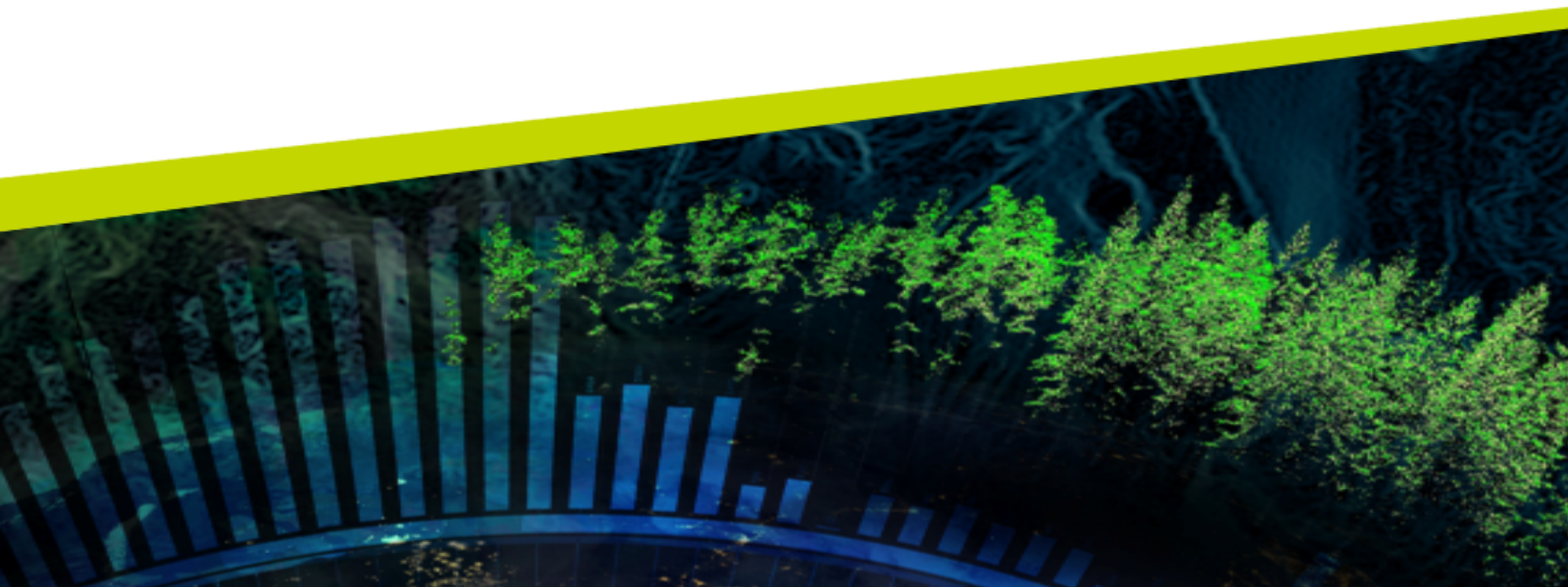
INFERENCIA Y MODELOS ESTADÍSTICOS

Jacqueline Köhler C. y José Luis Jara V.



CAPÍTULO 3. VARIABLES ALEATORIAS Y DISTRIBUCIONES DE PROBABILIDAD

Este capítulo hace una presentación resumida de muchos conceptos y, si bien son materias previas que deberían manejarse, si necesitas más material para completar su entendimiento, puedes consultar las fuentes en que se basa este capítulo: Diez et al. (2017, pp. 104-157) y Freund y Wilson (2003, pp. 104-106).

Comencemos con una definición de **variable aleatoria** que entenderemos como a un tipo de variable, usualmente numérica, cuyos posibles valores dependen de un **proceso aleatorio**. Dichas variables se nombran con letras mayúsculas (e.g. X, Y, Z) y denotamos sus posibles valores por la letra minúscula correspondiente, acompañada de un subíndice (e.g. x_1, x_2, x_3). Las variables aleatorias tienen una **función de probabilidad**, la cual define la probabilidad de que ocurran los diferentes valores que dicha variable puede tomar.

3.1 VARIABLES ALEATORIAS DISCRETAS

Definimos como **variable aleatoria discreta** a aquella cuya función de probabilidad solo toma valores distintos de cero en un conjunto finito, o infinito numerable, de valores. A esta función de probabilidad se le llama **función de masa de probabilidad**.

Un ejemplo típico de variable aleatoria discreta es el resultado del lanzamiento de un dado. Si el dado está bien balanceado, tendremos igual probabilidad de obtener cualquiera de las caras. Pero es sabido que algunos tramposos fabrican dados adulterados para favorecer algunos resultados, como la obtención de valores 1 y 6. Una distribución aleatoria de la variable "X: lanzamiento de un dado adulterado" podría ser la que se presenta en la tabla 3.1, la que iremos usando en los siguientes ejemplos.

i	1	2	3	4	5	6	Total
x_i	1	2	3	4	5	6	-
$P(X = x_i)$	0.250	0.125	0.125	0.125	0.125	0.250	1.000

Tabla 3.1: distribución de probabilidad para el lanzamiento de un dado adulterado.

El **valor esperado**, denotado como $\mathrm{E}[X]$ o la letra griega μ (mu), corresponde al resultado promedio de una variable aleatoria. Para una variable aleatoria discreta, se calcula sumando los valores posibles ponderados por su probabilidad, como muestra la ecuación 3.1.

$$\mathrm{E}[X] \equiv \mu = \sum_{i=1}^{n} x_i P(X = x_i) \tag{3.1}$$

También podemos calcular qué tan alejado podría estar el valor obtenido del valor esperado por medio de la **varianza**, denotada por $\mathrm{Var}(X)$ o el cuadrado de la letra griega sigma (σ^2), que se calcula como el valor esperado de los cuadrados de la diferencia con respecto a la media, como muestra la ecuación 3.2. Una vez más, la desviación estándar corresponde a la raíz cuadrada de la varianza y se denota σ.

$$\mathrm{Var}(X) \equiv \sigma^2 = \mathrm{E}[(x_i - \mu)^2] = \sum_{i=1}^{n} (x_i - \mu)^2 P(X = x_i) \tag{3.2}$$

En R, el script 3.1 ejemplifica cómo obtener estos valores, además de la desviación estándar, para el ejemplo del dado adulterado.

Script 3.1: estadísticas descriptivas de una variable aleatoria discreta.

```r
# Crear una variable discreta para representar el dado adulterado
resultados <- 1:6
probabilidades <- c(0.25, 0.125, 0.125, 0.125, 0.125, 0.25)

# Calcular el valor esperado
esperado <- sum(resultados * probabilidades)
cat("Valor esperado:", esperado, "\n")

# Calcular la varianza
varianza <- sum(((resultados - esperado) ^ 2) * probabilidades)
cat("Varianza:", varianza, "\n")

# Calcular la desviación estándar
desviacion <- sqrt(varianza)
cat("Desviación estándar:", desviacion, "\n")
```

Para ayudarnos a entender mejor la noción de distribución de probabilidad, veamos la figura 3.1 obtenida mediante el script 3.2. Ella nos muestra, de izquierda a derecha, las distribuciones de probabilidad para la puntuación total obtenida al lanzar 5, 10 y 20 dados adulterados, respectivamente.

