Como ya dijimos, los parámetros y los estadísticos son valores que resumen, respectivamente, una población y una muestra. En consecuencia, podemos decir que un estadístico corresponde a un **estimador puntual** de un parámetro. El valor de un estimador puntual cambia dependiendo de la muestra que usemos para obtenerlo. Así, por más que su valor se acerque al parámetro de la población, difícilmente será igual a este último. Sin embargo, el estimador tiende a mejorar a medida que aumentamos el tamaño de la muestra, por efecto de la **ley de los grandes números**. Para ilustrar este fenómeno, consideremos la **media móvil**, que es una secuencia de medias muestrales en que cada una de ellas toma un elemento más de la población que su antecesora. La figura 4.1, elaborada con el script 4.1, ejemplifica este fenómeno.

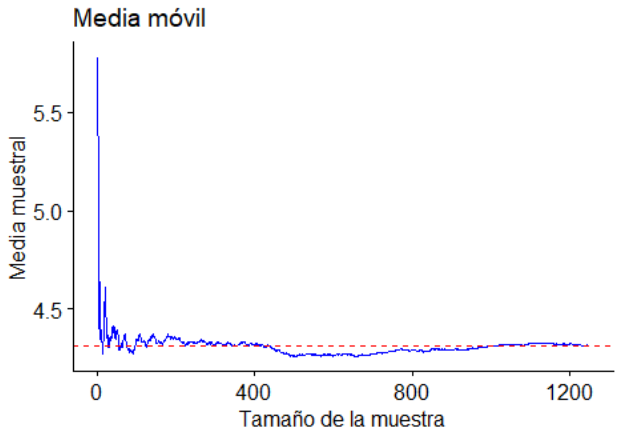


Figura 4.1: medias obtenidas al agregar a la muestra un elemento cada vez.

Script 4.1: representación gráfica de la media móvil.

```
library(ggpubr)

# Establecer la semilla para generar números aleatorios.
set.seed(9437)

# Generar aleatoriamente una población de tamaño 1500
# (en este caso, con una distribución cercana a la normal).
```

```r
poblacion <- rnorm(n = 1500, mean = 4.32, sd = 0.98)

# Calcular la media de la población.
media_poblacion <- mean(poblacion)
cat("Media de la población:", media_poblacion, "\n")

# Tomar una muestra de tamaño 1250.
tamano_muestra <- 1250
muestra <- sample(poblacion, tamano_muestra)

# Calcular las medias acumuladas (es decir, con muestras de
# 1, 2, 3, ... elementos).
n <- seq(along = muestra)
media <- cumsum(muestra) / n

# Crear una matriz de datos con los tamaños y las medias muestrales.
datos <- data.frame(n, media)

# Graficar las medias muestrales.
g <- ggline(data = datos,
            x = "n",
            y = "media",
            plot_type = "l",
            color = "blue",
            main = "Media móvil",
            xlab = "Tamaño de la muestra",
            ylab = "Media muestral")

# Añadir al gráfico una recta con la media de la población.
g <- g + geom_hline(aes(yintercept = media_poblacion),
                    color = "red", linetype = 2)

print(g)
```

Para determinar qué tan adecuado es un estimador, necesitamos saber cuánto cambia de una muestra a otra. Si esta variabilidad es pequeña, es muy probable que la estimación sea buena. Podemos estudiar la variabilidad de la muestra con ayuda de la **distribución muestral**, que representa la distribución de estimadores puntuales obtenidos con **todas** las diferentes muestras de igual tamaño de una misma población. La figura 4.2 (construida con el script 4.2) representa las medias para diferentes muestras de una población, aunque solo una selección aleatoria de todas las posibles muestras, incluyendo además una línea vertical roja que señala la media de la población. Podemos destacar que las medias muestrales tienden a aglutinarse en torno a la media poblacional, pues de acuerdo al **teorema del límite central**, la distribución de $\overline{x}$ se aproxima a la normalidad. Esta aproximación mejora a medida que aumenta el tamaño de la muestra.

Script 4.2: distribución de la media muestral.

```r
library(ggpubr)

# Establecer la semilla para generar números aleatorios.
set.seed(94)

# Generar aleatoriamente una población de tamaño 1500
# (en este caso, con una distribución cercana a la normal).
poblacion <- rnorm(n = 1500, mean = 4.32, sd = 0.98)

# Calcular la media de la población.
media_poblacion <- mean(poblacion)
cat("Media de la población:", media_poblacion, "\n")

