Lectura 4
Fundamentos de la inferencia

Estimadores puntuales

-> Un estadístico corresponde a un estimador puntual de un parámetro.

->Este valor cambia dependiendo de la muestra que usemos para obtenerlo. Así por más que el valor se acerque al parámetro de la población, difícilmente será igual a este último

-> Mejora la medida por la **Ley de los grandes números**

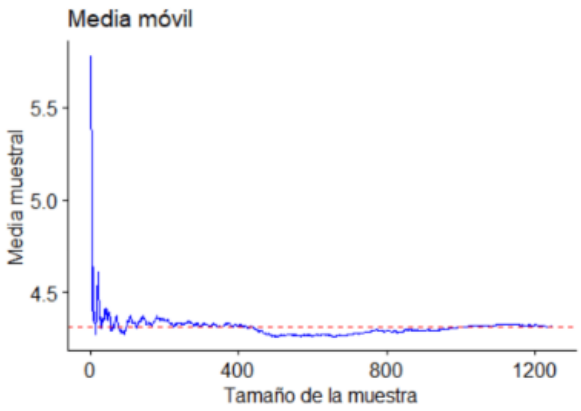

Figura 4.1: medias obtenidas al agregar a la muestra un elemento cada vez.

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 la determinación de un buen estimador se necesita saber cuánto cambia de una muestra a otra. Si esa variabilidad es pequeña, es muy probable que la estimación sea buena.

Distribución Muestral -> Ayuda a estudiar la variabilidad de una muestra, ya que esta representa la distribución de estimadores puntuales obtenidos con todas las diferentes muestras de igual tamaño de una misma población

-> Línea vertical roja que señala la media de la población

-> Las media muestrales tienden a aglutinarse en torno a la media poblacional (Teorema límite central

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")

```

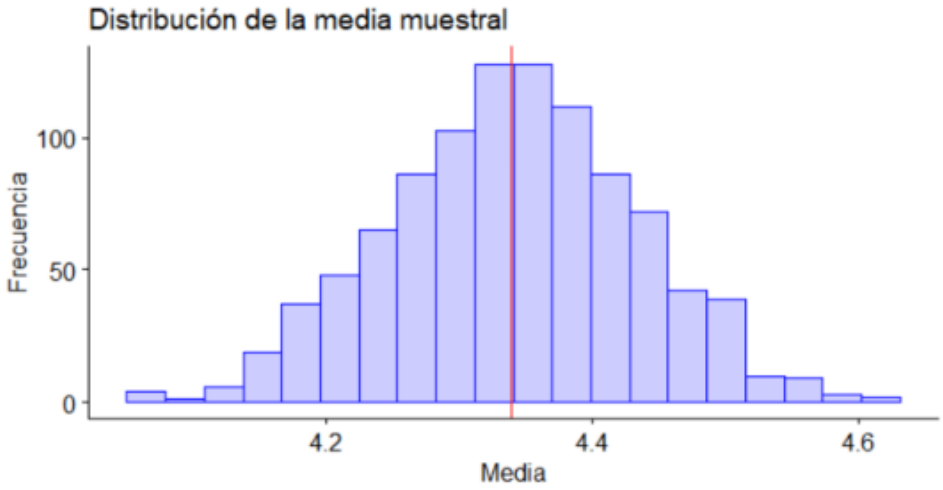

Figura 4.2: distribución muestral de la media para muestras con 100 observaciones.

```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)
```

Modelos estadísticos

Es la descripción de un **proceso probabilístico** con **parámetros desconocidos** que deben ser **estimados** en base a **suposiciones** y un conjunto de datos **observados.**

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

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.

Error estándar

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.

Antes de usarla

Condiciones que deben cumplirse

1) Las observaciones de la muestra son independientes
2) La muestra es grande (en general **n** mayor o igual a **30**)

3) La distribución de la muestra no es significativamente asimétrica (Valores atípicos) Mientras mayor sea el tamaño de la muestra, más se puede relajar esta condición

Si estas condiciones no se cumplen, se deben 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.

Intervalos de confianza

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.

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

Prueba 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: $\mu_N \neq 530$

Prueba formal de hipótesis con valores p

(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.

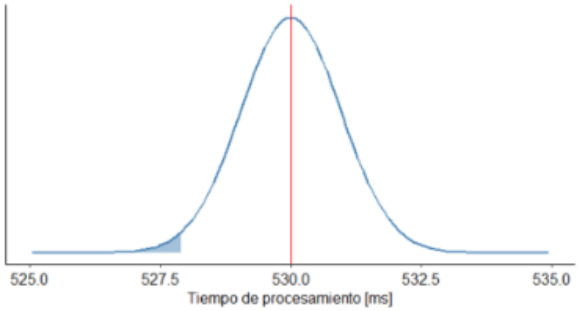

Figura 4.3: probabilidad de encontrar una media igual o menor que $\overline{x} = 527,9$ [ms] en la distribución muestral con $\mu_{\overline{x}} = 530$ y $\sigma_{\overline{x}} = 1,2$.

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.

```r
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)
```

```r
# 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


# 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)
```

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**.

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.

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.