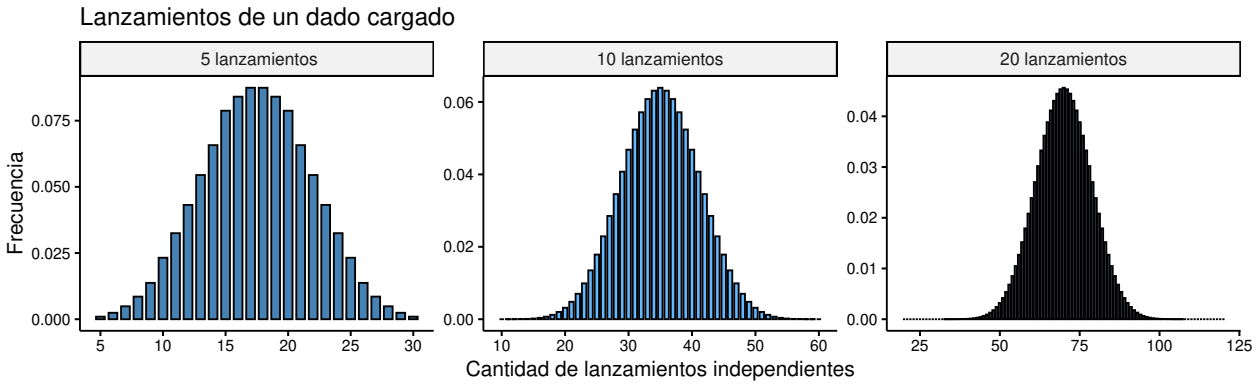


Figura 3.1: distribución de probabilidad para varios lanzamientos de un dado cargado.

Script 3.2: histogramas de variables aleatorias discretas en R.

```r
library(ggpubr)

# Suma de variables aleatorias independientes e idénticamente distribuidas (IID).
# Es decir, cada variable aleatoria tiene la misma distribución de probabilidad
# y todas son mutuamente independientes.
# Parámetros:
#   pr: un vector de con la distribución de probabilidad. Los nombres asociados
#       a cada elemento del vector corresponde a los posibles resultados de la
#       variable aleatoria, que deben ser valores enteros distintos.
#    n: el número de variables IID a sumar
SumaIID <- function(pr, n = 2) {
  probs <- pr
  i <- 2
  # Como un resultado depende del resultado anterior, se usa un ciclo tradicional
  while(i <= n) {
    # Producto de los vectores de probabilidades
    npr <- outer(probs, pr, FUN="*")
    # Obtiene a qué salida pertenece cada probabilidad
    nout <- outer(as.numeric(names(probs)), as.numeric(names(pr)), FUN = "+")
    # Suma las probabilidades correspondientes a cada salida
```

```
    tmp <- tapply(npr, nout, sum)

    probs <- tmp
    i <- i + 1
  }

  invisible(probs)
}

# Crear una variable discreta para representar el dado adulterado
resultados <- 1:6
probabilidades = c(0.25, 0.125, 0.125, 0.125, 0.125, 0.25)
names(probabilidades) <- resultados

# Crear vector con los resultados de 5 lanzamientos del dado
lanzar_5 <- SumaIID(probabilidades, n = 5)
lanzar_5_df <- data.frame(Salida = names(lanzar_5), Prob = lanzar_5, N = "05")

# Crear vector con los resultados de 10 lanzamientos del dado
lanzar_10 <- SumaIID(probabilidades, n = 10)
lanzar_10_df <- data.frame(Salida = names(lanzar_10), Prob = lanzar_10, N = "10")

# Crear vector con los resultados de 20 lanzamientos del dado
lanzar_20 <- SumaIID(probabilidades, n = 20)
lanzar_20_df <- data.frame(Salida = names(lanzar_20), Prob = lanzar_20, N = "20")

# Juntar las matrices de datos con los resultados
lanzamientos <- rbind(lanzar_5_df, lanzar_10_df, lanzar_20_df)
lanzamientos[["Salida"]] <- as.integer(lanzamientos[["Salida"]])

# Graficar las los resultados
g <- ggbarplot(lanzamientos, x = "Salida", y = "Prob",
               fill = "N", palette = c("steelblue", "steelblue1", "slategray4"),
               title = "Lanzamientos de un dado cargado",
               xlab = "Cantidad de lanzamientos independientes",
               ylab = "Frecuencia")
g <- g + scale_x_continuous(breaks = get_breaks(n = 6))
g <- ggpar(g, legend = "none", font.tickslab = c(9, "plain", "black"))
g <- facet(g, facet.by = "N", scales = "free",
              panel.labs = list(N = c(" 5 lanzamientos",
                                      "10 lanzamientos",
                                      "20 lanzamientos")))

print(g)
```

Conocer la función de masa de probabilidad de una variable discreta nos ayuda a hacer estimaciones útiles. A modo de ejemplo, supongamos que un ingeniero de software debe crear un programa que resuelva un problema (siempre con instancias del mismo tamaño) con un tiempo de respuesta no mayor a 25 segundos. El histograma de la figura 3.2 muestra los tiempos de ejecución obtenidos para 500 pruebas de la solución propuesta, donde se observa que 30 de ellas tardaron en realidad más de 25 segundos, con un rango que va de los 10 a los 30 segundos. Así, podemos estimar la probabilidad de que el tiempo de ejecución sea mayor a 25 segundos dividiendo la cantidad de observaciones que cumplen este criterio por la cantidad total de instancias, como muestra la ecuación 3.3.

$$P(X > 25) = \frac{30}{500} = 0.06 \tag{3.3}$$

Frecuentemente resulta más adecuado expresar o modelar un fenómeno como una combinación de dos o más variables aleatorias. Por ejemplo, un jugador de baloncesto puede anotar canastas de uno, dos o tres puntos

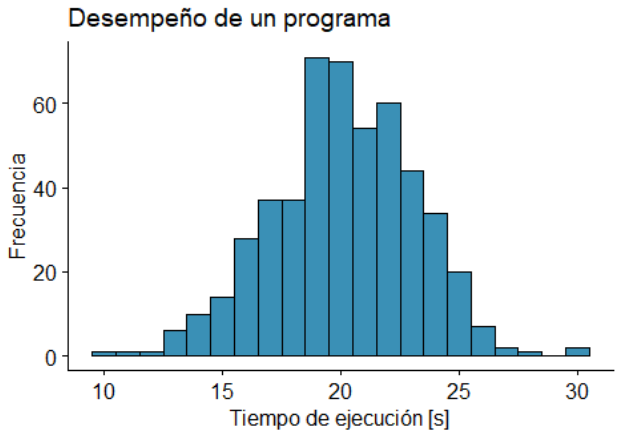


Figura 3.2: histograma para el desempeño del programa.

dependiendo de si encesta con un tiro libre, un lanzamiento desde dentro del área o desde fuera del área, respectivamente. Así, se tienen tres variables aleatorias:

1. X: Anotaciones por tiro libre.
2. Y: Anotaciones desde dentro del área.
3. Z: Anotaciones desde fuera del área.

Podemos representar el total de puntos anotados por el jugador como la suma de los puntos anotados de las tres formas posibles, lo que corresponde a una **combinación lineal** de las variables X, Y y Z. La fórmula general de una combinación lineal de n variables está dada por la ecuación 3.4, donde cada X_i corresponde a una variable aleatoria y cada c_i es una constante conocida.

$$\sum_{i=1}^{n} c_i X_i \tag{3.4}$$

Cuando las variables de una combinación lineal son **independientes**[1], es decir que conocer el valor que toma una de ellas no aporta ninguna información acerca de la distribución de la otra, podemos calcular el valor esperado y la varianza de la combinación lineal usando las ecuaciones 3.5 y 3.6. Una vez más, la desviación estándar está dada por la raíz cuadrada de la varianza.

$$\mathrm{E}\left[\sum_{i=1}^{n} c_i X_i\right] = \sum_{i=1}^{n} c_i \,\mathrm{E}[X_i] \tag{3.5}$$

$$\mathrm{Var}\left(\sum_{i=1}^{n} c_i X_i\right) = \sum_{i=1}^{n} c_i^2 \,\mathrm{Var}(X_i) \tag{3.6}$$

3.2 VARIABLES ALEATORIAS CONTINUAS

Definimos como **variable aleatoria continua** a aquella variable aleatoria cuya función de probabilidad es una función continua que asigna una probabilidad de ocurrencia a todos los infinitos valores posibles dentro de un intervalo. Dicha función de probabilidad recibe el nombre de **función de densidad de probabilidad** o simplemente **distribución** o **densidad**.