```

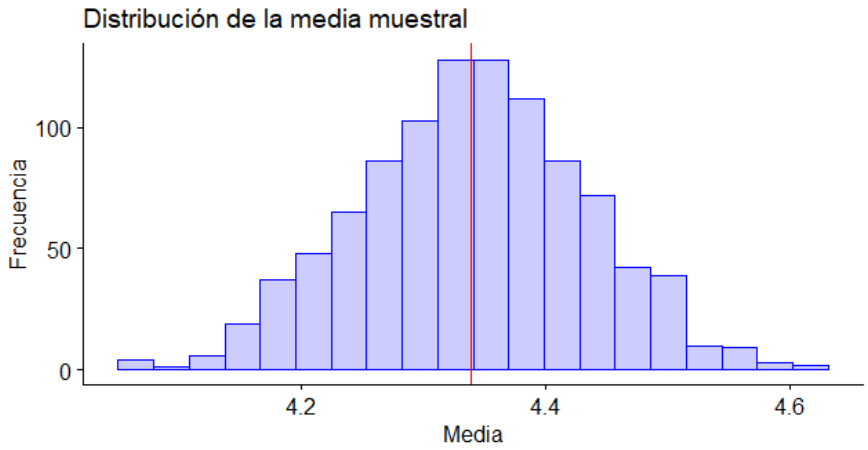


Figura 4.2: distribución muestral de la media para muestras con 100 observaciones.

```r
# Tomar 1000 muestras de tamaño 100. Quedan almacenadas
# como una matriz donde cada columna es una muestra.
tamano_muestra <- 100
repeticiones <- 1000

muestras <- replicate(repeticiones,
                      sample(poblacion, tamano_muestra))

# Calcular medias muestrales y almacenar los resultados
# en forma de data frame.
medias <- colMeans(muestras)
medias <- as.data.frame(medias)

# Construir un histograma de las medias muestrales.
g <- gghistogram(data = medias,
                 x = "medias",
                 bins = 20,
                 title = "Distribución de la media muestral",
                 xlab = "Media",
                 ylab = "Frecuencia",
                 color = "blue",
                 fill = "blue",
                 alpha = 0.2)

# Agregar línea vertical con la media de la población.
g <- g + geom_vline(aes(xintercept = media_poblacion),
                    color = "red", linetype = 1)

print(g)
```

4.2 MODELOS ESTADÍSTICOS

Ahora que hemos conocido más conceptos, podemos definir con precisión qué es un **modelo estadístico**. En el capítulo 1 dijimos que un modelo es simplemente una representación y que los modelos estadísticos pueden

emplearse para diversos propósitos:

- Describir o resumir datos.
- Clasificar objetos o predecir resultados.
- Anticipar los resultados de intervenciones (en ocasiones).

Más formalmente, un modelo estadístico es una descripción de un **proceso probabilístico** con **parámetros desconocidos** que deben ser **estimados** en base a **suposiciones** y un conjunto de datos **observados**. En general, tiene la forma dada en la ecuación 4.1:

$$y_i = (\text{modelo}) + \varepsilon_i \tag{4.1}$$

Donde:

- y_i es el i-ésimo valor observado de la variable respuesta Y (también llamada variable de salida o variable dependiente).
- modelo es el resultado de una función determinista basada en un conjunto de argumentos.
- ε_i es el error, correspondiente a la **variación natural**, y no a una equivocación, existente entre los valores observados y los valores pronosticados por el modelo. También recibe los nombres de variación no sistemática, variación aleatoria, residuos o incluso, residuales.

El error ε_i en la ecuación 4.1 se relaciona entonces con la calidad del modelo. Mientras menor sea el error, mejor será el modelo. Por el contrario, un error grande es señal de un modelo fallido, que no describe bien los datos, no ayuda a predecirlos bien, o no ayuda a su correcta clasificación.

La media y la proporción, y cualquier estadístico en general, son, en sí mismos, modelos estadísticos, aunque bastante simples.

4.3 ERROR ESTÁNDAR

En el capítulo 2 conocimos la desviación estándar como medida que estima la distancia de las observaciones respecto de la media. El **error estándar**, denotado usualmente por $SE_{\hat{\theta}}$ o $\sigma_{\hat{\theta}}$, corresponde a la desviación estándar de la distribución de un estimador muestral $\hat{\theta}$ de un parámetro θ. Por ejemplo, el error estándar de la media, es decir la desviación estándar de la distribución de las medias de todas las posibles muestras de n observaciones independientes, se calcula de acuerdo a la ecuación 4.2.

$$SE_{\overline{x}} = \frac{s}{\sqrt{n}} \tag{4.2}$$

Donde s es la desviación estándar de la muestra (ecuación 2.3) y n corresponde al tamaño de la muestra. En esta ecuación queda en evidencia que el error estándar de la media disminuye a medida que el tamaño de la muestra aumenta. Un método confiable que podemos usar para asegurar que las observaciones sean independientes es realizar un muestreo aleatorio simple[1] que abarque menos del 10 % de la población.

Volviendo a la ecuación para calcular el error estándar de la media muestral (equación 4.2), ¡debemos tener cuidado antes de usarla! Ya hemos mencionado antes que la distribución de las medias muestrales tiende a ser cercana a la normal, por lo que en dicho caso es posible usar el **modelo normal**, sustentado en el teorema del límite central. Las condiciones que deben cumplirse para usar este modelo y que, en consecuencia, el error estándar sea preciso, son:

1. Las observaciones de la muestra son independientes.

[1] Es decir, una muestra en que todos los elementos de la población tengan igual probabilidad de ser escogidos. Las técnicas de muestreo se abordan con más detalle en capítulos posteriores.

2. La muestra es grande (en general $n \geq 30$).
3. La distribución de la muestra no es significativamente asimétrica. Esto último suele además relacionarse con la presencia de valores atípicos. Mientras mayor sea el tamaño de la muestra, más se puede relajar esta condición.

Si no se cumplen las condiciones anteriores, debemos considerar otras opciones: para muestras pequeñas, se deben considerar métodos alternativos, y si la distribución de la muestra presenta una asimetría significativa, entonces tendremos que incrementar el tamaño de la muestra para compensar el efecto de la desviación.

4.4 INTERVALOS DE CONFIANZA

Hasta ahora sabemos que un estimador puntual es un único valor (obtenido a partir de una muestra) que, como su nombre indica, estima un parámetro de la población. Por ende, dicho valor rara vez es exacto. En consecuencia, lo lógico sería establecer un rango de valores plausibles para el parámetro estimado, que llamaremos **intervalo de confianza**, y que se construye en torno al estimador puntual. Dado que el error estándar representa la desviación estándar asociada al estimador, tiene sentido que lo usemos como guía en este proceso.