Debemos hacer dos observaciones importantes:

[1] Si las variables no son independientes, se requieren métodos más complejos fuera del alcance de este libro.

- el área bajo la de una función de densidad de probabilidad tiene valor 1, y
- por su naturaleza continua, $P(X = x) = 0$ para todo valor x.

Un concepto importante en este punto es la **función de distribución de probabilidad** o **distribución de probabilidad** o simplemente **distribución**, que corresponde a la probabilidad $P(X \leq x)$ y que se define para variables aleatorias continuas según la ecuación 3.7, en donde $f(x)$ es la función de densidad.[2]. Con esta definición, puede verse que cualquier probabilidad que necesite calcularse se obtiene estimando la correspondiente área bajo la curva de densidad.

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\, dt \tag{3.7}$$

Volvamos a considerar el ejemplo del programa con un tiempo de respuesta máximo visto en la sección anterior. La variable aleatoria "X: tiempo de ejecución para resolver la instancia de prueba" es en realidad una variable continua. Así, suponiendo una función de densidad basada en el histograma de la figura 3.2, la probabilidad de que el tiempo de ejecución sea mayor a 25 segundos corresponde al área coloreada en el gráfico de la figura 3.3, con un valor de 0,048. El cálculo de esta probabilidad se aborda un poco más adelante.

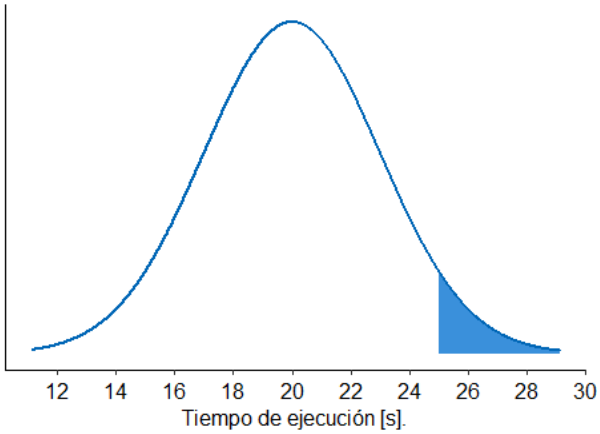


Figura 3.3: distribución para el desempeño del programa.

Existen múltiples funciones de distribución continua que son de uso frecuente en estadística, las cuales se describen a continuación.

3.2.1 Distribución normal

También conocida como **distribución gaussiana**, la **distribución normal** es la más ampliamente empleada en estadística, pues muchas variables se acercan a esta distribución. Se caracteriza por ser unimodal y simétrica, con forma de campana. La figura 3.3 ejemplifica una de estas distribuciones.

La distribución normal se usa para modelar diversos fenómenos y podemos ajustarla mediante dos parámetros:

- μ: la media, que desplaza el centro de la curva a lo largo del eje x.
- σ: la desviación estándar, que modifica qué tan dispersos están los datos con respecto a la media.

Así, denotamos este tipo de distribución por $\mathcal{N}(\mu, \sigma^2)$. La figura 3.4, creada mediante el script 3.3, muestra ejemplos superpuestos de distribuciones normales con media 10 y diferentes desviaciones estándar.

[2]En el caso de variables aleatorias continuas, se suele escribir x para denotar un valor en particular en vez de x_i, que se usa comúnmente en el caso de variables aleatorias discretas

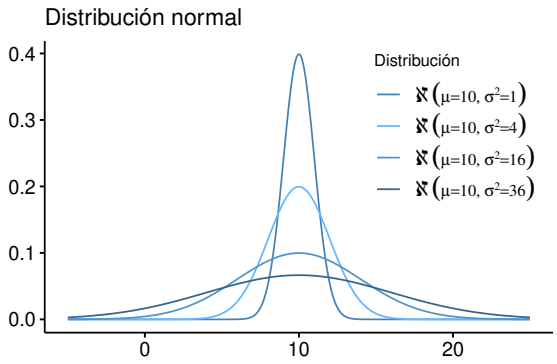


Figura 3.4: ejemplos superpuestos de distribuciones normales.

Script 3.3: graficando dos ejemplos de distribución normal.

```r
library(ggpubr)
library(latex2exp)

# Definir las distribuciones normales del ejemplo
medias <- c(10, 10, 10, 10)
sigmas <- c(1, 2, 4, 6)
vars <- sigmas^2
# Crear las etiquetas para cada ejemplo
labels_str <- sprintf(r'($\aleph(\small{\mu=%d,\sigma^2=%d})$)', medias, vars)

# Construir las distribuciones normales
x <- seq(-5, 25, 0.01)
y1 <- dnorm(x, mean = medias[1], sd = sigmas[1])
y2 <- dnorm(x, mean = medias[2], sd = sigmas[2])
y3 <- dnorm(x, mean = medias[3], sd = sigmas[3])
y4 <- dnorm(x, mean = medias[4], sd = sigmas[4])

# Juntar las distribuciones de ejemplo en una sola matriz de datos
n <- length(x)
y = c(y1, y2, y3, y4)
ejemplo <- factor(rep(sigmas, each = n))
datos_norm <- data.frame(x, y, ejemplo)

# Graficar las distribuciones normales
g <- ggline(datos_norm, "x", "y", color = "ejemplo"
            , numeric.x.axis = TRUE, plot_type = "l"
            , xlab = FALSE, ylab = FALSE
            , title = "Distribución normal"
)
# Desplegar las leyendas usando una interfaz tipo LaTeX, que la
# función TeX() trasforma en expresiones que pueden ser entendidas
# por los diferentes dispositivos gráficos (como SVG, PDF, PostScript, etc.),
# cambiando los colores usados por defecto.
colores <- c("steelblue", "steelblue1", "steelblue3", "steelblue4")
g <- ggpar(g, legend = c(0.8, 0.7))
g <- ggpar(g, legend.title = "Distribución")
g <- ggpar(g, font.legend = c(10, "plain", "black"))
g <- g + scale_color_discrete(labels = lapply(labels_str, TeX),
                              type = colores)
g <- g + theme(legend.text = element_text(family = "serif", size = 14))
# Mostrar el gráfico
print(g)
```

Antes de continuar, fijémonos en las líneas 13 a la 16 del script 3.3, donde se usa la función `dnorm(x, mean, sd)` que corresponde a la función de densidad de una distribución normal con media $\mu =$ `mean` y desviación estándar $\sigma =$ `sd`. Además de `dnorm()`, R nos ofrece otras funciones que también resultan de mucha ayuda:

- `pnorm(q, mean, sd, lower.tail)`: corresponde a la distribución de probabilidad de la distribución normal especificada. Así, devuelve $P(X \leq$ `q`$)$ cuando `lower.tail = TRUE` (que es el valor por omisión) y $P(X >$ `q`$)$ cuando `lower.tail = FALSE`.
- `qnorm(p, mean, sd, lower.tail)`: encuentra el cuantil que determina un área de tamaño `p` bajo la curva de densidad de la distribución normal especificada. El área considerada está a la izquierda de `q` cuando `lower.tail = TRUE` (valor por omisión) y a su derecha cuando `lower.tail = FALSE`. Así, `qnorm()` corresponde la función inversa de `pnorm()` para una misma distribución normal.
- `rnorm(n, mean, sd)`: genera aleatoriamente `n` observaciones de la distribución normal especificada.

Los argumentos de esta familia de funciones son:

- `x`, `q`: vector de cuantiles.
- `p`: vector de probabilidades.
- `mean`: media de la distribución normal.
- `sd`: desviación estándar de la distribución normal.
- `lower.tail`: valor lógico que señala cuál de los dos extremos o colas de la distribución emplear.
- `n`: tamaño del vector resultante.

Es importante observar que en estas funciones proporcionadas por R, por omisión el argumento `lower.tail` toma el valor verdadero, por lo que operan con la cola inferior de la distribución. Si, en cambio, se especifica `lower.tail = FALSE`, dichas funciones operan con la cola superior. También notemos que, como la gran mayoría de funciones en R, estas funciones son **vectorizadas**, por lo que algunos parámetros pueden corresponder a vectores de valores. Si se cumplen las restricciones de coincidencia de tamaños entre estos vectores, la función devuelve un vector con los resultados de cada caso especificado. Debe consultarse los manuales de R para ver en detalle cómo se interpretan vectores en cada combinación. Por ejemplo, la llamada `pnorm(c(10, 15, 20, 25), mean = 16, sd = 3)` devolverá el vector de largo 4 con las probabilidades de observar los valores 10, 15, 20 y 25, respectivamente, en una distribución normal con media 16 y desviación estándar 3; mientras que `pnorm(c(10, 15, 20, 25), mean = c(16, 18), sd = (3, 4))` devolverá el vector (p_1, p_2, p_3, p_4) donde p_1 y p_3 son, respectivamente, las probabilidades de observar los valores 10 y 20 en una distribución normal con media 16 y desviación estándar 3, y p_2 y p_4 son las probabilidades de observar los valores 15 y 25, respectivamente, en una distribución normal con media 18 y desviación estándar 4.

Una **regla empírica** muy útil al momento de trabajar con distribuciones normales es la llamada regla 68-95-99.7, ilustrada en la figura 3.5, la cual establece que:

- Cerca de 68 % de las observaciones se encuentran a una distancia de una desviación estándar de la media.
- Alrededor de 95 % de las observaciones se encuentran a una distancia de dos desviación estándar de la media.
- Aproximadamente 99.7 % de las observaciones se encuentran a una distancia de tres desviación estándar de la media.

Muchas de las pruebas estadísticas que estudiaremos más adelante, operan bajo el supuesto de que los datos siguen una distribución normal. Como se insinuó en párrafos anteriores, la normalidad es siempre una aproximación, por lo que debemos verificar que el supuesto de una distribución normal sea aceptable. Una buena herramienta para ello es el **gráfico cuantil-cuantil**, también llamado **gráfico Q-Q**, *quantile–quantile*, que se muestra en la figura 3.6 y que podemos construir en R como muestra el script 3.4. En él podemos distinguir los siguientes elementos: un grupo de puntos, una recta y una región coloreada. Los puntos corresponden a las observaciones, mientras que la recta representa la distribución normal. En consecuencia, mientras más se asemeje el patrón que forman los puntos a la recta, más parecida será la distribución a la normal. La banda coloreada establece el margen aceptable para suponer normalidad en el conjunto de datos. Así, para el conjunto de datos de la figura 3.6 corresponderá al analista decidir si es imprudente aceptar el supuesto de normalidad con 3 valores que están, muy cerca, pero fuera de la región que se espera. Eso depende de la criticidad de tarea en manos y de la cantidad de datos, puesto que en los extremos es usual encontrar

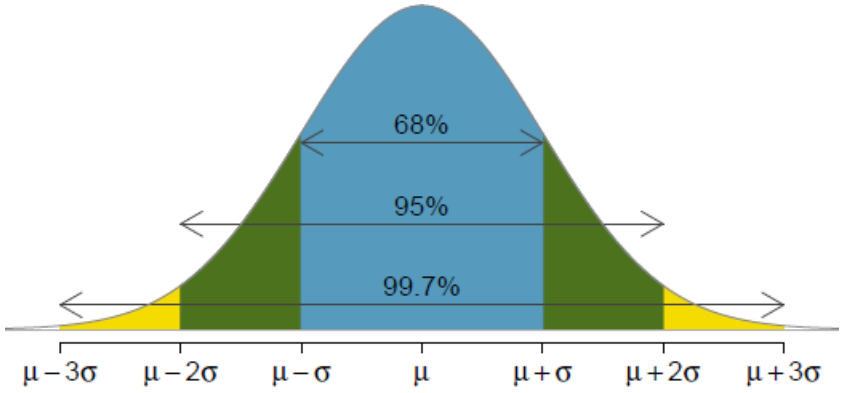


Figura 3.5: regla empírica de la distribución normal. Fuente: Diez et al. (2017, p. 136).

mayor variabilidad (de ahí el mayor ancho de la banda aceptable) y que estos valores se estiman a partir de la muestra que se está analizando.

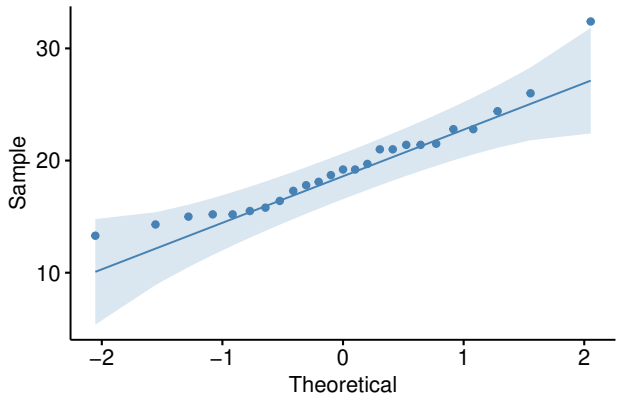


Figura 3.6: gráfico cuantil-cuantil.

Script 3.4: creación de un gráfico cuantil-cuantil.

```
library(ggpubr)

# Cargar datos.
datos <- read.csv2("Mtcars.csv", stringsAsFactors = TRUE,
                   row.names = 1)

# Gráfico Q-Q para la variable Rendimiento.
g <- ggqqplot(datos, x = "Rendimiento", color = "steelblue")

print(g)
```

3.2.2 Distribución Z

Al trabajar con distribuciones, especialmente las simétricas, a menudo usaremos **técnicas de estandarización** para determinar qué tan usual o inusual es un determinado valor en una escala única. Así, para la distribución normal usamos como estandarización la **distribución Z** o **distribución normal estándar**, que no es más que una distribución normal centrada en 0 y con desviación estándar 1, que para datos reales podemos obtener de manera bastante sencilla como muestra la ecuación 3.8.

$$Z = \frac{x - \mu}{\sigma} \sim \mathcal{N}(0, 1) \tag{3.8}$$

Al aplicar la ecuación 3.8 a una observación x en una distribución normal obtenemos, entonces, su **valor z**, que determina cuán por encima o por debajo de la media (en términos de la desviación estándar) se encuentra dicha observación x. Así, observaciones cuyos valores z sean negativos estarán por debajo de la media. Análogamente, un valor z positivo indica que la observación está por sobre la media. Mientras mayor sea el valor absoluto de su valor z ($|z|$), más inusual será la observación.

Mencionemos que las funciones `dnorm()`, `pnorm()`, `qnorm()` y `rnorm()` asumen por omisión que `mean = 0` y `sd = 1`, permitiendo trabajar fácilmente con esta distribución estándar.

3.2.3 Distribución chi-cuadrado

También llamada **ji-cuadrado** o χ^2, la distribución **chi-cuadrado** (Devore, 2008) se usa para caracterizar valores siempre positivos y habitualmente con asimetría positiva (la cola se alarga hacia la derecha). El único parámetro de esta distribución corresponde a los **grados de libertad**, usualmente representada por la letra griega ν (nu), que son una estimación de la cantidad de observaciones empleadas para calcular un estimador. Otra forma de entender esta idea es como la cantidad de valores que pueden cambiar libremente en un conjunto de datos. Como ejemplo, supongamos que necesitamos una muestra de tres elementos cuya media sea 10. Una vez escogidos los primeros dos, solo queda una posibilidad para el tercero de modo que se cumpla con la media deseada. Así, solo los dos primeros valores pueden cambiar libremente, por lo que se tienen dos grados de libertad.