Recordemos que en el capítulo 3 vimos una regla empírica para la distribución normal (figura 3.5), la cual señala que (para distribuciones normales) alrededor de 95 % de las veces el estimador puntual se encontrará en un rango de 2 errores estándar del parámetro. Es decir, al considerar un intervalo de confianza de dos errores estándar (4.3), tendremos 95 % de **confianza** de haber capturado el parámetro real.

$$\overline{x} \pm 2 \cdot SE_{\overline{x}} \tag{4.3}$$

Podemos generalizar la ecuación 4.3 para calcular el intervalo de confianza para la media con cualquier **nivel de confianza** como muestra la ecuación 4.4.

$$\overline{x} \pm z^* \cdot SE_{\overline{x}} \tag{4.4}$$

El término z^* en la ecuación 4.4 corresponde, usualmente, al valor z tal que el área bajo la curva normal estándar comprendida entre $-z^*$ y z^* es igual al nivel de confianza deseado. La expresión $z^* \cdot SE$ recibe el nombre de **margen de error**.

Tomemos como ejemplo un **nivel de confianza** (que, por razones que veremos en la sección siguiente, denotaremos por $1 - \alpha$) de 90 % (es decir, $1 - \alpha = 0,9$). Eso significa, entonces, que nuestro intervalo de confianza excluye el 5 % del área correspondiente a la cola inferior (es decir, el percentil con valor 0,05) e igual porcentaje del área correspondiente a la cola superior (que, como la distribución Z es simétrica, es igual al área anterior). Puesto que conocemos el percentil, $(1 - \alpha)/2 = 0,05$, en R podemos usar la llamada `qnorm(0.05, mean = 0, sd = 1, lower.tail = FALSE)` y obtenemos $z^* = 1,64$. Es importante indicar que en esta llamada estamos en realidad trabajando con la cola superior para que z^* sea positivo. Si hacemos la llamada para la cola inferior, obtenemos $z^* = -1,64$.

Es importante destacar que, una vez más, debemos ser cuidadosos al interpretar un intervalo de confianza del x % (x $= 1 - \alpha$). Su significado es, sencillamente, "se tiene x % de certeza de que el parámetro de la población se encuentra entre..." (Diez y col., 2017, p. 180), es decir, que, en promedio, x % de los intervalos de confianza que se construyan en torno a un estadístico, con muestras de un tamaño fijo, capturarán el verdadero valor del parámetro. Esto **no es equivalente** a decir que el valor del parámetro tiene una "probabilidad de x %" de estar entre los valores del intervalo calculado, lo que sería incorrecto. Por otra parte, los intervalos de confianza no dicen nada acerca de observaciones individuales, sino que solo hablan del parámetro en cuestión.

4.5 PRUEBAS DE HIPÓTESIS

Supongamos que un banco ha desarrollado un nuevo sistema computacional para gestionar sus transacciones. El nuevo sistema (N) se ha puesto a prueba durante un mes, funcionando (con iguales condiciones de hardware) en paralelo con el sistema antiguo (A) y el banco ha llevado un registro del tiempo que tarda cada sistema en efectuar cada transacción. El gerente ha determinado que autorizará la migración al nuevo sistema únicamente si este es más rápido que el antiguo para procesar las transacciones. Se sabe que el sistema antiguo tarda en promedio $\mu_A = 530$ milisegundos en procesar una transacción. Para el sistema nuevo se han registrado $n = 1.600$ transacciones, realizadas en un tiempo promedio de $\overline{x}_N = 527,9$ [ms] con desviación estándar $s_N = 48$ [ms].

Una primera aproximación para tomar la decisión puede ser investigar si existe diferencia en los tiempos de ejecución de ambos sistemas, lo que puede expresarse en torno a dos **hipótesis** (palabra que la Real Academia Española (2014) define como "Suposición de algo posible o imposible para sacar de ello una consecuencia") que compiten entre sí:

H_0: El nuevo sistema, en promedio, tarda lo mismo que el antiguo en procesar las transacciones, es decir: $\mu_N = \mu_A$.
H_A: Los sistemas requieren, en promedio, cantidades de tiempo diferentes para procesar las transacciones, es decir: $\mu_N \neq \mu_A$

La primera hipótesis, H_0, recibe el nombre de **hipótesis nula** y suele representar una postura escéptica, es decir, que no hay cambios, por lo que **¡la hipótesis nula siempre se formula como una igualdad!**. La segunda (H_A), llamada **hipótesis alternativa**, representa en cambio una nueva perspectiva. Esta primera aproximación corresponde a una **prueba bilateral** o de dos colas, pues la diferencia puede ser en ambos sentidos: H_0 no parece correcta si $\mu_N < \mu_A$ o si $\mu_N > \mu_A$.

Como en este caso conocemos el valor de $\mu_A = 530$ [ms], también podríamos escribir la formulación matemática de las hipótesis de la siguiente manera:

H_0: $\mu_N = 530$
H_A: $\mu_N \neq 530$

En este planteamiento, "530" recibe el nombre de **valor nulo**, pues representa el valor del parámetro cuando se cumple la hipótesis nula.