Esta distribución está relacionada con la ya conocida distribución Z, pues si sumamos los cuadrados de k variables aleatorias independientes que siguen una distribución Z, dicha suma sigue una distribución χ^2 con k grados de libertad:

$$\sum_{i=1}^{k} Z_i^2 \sim \chi^2(\nu = k) \tag{3.9}$$

La media de la distribución χ^2 es $\mu = \nu$, y su desviación estándar, $\sigma = 2\nu$.

Las funciones de R para esta distribución, similares a las descritas para la distribución normal, son:

- `dchisq(x, df)`.
- `pchisq(q, df, lower.tail)`.
- `qchisq(p, df, lower.tail)`.
- `rchisq(n, df)`.

Donde:

- `x`, `q` son vectores de cuantiles (enteros no negativos).
- `p` es un vector de probabilidades.
- `n` es la cantidad de observaciones.
- `df` son los grados de libertad.
- `lower.tail` es análogo al de la función `pnorm`.

La figura 3.7 muestra ejemplo de distribuciones χ^2 con diferentes grados de libertad.

3.2.4 Distribución t de Student

Ampliamente empleada cuando se trabaja con muestras pequeñas, la **distribución t de Student**, o simplemente **distribución t**, tiene, al igual que la distribución χ^2, los grados de libertad como único parámetro. A

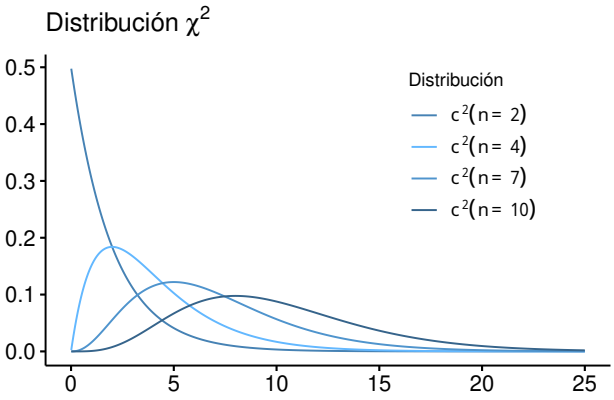


Figura 3.7: ejemplo de distribución χ^2 con 2 grados de libertad.

medida que los grados de libertad aumentan, esta distribución se asemeja cada vez más a la normal, aunque sus colas son más gruesas, como ilustra la figura 3.8.

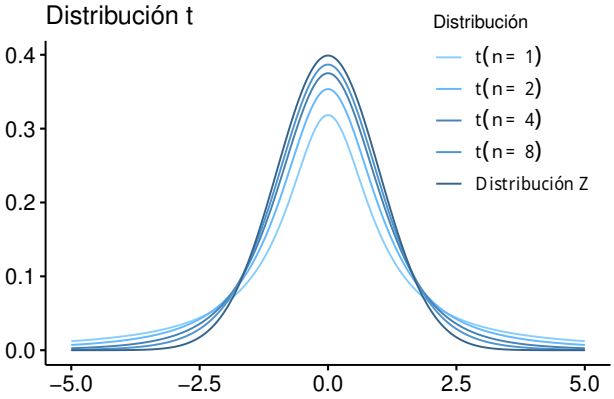


Figura 3.8: ejemplo de distribuciones t con diferentes grados de libertad comparadas con una distribución Z.

La distribución t se encuentra relacionada con las distribuciones vistas anteriormente de acuerdo a la ecuación 3.10, donde Z es una distribución normal estándar y $\chi^2(\nu)$ es una distribución χ^2 con ν grados de libertad.

$$Z\sqrt{\frac{v}{\chi^2(v)}} \sim t(\nu) \tag{3.10}$$

La media de la distribución t, para $\nu > 1$, es $\mu = 0$. Su desviación estándar, para $\nu > 2$, está dada por la ecuación 3.11.

$$\sigma = \sqrt{\frac{\nu}{\nu - 2}} \tag{3.11}$$

Las funciones de R para esta distribución, cuyos argumentos son análogos a los que hemos visto para las distribuciones anteriores, son:

- `dt(x, df)`.
- `pt(q, df, lower.tail)`.
- `qt(p, df, lower.tail)`.
- `rt(n, df)`.

3.2.5 Distribución F

Otra distribución que usaremos a lo largo de este libro es la **distribución F**, ampliamente usada para comparar varianzas. La distribución F se relaciona con las anteriores de acuerdo a la ecuación 3.12, donde $\chi_1^2(\nu_1)$ y $\chi_2^2(\nu_2)$ son dos distribuciones χ^2 con ν_1 y ν_2 grados de libertad, respectivamente. Ejemplos de distribuciones F se pueden encontrar en la figura 3.9.

$$\frac{\frac{X_1^2(\nu_1)}{\nu_1}}{\frac{X_2^2(\nu_2)}{\nu_2}} \sim F(\nu_1, \nu_2) \tag{3.12}$$

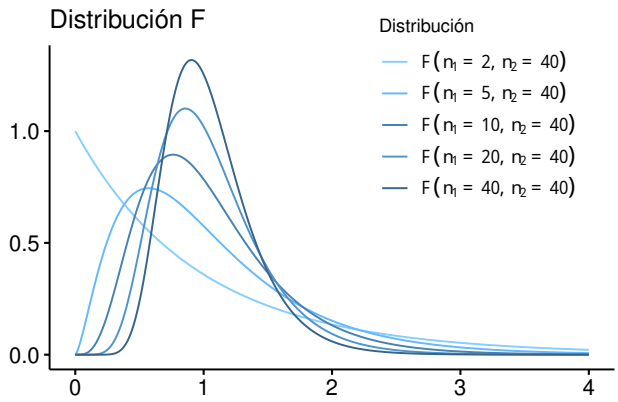


Figura 3.9: ejemplo de distribuciones F con diferentes grados de libertad.

Para $\nu_2 > 2$, la media de esta distribución está dada por la ecuación 3.13, y la desviación estándar corresponde a la ecuación 3.14 para $\nu_2 > 4$.

$$\mu = \frac{\nu_2}{\nu_2 - 2} \tag{3.13}$$

$$\sigma = \sqrt{\frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)}} \tag{3.14}$$

Las funciones de R para esta distribución son:

- `df(x, df1, df2)`.
- `pf(q, df1, df2, lower.tail)`.
- `qf(p, df1, df2, lower.tail)`.
- `rf(n, df1, df2)`.

Donde `df1` como `df2` corresponden a grados de libertad y los argumentos restantes son los mismos que ya hemos visto anteriormente.

3.3 DISTRIBUCIONES DISCRETAS

Al igual que con las variables aleatorias continuas, también existen diversas distribuciones discretas de uso frecuente en estadística, las que veremos a continuación.

Pero primero debemos mencionar que para estas distribuciones también existe la **función de distribución de probabilidad**, o **función de distribución acumulada**, que se define según la ecuación 3.15, en donde $f(x)$ es la función de masa de probabilidad.

$$F(x_i) = P(X \leq x_i) = \sum_{x_j \leq x_i} f(x_j) \tag{3.15}$$

3.3.1 Distribución de Bernoulli

Un **ensayo de Bernoulli** es un experimento aleatorio con solo dos resultados posibles: "éxito" y "fracaso". "Lanzar una moneda al aire esperando observar cara" y "lanzar un dado de 20 lados esperando obtener un 20 como resultado" son dos ejemplos de ensayos Bernulli. Notemos que la selección de qué resultado se considera como éxito o fracaso suele ser arbitraria. Consideremos, para ilustrar esta idea, que si dos personas lanzan una moneda al aire para sortear al ganador, cada una de ellas considerará un lado diferente de la moneda como un éxito.

Una **variable aleatoria de Bernoulli** es aquella que modela un ensayo de Bernoulli, por lo que puede tomar solo dos valores: 1 y 0, que comúnmente representan a un éxito y a un fracaso, respectivamente. La probabilidad de que esta variable tomo valor 1 es fija y se denota p, mientras que toma valor 0 con probabilidad $q = 1 - p$. Luego, la distribución de probabilidad asociada a esta variable aleatoria, llamada **distribución Bernoulli** tiene media $\mu = p$ y desviación estándar $\sigma = \sqrt{pq} = \sqrt{p(1-p)}$.

Entenderemos por un **proceso Bernoulli** como una secuencia, finita o infinita contable, de repeticiones **independientes** e **idénticamente distribuidas** de ensayos de Bernulli. Así, lanzar varias veces un dado de 20 caras esperando obtener un 20 como resultado es un ejemplo obvio de un proceso Bernulli, ya que cada uno de los lanzamientos tiene probabilidad de éxito (obtener 20) $p = 1/20 = 0.05$ y probabilidad de fracaso (obtener otro valor) $q = 1 - p = 0.95$. Además, estos lanzamientos del dado son independientes puesto que el resultado de un lanzamiento no afecta el resultado de los demás.

Definimos la **proporción de la muestra** $\hat{p}$ para un proceso Bernoulli como la cantidad de éxitos dividida por la cantidad de intentos. Mientras mayor sea la cantidad de intentos, más cercano será el valor de $\hat{p}$ a la real probabilidad de éxito p.

El paquete `extraDistr` de R ofrece 4 funciones, similares a las ya conocidas, para la distrubución de Bernoulli:

- `dbern(x, prob)`.
- `pbern(q, prob, lower.tail)`.
- `qbern(p, prob, lower.tail)`.
- `rbern(n, prob)`.

3.3.2 Distribución geométrica

Una **variable aleatoria geométrica** corresponde al número de fracasos que se observan antes de obtener un éxito en un proceso Bernoulli. La **distribución geométrica** describe la probabilidad para cada posible valor de esta variable aleatoria. La variable toma valor $X = n$ si los primeros n ensayos de Bernoulli resultaron en fracaso (con probabilidad q cada uno) y el siguiente en éxito (con probabilidad p). La probabilidad asociada a este evento está dada por la ecuación 3.16, donde podemos ver que esta decrece exponencialmente rápido en esta distribución, como ilustra la figura 3.10. La media y la desviación estándar de la distribución geométrica están dadas, respectivamente, por las ecuaciones 3.17 y 3.18.

$$\Pr(n \text{ fracasos antes del primer éxito}) = P(X = n) = p(1-p)^n \tag{3.16}$$

$$\mu = \frac{1-p}{p} \tag{3.17}$$

$$\sigma = \sqrt{\frac{1-p}{p^2}} \tag{3.18}$$

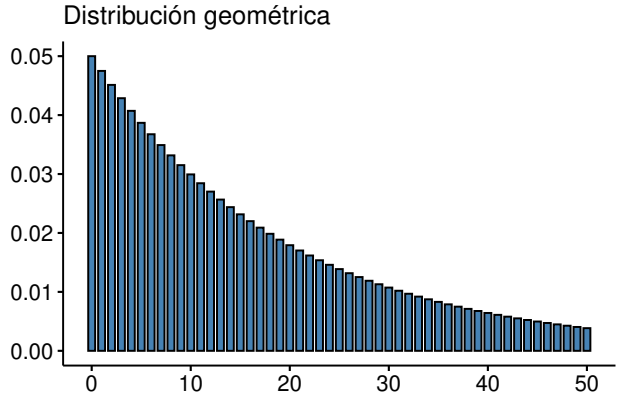


Figura 3.10: distribución geométrica para obtener un valor específico lanzando un dado de 20 caras.

Para entender mejor la utilidad de la distribución geométrica, consideremos la pregunta: ¿cuántas veces tenemos que lanzar un dado de 20 caras para obtener un 20? La respuesta se obtiene considerando el valor esperado la distribución geométrica con $p = 1/20$, que corresponde a su media, que en este caso sería como muestra la ecuación 3.19. Luego, en promedio necesitaríamos 20 tiros del dado para conseguir el valor deseado.

$$\mu = \frac{1-p}{p} = \frac{19/20}{1/20} = 19 \tag{3.19}$$

Una vez más, R ofrece funciones similares a las presentadas anteriormente para trabajar con distribuciones geométricas:

- `dgeom(x, prob)`.
- `pgeom(q, prob, lower.tail)`.
- `qgeom(p, prob, lower.tail)`.
- `rbern(n, prob)`.

Así, siguiendo con el ejemplo del dado de 20 caras, la probabilidad de obtener un valor 20 después de 10 tiros fallidos puede calcularse con la llamada `dgeom(10, prob = 1/20)`, mientras que `pgeom(10, prob = 1/20)` nos entrega la probabilidad de obtener el valor 20 con a lo más 10 tiros fallidos, y `qgeom(0.5, prob = 1/20)` nos dirá la cantidad n de tiros fallidos que nos da al menos un 50 % de probabilidad de observar el valor 20 en el tiro $(n + 1)$. Al tratarse de una distribución discreta, existen dos consideraciones que no teníamos con las continuas. Primero, si `x` no es un valor entero, la función `dgeom()` devuelve probabilidad cero con un mensaje de advertencia. Y segundo, el cuantil que devuelve la función `qgeom()` corresponde al menor valor entero x_i que cumple $F(x_i) \geq$ `p` (recordemos que F es la función de distribución que implementa `pgeom()`).

Una nota importante para no caer en confusión: la definición que usamos de una variable aleatoria geométrica "X: cantidad de fracasos antes de obtener un éxito" fue seleccionada porque corresponde a la distribución implementada por las funciones `dgeom()`, `pgeom()`, `qgeom()` y `rgeom()`. Pero en la literatura **existe otra definición** para esta variable: "Y: cantidad de ensayos requeridos para obtener el primer éxito". Si bien esta definición y la definición usada aquí se relacionan de forma muy simple: $Y = X + 1$, la masa de probabilidad y, en consecuencia, su distribución de probabilidad y medidas estadísticas son diferentes (de hecho el valor 0 no estaría en el dominio de Y).

3.3.3 Distribución binomial

Una **variable aleatoria binomial** corresponde al número de éxitos que se observan en un proceso Bernoulli. La **distribución binomial**, entonces, describe la probabilidad de tener exactamente k éxitos en n ensayos de Bernoulli independientes cada uno con probabilidad de éxito fija p. La función de masa de probabilidad de esta distribución está dada por la ecuación 3.20, donde $\binom{n}{k}$ corresponde a la cantidad de formas de obtener

k éxitos en un total de n intentos y $p^k(1-p)^{n-k}$ es la probabilidad de tener un único éxito en una de las $\binom{n}{k}$ maneras posibles. Su media corresponde a $\mu = np$ y su desviación estándar a $\sigma = \sqrt{np(1-p)}$. Ejemplo de esta distribución se presentan en la figura 3.11

$$f(k, n, p) = \Pr(k; n, p) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{3.20}$$