Una aproximación más cercana al problema descrito puede ser investigar si el nuevo sistema es efectivamente **más rápido** que el antiguo. En este caso, se habla de una **prueba unilateral** o de una cola, pues solo interesa saber si el tiempo promedio empleado por el nuevo sistema es menor que el empleado por el sistema antiguo. Las hipótesis, en este caso, serían:

H_0: El nuevo sistema tarda, en promedio, lo mismo que el antiguo en procesar las transacciones, es decir: $\mu_N = \mu_A$.
H_A: El nuevo sistema tarda, en promedio, menos que el antiguo en procesar las transacciones, es decir: $\mu_N < \mu_A$

Obviamente en otros casos podría interesar solamente si valor alternativo es mayor que el valor nulo.

Teniendo las hipótesis planteadas, sigue decidir si la hipótesis nula parece o no plausible a través de una **prueba de hipótesis**. El marco para la prueba de hipótesis es **escéptico**: no se rechaza la hipótesis nula a menos que haya suficiente evidencia para rechazarla en favor de la hipótesis alternativa. Esta idea es muy parecida a la expresada en la expresión de uso común "se presume inocente mientras no se demuestre lo contrario". Sin embargo, el que no se logre rechazar H_0 **no significa aceptarla** como verdadera o como correcta sin más. Por eso se usa un lenguaje bastante peculiar, señalando que *se falla al rechazar* H_0 o bien que *se rechaza* H_0 *en favor de* H_A. Retomando la analogía con la expresión anterior, que no haya pruebas suficientes para la culpabilidad, no significa que una persona sea en verdad inocente.

Volvamos al escenario del ejemplo para la prueba de hipótesis bilateral (es decir, aquella en que solo queremos

ver si hay diferencias en el tiempo de procesamiento de transacciones entre ambos sistemas del banco). El valor de $\overline{x}_N = 527,9$ [ms] es, en efecto, distinto de $\mu_A = 530$ [ms]. No obstante, al ser una estimación puntual, como ya hemos estudiado, esta diferencia podría deberse simplemente a la muestra escogida, por lo que el parámetro real μ_N podría ser igual a μ_A [ms]. En consecuencia, resulta útil calcular el intervalo de confianza para $\overline{x}_N$.

Comencemos por determinar el error estándar:

$$SE_{\overline{x}} = \frac{s_N}{\sqrt{n}} = \frac{48}{\sqrt{1600}} = 1,2$$

Ahora fijemos un nivel de confianza, por ejemplo 95 %, y usemos el valor z^* correspondiente para calcular el intervalo de confianza:

$$\overline{x}_N \pm z^* \cdot SE_{\overline{x}} = 527,9 \pm 1,96 \cdot 1,2 = [525,548; 530,252]$$

Como el parámetro del sistema antiguo ($\mu_A = 530$ [ms]) cae (a penas) dentro de este intervalo, se puede suponer que no existe una diferencia significativa entre los tiempos promedio requeridos por ambos sistemas, por lo que no se rechaza H_0. Así, tenemos un 95 % de confianza en que no existe una diferencia entre los tiempos que requieren ambos sistemas para procesar transacciones. Sin embargo, esta decisión es un tanto apresurada ya que el resultado está cerca del borde de rechazo y, en este caso, lo lógico sería investigar más (hacer crecer la muestra).

Revisemos ahora el caso planteado con hipótesis alternativa unilateral (es decir, queremos ver si el nuevo sistema es, en efecto, más rápido). Manteniendo nuestro nivel de confianza $1-\alpha = 0,95$, en este caso debemos considerar los valores menores a $\mu_A = 530$ [ms] para el cálculo de z^*. En otras palabras, el 5 % que descartamos corresponde únicamente a la cola superior. Así, nuestro valor para z^* está dado por la llamada `qnorm(0.05, mean = 0, sd = 1, lower.tail = FALSE)`, obteniéndose $z^* = 1,64$ (aprox.) por lo que se tiene que la cota superior es:

$$\overline{x}_N - z^* \cdot SE_{\overline{x}} = 527,9 - 1,64 \cdot 1,2 = 529.874$$

Luego, el intervalo de confianza va desde "cualquier valor" bajo la media observada en la muestra hasta el valor calculado arriba, por lo que el intervalo con 95 % confianza sería: $[-\infty; 529,874]$.

Ahora el valor $\mu_A = 530$ [ms] cae (apenas) fuera del intervalo y podemos decir que existe evidencia de que el nuevo sistema tarda en promedio menos tiempo que el antiguo en procesar las transacciones.

Ahora bien, siempre que se prueban hipótesis podemos cometer un error al momento de decidir si rechazar o no la hipótesis nula. Afortunadamente, la estadística ofrece herramientas para cuantificar cuán frecuentes son dichos errores. Existen cuatro posibles escenarios, los cuales se presentan en la tabla 4.1. El **error tipo I** corresponde a rechazar H_0 cuando en realidad es verdadera, mientras que el **error tipo II** corresponde a no rechazarla cuando en realidad H_A es verdadera.

		Conclusión de la prueba	
		No rechazar H_0	Rechazar H_0 en favor de H_A
Verdad	H_0 verdadera	Decisión correcta	Error tipo I
	H_A verdadera	Error tipo II	Decisión correcta

Tabla 4.1: posibles escenarios para una prueba de hipótesis.

Como ya hemos señalado, la prueba de hipótesis se basa en no rechazar H_0 a menos que se tenga evidencia contundente. Por regla general, no se desea cometer el error de rechazar incorrectamente la hipótesis nula (error tipo I) en más de 5 % de los casos. Esto corresponde a un **nivel de significación** de 0,05, denotado por $\alpha = 0,05$. Si usamos un intervalo de confianza de 95 % para evaluar una prueba de hipótesis en que la

hipótesis nula es verdadera, cometeremos un error tipo I cada vez que el estimador puntual esté a 1,96 o más errores estándar del parámetro de la población. Esto puede ocurrir un 5 % de las veces (2,5 % en cada cola de la distribución para el caso bilateral). Del mismo modo, un intervalo de confianza del 99 % es equivalente a un nivel de significación $\alpha = 0,01$.

El intervalo de confianza es de mucha ayuda para decidir si rechazar o no H_0. No obstante, no aporta información directa acerca de cuán fuerte es la evidencia para la decisión tomada.

4.5.1 Prueba formal de hipótesis con valores p

Antes de que la computación se hiciera masiva, las personas tenían dos procedimientos posibles para decidir una prueba de hipótesis. El primero es el realizado en la sección anterior, esto es, calcular el intervalo con $(1 - \alpha)$ % de confianza de acuerdo a los estadísticos observados en una muestra y revisar si el valor nulo cae o no dentro de este intervalo. El otro procedimiento clásico, que podemos encontrar en muchos libros y sitios en Internet, es estimar a qué valor z corresponde la media observada en la distribución normal estandarizada que define el valor nulo y el error estándar: si este estadístico z es mayor que z^*, entonces el estadístico cae en una "zona de rechazo" de H_0; en caso contrario ($|z| < z^*$), se falla en rechazar la hipótesis nula.

Si bien estos procedimientos siguen siendo útiles, su diseño respondía a la existencia de **tablas de probabilidad** en que se tabulaban probabilidades para algunos valores de percentiles de uso común, como 90 %, 95 %, 0,975 % o 0,99 %.

Con la llegada de los computadores, y en particular de entornos como R, es posible obtener probabilidades (casi) exactas para cualquier percentil. Esto hizo que un tercer método para decidir una prueba de hipótesis haya ido ganando popularidad: el uso del **valor p**, también llamado **p-valor**, que es definido por Diez y col. (2017, p. 186) como "la probabilidad de observar datos al menos tan favorables como la muestra actual para la hipótesis alternativa, si la hipótesis nula es verdadera". De esta forma, un p-valor permite cuantificar cuán fuerte es la evidencia en contra de la hipótesis nula (y en favor de la hipótesis alternativa).

Consideremos ahora el escenario de la hipótesis unilateral del ejemplo, con un nivel de significación $\alpha = 0,05$, bajo el supuesto de que H_0 es verdadera y que la muestra a su vez tiene una distribución cercana a la normal. Recordemos que $\overline{x}_N = 527,9$ [ms] y $s_N = 48$ [ms] en $n = 1600$ observaciones. Esta distribución se vería como muestra la figura 4.3, creada mediante el script 4.3.

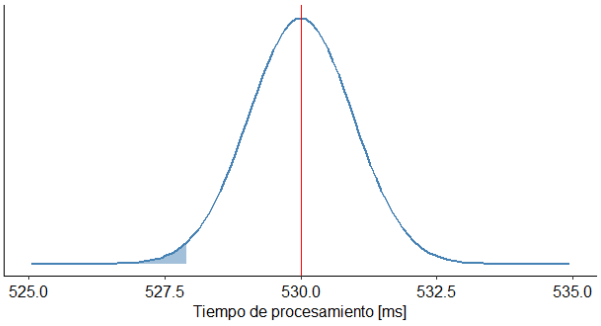


Figura 4.3: probabilidad de encontrar una media igual o menor que $\overline{x} = 527,9$ [ms] en la distribución muestral con $\mu_{\overline{x}} = 530$ y $\sigma_{\overline{x}} = 1,2$.

En este punto, resulta importante hacer una aclaración en relación al valor p. El área bajo la sección de la curva con valores menores o iguales a un estimador se calcula usando para ello el **valor z**, definido en la ecuación 4.5, como **estadístico de prueba**.

$$z = \frac{\text{estimador puntual} - \text{valor nulo}}{SE_{\text{estimador puntual}}} = \frac{\hat{\theta} - \theta_0}{SE_{\hat{\theta}}} \tag{4.5}$$

Un **estadístico de prueba** es un estadístico de resumen que resulta especialmente útil para evaluar hipótesis o calcular el valor p. El valor z se usa cuando el estimador puntual se acerca a la normalidad, aunque existen otros estadísticos de prueba adecuados para otros escenarios.

Script 4.3: cálculo del valor p para una prueba de una cola.

```
library(ggpubr)

# Generar una muestra donde la media cumpla con la hipótesis nula.
set.seed(872)

media_poblacion_antiguo <- 530
media_muestra_nuevo <- 527.9
desv_est <- 48
n <- 1600
error_est <- desv_est / sqrt(n)

x <- seq(media_poblacion_antiguo - 5.2 * error_est,
         media_poblacion_antiguo + 5.2 * error_est,
         0.01)

y <- dnorm(x, mean = media_poblacion_antiguo, sd = error_est)

datos <- data.frame(x, y)

# Graficar la muestra.
g <- ggplot(data = datos, aes(x))

g <- g + stat_function(fun = dnorm,
                       args = list(mean = media_poblacion_antiguo,
                                   sd = error_est),
                       colour = "steelblue", size = 1)

g <- g + ylab("")
g <- g + scale_y_continuous(breaks = NULL)
g <- g + scale_x_continuous(name = "Tiempo de procesamiento [ms]")
g <- g + theme_pubr()

# Colorear el área igual o menor que la media observada.
g <- g + geom_area(data = subset(datos,
                                 x < media_muestra_nuevo),
                   aes(y = y),
                   colour = "steelblue",
                   fill = "steelblue",
                   alpha = 0.5)

# Agregar una línea vertical para el valor nulo.
g <- g + geom_vline(aes(xintercept = media_poblacion_antiguo),
                    color = "red", linetype = 1)

print(g)

# Calcular el valor Z para la muestra.
Z <- (media_muestra_nuevo - media_poblacion_antiguo) / error_est