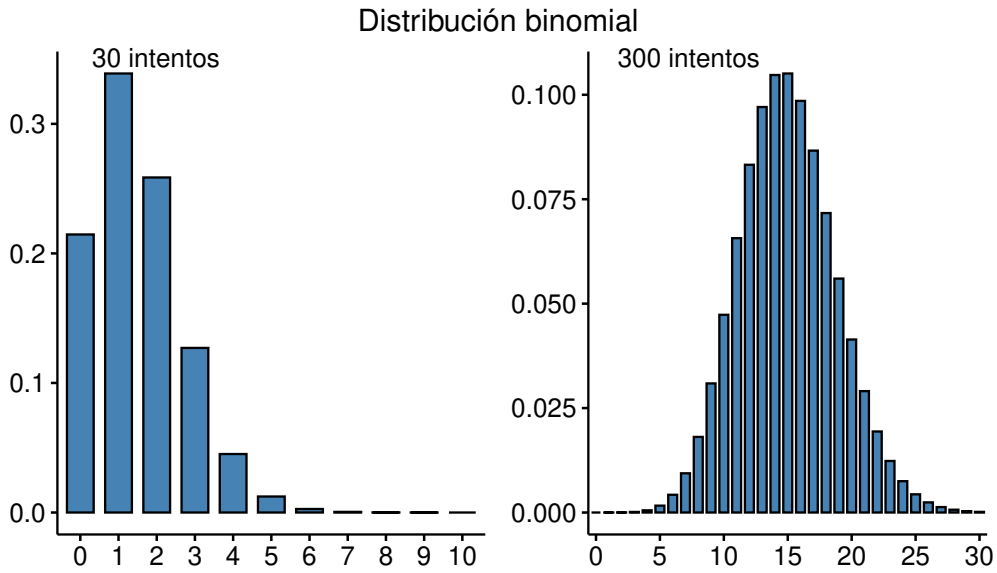


Figura 3.11: distribuciones binomiales de cuántas veces se observa un valor específico (como el valor 20) en 30 (a la izquierda) y 300 (a la derecha) lanzamientos de un dado de 20 lados balanceado.

Aunque las definiciones anteriores las incluyen, es mejor si hacemos más explícitas las condiciones que tenemos verificar antes de "decidir" usar la distribución binomial para modelar la ocurrencia de eventos aleatorios:

1. Los intentos son independientes.
2. La cantidad de intentos (n) es fija.
3. El resultado de cada intento puede ser clasificado como éxito o fracaso.
4. La probabilidad de éxito (p) es la misma para cada intento.

Las funciones que ofrece R para trabajar con esta distribución son:

- `dbinom(x, size, prob)`.
- `pbinom(x, size, prob)`.
- `qbinom(p, size, prob)`.
- `rbinom(n, size, prob)`.

Donde:

- `x` es un vector numérico.
- `p` es un vector de probabilidades.
- `n` es la cantidad de observaciones.
- `size` corresponde al número de intentos.
- `prob` es la probabilidad de éxito de cada intento.

Así, siguiendo con el ejemplo del dado de 20 caras, la probabilidad de obtener exactamente 3 veces un valor 20 en 30 intentos puede calcularse con la llamada `dbinom(3, size = 30, prob = 1/20)`, mientras que `pbinom(3, size = 30, prob = 1/20)` nos entrega la probabilidad de obtener el valor deseado 3 o menos veces en 30 tiros, y `qbinom(0.5, size = 30, prob = 1/20)` nos dirá el valor entero k que nos entrega al menos un 50 % de probabilidad de observar el valor 20 una cantidad menor o igual a k veces en 30 tiros. Las

consideraciones a tener con estas funciones mencionadas para la distribución geométrica también aplican en este caso con la distribución binomial.

En la figura 3.11 podemos observar que la forma de esta distribución depende fuertemente del número de intentos que se considere (`size` en las funciones de R). Cuando este valor es alto, la distribución binomial luce bastante simétrica y se asemeja a una distribución normal, aunque no tiene exactamente la forma de campana de la distribución gaussiana. Esta similitud ofrece una importante ventaja, pues en estos casos es posible usar la distribución normal para **estimar** probabilidades binomiales, evitando así el uso de la compleja fórmula de la ecuación 3.20. Una regla empírica en este respecto es que esta aproximación es válida cuando el tamaño de la muestra de observaciones (n) es lo suficientemente grande para que tanto np como $n(1-p)$ sean mayores o iguales que 10. En este caso, los parámetros de la distribución normal aproximada son los mismos de la distribución binomial ($\mu = np$ y $\sigma = \sqrt{np(1-p)}$).

3.3.4 Distribución binomial negativa

Una **variable aleatoria binomial negativa** corresponde al número de fracasos que se observan en un proceso Bernoulli antes de obtener un número determinado de éxitos. La **distribución binomial negativa**, entonces, describe la probabilidad de encontrar el k-ésimo éxito al n-ésimo intento. Puede considerarse una generalización de una distribución geométrica que está limitada a observar el primer éxito. Tampoco se puede modelar como una distribución binomial, puesto que como señalan Diez et al. (2017, p. 155), "en el caso binomial, en general se tiene una cantidad fija de intentos y se considera la cantidad de éxitos. En el caso binomial negativo, se examina cuántos intentos se necesitan para observar una cantidad fija de éxitos y se requiere que la última observación sea un éxito"[3].

Como en la distribución anterior, es conveniente hacer explícitas las condiciones que debemos verificar para la distribución binomial negativa:

1. Los intentos son independientes.
2. El resultado de cada intento puede ser clasificado como éxito o fracaso.
3. La probabilidad de éxito (p) es la misma para cada intento.
4. El último intento debe ser un éxito.

La función de probabilidad para esta distribución está dada por la ecuación 3.21. La varianza y la desviación estándar están dadas por las ecuaciones 3.22 y 3.23 (Devore, 2008, p. 120). Esto queda ejemplificado en la figura 3.12 que corresponde a las probabilidades del número de tiros fallidos antes de obtener el valor 20 dos veces.

$$\Pr(k\text{-ésimo éxito al } n\text{-ésimo intento}) = \binom{n-1}{k-1} p^k (1-p)^{n-k} \tag{3.21}$$

$$\mu = \frac{k(1-p)}{p} \tag{3.22}$$

$$\sigma = \sqrt{\frac{k(1-p)}{p^2}} \tag{3.23}$$

Nuevamente, R dispone de cuatro funciones que permiten trabajar con esta distribución:

- `dnbinom(x, size, prob)`.
- `pnbinom(q, size, prob, lower.tail)`.
- `qnbinom(p, size, prob, lower.tail)`.
- `rnbinom(n, size, prob)`.

Donde:

[3] Traducción libre de los autores.

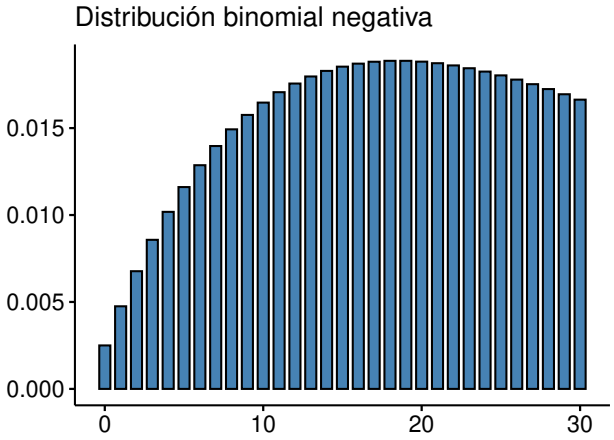


Figura 3.12: ejemplo de distribución binomial negativa.

- `x, q` son vectores de cuantiles (enteros no negativos).
- `p` es un vector de probabilidades.
- `n` es la cantidad de observaciones requerida.
- `size` corresponde al de éxitos buscado.
- `prob` es la probabilidad de éxito de cada intento.
- `lower.tail` es análogo al de la función `pnorm`.

Siguiendo con el ejemplo del dado de 20 caras, la probabilidad de observar exactamente 10 fracaso antes de obtener 3 veces el valor 20 se obtiene con la llamada `dnbinom(10, size = 3, prob = 1/20)`, mientras que `pnbinom(10, size = 3, prob = 1/20)` nos entrega la probabilidad de hacer 10 o menos intentos fallidos antes de observar el valor deseado 3 veces, y `qnbinom(0.5, size = 3, prob = 1/20)` nos dirá el valor entero k con al menos un 50 % de probabilidad de observar obtener k o menos tiros fallidos antes que el valor deseado 3 veces. Las consideraciones a tener con estas funciones mencionadas para las distribuciones discretas anteriores también aplican a la distribución binomial negativa.