```

```r
# Calcular el valor p.
p_1 <- pnorm(Z, lower.tail = TRUE)

cat("Valor p: ", p_1, "\n")

# También se puede calcular el valor p directamente a partir de la
# distribución muestral definida por el valor nulo y el error
# estándar.
p_2 <- pnorm(media_muestra_nuevo, mean = media_poblacion_antiguo,
             sd = est_err)

cat("Valor p: ", p_2)
```

El valor p, en este caso $p = 0,040$, corresponde al área coloreada en la figura 4.3, y se calcula en la línea 51 del script 4.3. Esto nos indica, en este caso, que si H_0 fuera verdadera y el nuevo sistema tarda en promedio lo mismo que el antiguo en procesar las transacciones, la probabilidad de encontrar una media de a lo más 527,9 [ms] para una muestra de 1.600 transacciones es de 4 %, lo que sería bastante poco frecuente.

Cuanto menor sea el valor p, más fuerte será la evidencia en favor de H_A por sobre H_0. Y aquí la ventaja de usar este método para decidir: el valor p se puede **comparar directamente** con el nivel de significación α, y si p es menor que el nivel de significación se considera evidencia suficiente para rechazar la hipótesis nula en favor de la hipótesis alternativa. En este ejemplo, $p = 0,040 < \alpha = 0,05$, por lo que se falla al rechazar H_0 en favor de H_A. Pero como se dijo cuando usamos intervalos de confianza, el valor p está cerca del valor α y convendría ser menos tajante en la decisión y evaluar la posibilidad de ampliar la muestra para conseguir evidencia más definitiva.

Siempre es recomendable formular la conclusión de la prueba de hipótesis en lenguaje llano, para facilitar su comprensión. Así, en este caso concluimos que los datos sugieren que el nuevo sistema tarda menos que el antiguo en procesar transacciones, pero que es necesario hacer un estudio con más observaciones para tener un diagnóstico más definitivo.

Volvamos nuevamente al escenario de la prueba de hipótesis bilateral para el ejemplo, manteniendo el nivel de significación $\alpha = 0,05$. Puesto que en este caso nos interesa la diferencia en ambas direcciones, ya que la evidencia en ambas direcciones es favorable para H_A, debemos considerar el área bajo las dos colas de la curva normal, a diferencia del caso de la prueba de hipótesis unilateral en que solo se consideramos la cola correspondiente a la dirección de interés de la diferencia. Dado que el modelo normal es simétrico, el área bajo ambas colas es la misma (figura 4.4, script 4.4). El valor p, entonces, ahora es igual a dos veces el área de la cola inferior, es decir, $p = 0,080$. Puesto que $p > \alpha$, se falla en rechazar H_0. Es decir, no hay evidencia suficiente para concluir que existe una diferencia entre los tiempos promedio requeridos por ambos sistemas para procesar transacciones.

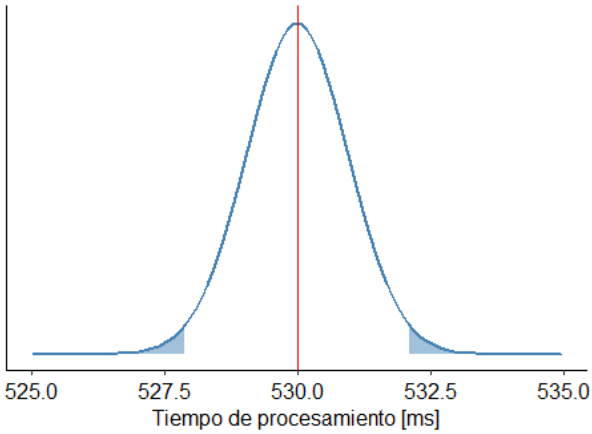


Figura 4.4: cuando la prueba de hipótesis es bilateral, se deben colorear ambas colas.

Script 4.4: cálculo del valor p para una prueba de dos colas.

```
library(ggpubr)

# Generar una muestra donde la media cumpla con la hipótesis nula.
set.seed(208)

media_poblacion_antiguo <- 530
media_muestra_nuevo <- 527.9
desv_est <- 48
n <- 1600
error_est <- desv_est / sqrt(n)

x <- seq(media_poblacion_antiguo - 5.2 * error_est,
         media_poblacion_antiguo + 5.2 * error_est,
         0.01)

y <- dnorm(x,
           mean = media_poblacion_antiguo,
           sd = error_est)

dataframe <- data.frame(x, y)

# Graficar la muestra.
g <- ggplot(data = dataframe, aes(x))

g <- g + stat_function(fun = dnorm,
                       args = list(mean = media_poblacion_antiguo,
                                   sd = error_est),
                       colour = "steelblue", size = 1)

g <- g + ylab("")
g <- g + scale_y_continuous(breaks = NULL)
g <- g + scale_x_continuous(name = "Tiempo de procesamiento [ms]")
g <- g + theme_pubr()

# Colorear el área igual o menor que la media observada.
g <- g + geom_area(data = subset(dataframe,
                                 x < media_muestra_nuevo),
                   aes(y = y),
                   colour = "steelblue",
                   fill = "steelblue",
                   alpha = 0.5)

# Calcular el área bajo la cola inferior.
area_inferior <- pnorm(media_muestra_nuevo,
                       mean = media_poblacion_antiguo,
                       sd = desv_est)


# Colorear igual área en la cola restante.
corte_x <- qnorm(1 - area_inferior,
                 mean = media_poblacion_antiguo,
                 sd = desv_est)

g <- g + geom_area(data = subset(dataframe,
                                 x > corte_x),
                   aes(y = y),
                   colour = "steelblue",
                   fill = "steelblue",
```

```
                    alpha = 0.5)

# Agregar una línea vertical para el valor nulo.
g <- g + geom_vline(aes(xintercept = media_poblacion_antiguo),
                    color = "red", linetype = 1)

print(g)

# Calcular el valor Z para la muestra.
Z <- (media_muestra_nuevo - media_poblacion_antiguo) / error_est

# Calcular el valor p (recordando ahora que la hipótesis es bilateral).
p <- 2 * pnorm(Z, lower.tail = TRUE)

cat("Valor p: ", p)
```

Un punto importante que debemos tener en cuenta es que **las pruebas unilaterales** se usan cuando se desea verificar un incremento o un decremento, pero no ambas. No obstante, esta decisión debe tomarse siempre **antes de examinar los datos**, pues de lo contrario se duplica la probabilidad de cometer errores de tipo I y se está cayendo en **prácticas poco éticas**.

4.5.2 El efecto del nivel de significación

Hemos visto que el nivel de significación (α) representa la proporción de veces en que se cometería un error de tipo I (es decir, rechazar H_0 en favor de H_A, cuando H_0 es en realidad verdadera). Si resulta costoso o peligroso cometer un error de este tipo, debemos requerir evidencia más fuerte para rechazar la hipótesis nula (es decir, reducir la probabilidad de que esto ocurra), lo que podemos lograr usando un valor más pequeño para el nivel de significación, por ejemplo, $\alpha = 0,01$. Sin embargo, esto necesariamente **aumentará** la probabilidad de cometer un error de tipo II.

Si, por el contrario, el costo o el peligro de cometer un error de tipo II (no rechazar H_0 cuando en realidad H_A es verdadera) es mayor, debemos escoger un nivel de significación más elevado (por ejemplo, $\alpha = 0,10$).

Así, **el nivel de significación seleccionado para una prueba siempre debe reflejar las consecuencias de cometer errores de tipo I o de tipo II**.

4.6 INFERENCIA PARA OTROS ESTIMADORES

Hasta ahora, solo hemos considerado la media como estimador para la inferencia. No obstante, muchos de los conceptos que hemos visto en este capítulo pueden aplicarse, con algunas ligeras modificaciones, usando otros estimadores.

4.6.1 Estimadores puntuales con distribución cercana a la normal

En realidad existen múltiples estimadores puntuales, además de la media, cuya distribución muestral es cercana a la normal si las muestras son lo suficientemente grandes, tales como las proporciones y la diferencia de medias. Si bien veremos con detalle la prueba de hipótesis con estos estimadores puntuales en capítulos posteriores, es importante contar con algunas orientaciones generales.

Un supuesto importante que debemos tener en cuenta es que el estimador puntual $\hat{\theta}$ debe ser **insesgado**. Esto significa que la distribución muestral de $\hat{\theta}$ tiene su centro en el valor del parámetro θ que estima. En otras palabras, un estimador insesgado (como la media) tiende a proveer una estimación cercana al parámetro real.

En términos generales, el intervalo de confianza para un estimador puntual insesgado cuya distribución es cercana a la normal (como la media, las proporciones o la diferencia de medias) está dado por la ecuación 4.6, donde z^* se escoge de manera tal que se condiga con el nivel de confianza seleccionado y y la lateralidad de la hipótesis alternativa. Como se dijo anteriormente, el valor $z^* \cdot SE_{\hat{\theta}}$ se denomina "margen de error". Debemos recordar que la ecuación 4.2 corresponde al error estándar de la media, pero los errores estándar para otros estimadores puntuales se estiman de manera diferente a partir de los datos.

$$\hat{\theta} \pm z^* \cdot SE_{\hat{\theta}} \tag{4.6}$$

El método de prueba de hipótesis usando valores p puede generalizarse para otros estimadores puntuales con distribución cercana a la normal. Para ello, Diez y col. (2017, p. 199) señalan que se debemos considerar los siguientes pasos:

Prueba de hipótesis usando el modelo normal:

1. Formular las hipótesis nula (H_0) y alternativa (H_A) en lenguaje llano y luego en notación matemática.
2. Identificar un estimador puntual (estadístico) adecuado e insesgado para el parámetro de interés.
3. Verificar las condiciones para garantizar que la estimación del error estándar sea razonable y que la distribución muestral del estimador puntual siga aproximadamente una distribución normal.
4. Calcular el error estándar. Luego, graficar la distribución muestral del estadístico bajo el supuesto de que H_0 es verdadera y sombrear las áreas que representan el valor p.
5. Usando el gráfico y el modelo normal, calcular el valor p para evaluar las hipótesis y escribir la conclusión en lenguaje llano.

4.6.2 Estimadores con otras distribuciones

Existen métodos de construcción de intervalos de confianza y prueba de hipótesis adecuados para aquellos casos en que el estimador puntual o el estadístico de prueba no son cercanos a la normal (por ejemplo, si la muestra es pequeña, se tiene una mala estimación del error estándar o el estimador puntual tiene una distribución distinta a la normal). No obstante, la selección de métodos alternativos debe hacerse siempre teniendo en cuenta la distribución muestral del estimador puntual o del estadístico de prueba.

Una consideración importante es que **siempre debemos verificar el cumplimiento de las condiciones requeridas por una herramienta estadística**, pues de lo contrario las conclusiones pueden ser erradas y carecerán de validez.

4.7 EJERCICIOS PROPUESTOS

1. ¿Es correcto afirmar que, si se lanza un dado una y otra vez, la media móvil simple del número de puntos que aparecen en la cara superior crece monótonamente? Justifica tu respuesta.
2. ¿Es correcto afirmar que, si se lanza un dado una y otra vez, la proporción de veces que aparece un número impar de puntos (1, 3 o 5) en la cara superior es siempre 0,5? Justifica su respuesta.
3. Si se calcula la media de diez muestras distintas extraídas de la misma población, ¿se espera ver el mismo valor cada vez? ¿Cómo se llama a este fenómeno?
4. Completa las siguiente oraciones:
   a) Una estimación _____ _____ es un _____ calculado con datos de una muestra como aproximación del valor desconocido de un _____ de la población en estudio.
   b) $\overline{X}$ o $\overline{x}$ se usan para denotar la _____ _____, que es una estimación puntual de μ, la _____ _____.
5. Se sabe que una prueba para medir el coeficiente intelectual de jóvenes de 18 años produce puntuaciones que siguen una distribución $\mathcal{N}(\mu = 100, \sigma^2 = 100)$.
   a) Dibuja el histograma de la distribución muestral de medias para muestras de tamaño 25 de esta población.
   b) Una de las muestras anteriores presentó $\overline{x} = 95$ y $s = 15$. Determina el intervalo con 95 % de confianza para este caso.
   c) Con otra de las muestras se pudo determinar que su intervalo con 99 % confianza era $[90,26; 105,74]$. ¿Qué significa esto?
   d) El intervalo anterior, ¿es más grande o más pequeño que uno con 90 % de confianza?
6. Una empresa de tecnología quiere promocionar un software especializado para almacenar y recuperar imágenes médicas digitales. Con esta idea, está financiando un estudio para determinar el tiempo (en segundos) que necesita un grupo de médicos para recuperar imágenes desde sus propios registros en sus portátiles personales y desde la base de datos central con el software ofrecido y una conexión a la Web.
   a) Enuncia las hipótesis nula y alternativa (en castellano común).
   b) Identifica la variable aleatoria que se va a estudiar, el parámetro de interés y el correspondiente estadístico.
   c) Enuncia, más formalmente, las hipótesis nula y alternativa para este caso.
   d) Supón que el intervalo con 95 % confianza para el tiempo de recuperación promedio de una imagen digital desde la base de datos central resultó ser $[24; 36]$ [s]. ¿Qué decisión tomarías ante la hipótesis nula: la media del tiempo de recuperación de una imagen digital con el nuevo software es de 25 segundos? En este caso, ¿cuál podría ser la hipótesis alternativa?
   e) Para el intervalo de confianza anterior, ¿cuál sería un error de tipo I?
   f) Conociendo el intervalo de confianza anterior, ¿es posible cometer un error de tipo II? Explica.
7. Si una hipótesis nula es falsa, aumentar el nivel de significación para un tamaño de muestra dado, ¿reduce la probabilidad de rechazarla?
8. Qué significa que un estadístico tenga un valor p de 0,025?
9. Si una hipótesis nula es rechazada a un nivel de significación de 0,01, ¿será rechazada a un nivel de significación 0,05? Explica.
10. Si una hipótesis nula es rechazada por una prueba unilateral (una cola), ¿será también rechazada por una prueba bilateral (dos colas)? Explica.
11. Acabas de leer un artículo que hace la siguiente aseveración: "*a 95 % confidence interval for mean reaction time is from 0.25 to 0.29 seconds. Thus, about 95 % of individuals will have reaction times in this interval.*" Comenta.
12. Da el ejemplo de un estudio en que es más dañino cometer un error tipo II que un error tipo I.
13. Lista las condiciones que deben verificarse para asegurar que el TLC (teorema del límite central) está rigiendo y es posible hacer una prueba de hipótesis o calcular un intervalo de confianza.
14. Si para un estudio de una determinada variable aleatoria numérica es igualmente dañino cometer errores de tipo I como errores tipo II:
   a) Dibuja la distribución de una muestra de tamaño 16 (un diagrama de caja, por ejemplo) para la que el contraste de hipótesis con nivel de significación 0,05 sea confiable.

b) Dibuja la distribución de una muestra de tamaño 30 en que se requiera de un nivel de significación más exigente ($\alpha < 0,05$) para hacer el contraste de hipótesis más confiable.

c) Dibuja la distribución de una muestra en que es mejor no confiar en el contraste de hipótesis con métodos estudiados hasta ahora.

15. Si un estudio sobre el tiempo promedio de búsqueda y recuperación de imágenes médicas con dos tecnologías distintas reporta: "existe una diferencia significativa (p <0,02) entre el tiempo invertido con la tecnología A ($33 \pm 4[s]$) que con la tecnología B ($30 \pm 6[s]$)", ¿significa que se debe adoptar la tecnología B? ¿Por qué?
16. Explica por qué se incrementa la probabilidad de cometer errores tipo I al cambiar de una prueba de hipótesis bilateral a otra unilateral.

REFERENCIAS

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

Field, A., Miles, J. & Field, Z. (2012). *Discovering statistics using R*. SAGE Publications Ltd.

Real Academia Española. (2014). *Diccionario de la lengua española* (23.ª ed.). Consultado el 30 de marzo de 2021, desde https://dle.rae.es