3.3.5 Distribución de Poisson

Una **variable aleatoria Poisson** corresponde a la probabilidad de observar un número determinado de ocurrencias de un evento durante un cierto periodo de tiempo, a partir de la frecuencia de ocurrencia media del evento. De esta forma, por ejemplo, es útil para conocer la probabilidad de que ocurran 100 contagios de influenza en una semana entre las y los usuarios de un recorrido de bus del transporte público de Santiago, sabiendo que en promedio se producen 12 contagios por día. También es el tipo de variable aleatoria que se usa para estimar la demanda en un Call Center, y en el mundo de la informática se usa para modelar el número de tareas que llegan a la CPU para su procesamiento.

La distribución de Poisson tiene una función de probabilidad definida por la ecuación 3.24, donde λ es la tasa o cantidad de eventos que se espera observar en un lapso de tiempo dado y k puede tomar cualquier valor entero no negativo. La media de esta distribución está dada por λ y la desviación estándar por $\sqrt{\lambda}$.

$$\Pr(\text{observar } k \text{ eventos}) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{3.24}$$

La figura 3.13 muestra la distribución obtenida para el ejemplo mencionado de los contagios de influenza en una semana. Podemos ver que es más probable que ocurran entre 80 y 90 contagios de que ocurran 100.

Las funciones de R para esta distribución son:

- `dpois(x, lambda)`.

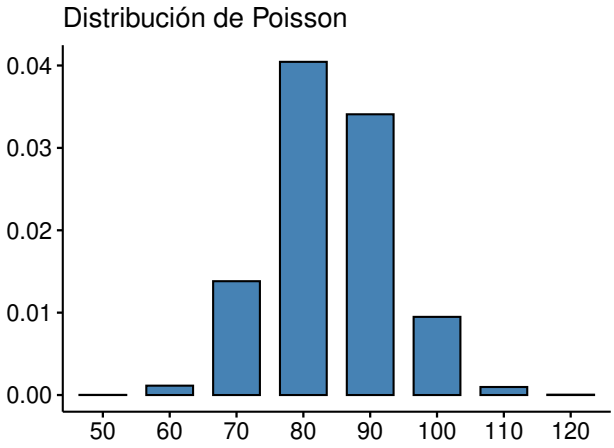


Figura 3.13: ejemplo de distribución de Poisson con $\lambda = 12 \cdot 7 = 84$.

- `ppois(q, lambda, lower.tail)`.
- `qqpois(p, lambda, lower.tail)`.
- `rpois(n, lambda)`.

Donde:

- `x, q` son vectores de cuantiles (enteros no negativos).
- `p` es un vector de probabilidades.
- `n` es la cantidad de observaciones.
- `lambda` es un vector no negativo de medias.
- `lower.tail` es análogo al de la función `pnorm`.

3.4 EJERCICIOS PROPUESTOS

3.1 Da un ejemplo de una pregunta de investigación sobre las asignaturas comunes en ingeniería que pueda modelarse con una variable aleatoria binomial. ¿Cuál sería el valor esperado? ¿Cuál sería la varianza? ¿Cómo te imaginas su función de masa de probabilidad?

3.2 Da un ejemplo de una pregunta de investigación sobre los conciertos realizados en Santiago que pueda modelarse con una variable aleatoria geométrica. ¿Cuál sería el valor esperado? ¿Cuál sería la varianza? ¿Cómo te imaginas su función de masa de probabilidad?

3.3 Da un ejemplo de una pregunta de investigación sobre el estado de la salud mental de estudiantes universitarios que pueda modelarse con una variable aleatoria binomial negativa. ¿Cuál sería el valor esperado? ¿Cuál sería la varianza? ¿Cómo te imaginas su función de masa de probabilidad?

3.4 Según el Reporte Mensual de Empleo, las siguientes son las estadísticas (media $\pm$ desviación estándar) para las seis variables relevantes que se han estudiado en los últimos cinco años:

1. Número de personas despedidas: 64.675 ± 8.321.
2. Número de personas renunciadas: 118.543 ± 17.936.
3. Número de personas jubiladas: 97.092 ± 11.147.
4. Número de empleos creados: 24.715 ± 10.832.
5. Número de personas contratadas: 301.345 ± 27.261.
6. Número de personas entrando a la fuerza de trabajo: 26.444 ± 29.440.

Con esta información, calcula la media y la desviación estándar de:

(a) Caída neta del empleo: $(d) - (a) - (b) - (c)$.

(b) Subida neta del empleo: $(e) - (a) - (b) - (c)$.

(c) Caída neta del desempleo: $(a) + (b) + (e) + (f)$.

(d) Vacancia del empleo: $(d) - (e)$.

3.5 La probabilidad de que un estudiante universitario chileno seleccionado al azar sea VIH positivo es 0,013. ¿Cuáles serían la media y la desviación estándar de esta variable?

3.6 Considerando la información anterior, en promedio ¿a cuántos estudiantes universitarios se debería revisar hasta encontrar a un VIH positivo?

3.7 Considerando la información anterior, y si el Departamento de Salud de una Universidad chilena controla a 50 estudiantes por día durante una semana de clases (lunes a viernes), ¿cuál sería el número promedio de VIH positivos detectados cada día? ¿Con qué varianza?

3.8 Si la Universidad del ejercicio anterior dispone de 10 paquetes de tratamiento de VIH para estudiantes, ¿cómo podría saber a cuántos estudiantes debería examinar para poder asignarlos (suponiendo que todo estudiante VIH positivo acepta el tratamiento)?

3.9 Da un ejemplo de una pregunta de investigación sobre las asignaturas comunes en ingeniería que pueda modelarse con una distribución normal. Indica y justifica los parámetros para tal distribución.

3.10 Da un ejemplo de una pregunta de investigación sobre los conciertos realizados en Santiago que pueda modelarse con una distribución normal. Indica y justifica los parámetros para tal distribución.

3.11 Da un ejemplo de una pregunta de investigación sobre el estado de la salud mental de estudiantes universitarios que pueda modelarse con una distribución normal. Indica y justifica los parámetros para tal distribución.

3.5 BIBLIOGRAFÍA DEL CAPÍTULO

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

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

Freund, R. J., & Wilson, W. J. (2003). *Statistical Methods* (2.ª ed.). Academic Press.

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 et al. (2017, pp. 168-202) y Field et al. (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.

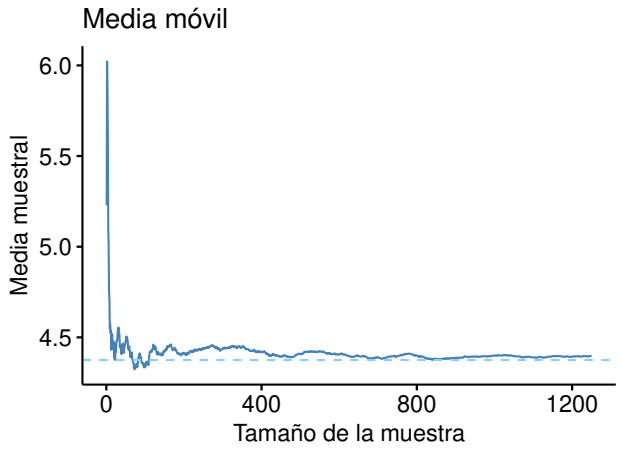


Figura 4.1: medias obtenidas al agregar a la muestra un elemento cada vez.

Script 4.1: representación gráfica de la media móvil.

```r
library(ggpubr)

# Establecer la semilla para generar los mismos números aleatorios
# cada vez que se ejecute el script.
set.seed(9437)

# Generar aleatoriamente una población de tamaño 1500
# (en este caso, siguiendo una distribución 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")
```

```r

# 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 = "steelblue",
            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 = "skyblue1", 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 punteada 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.

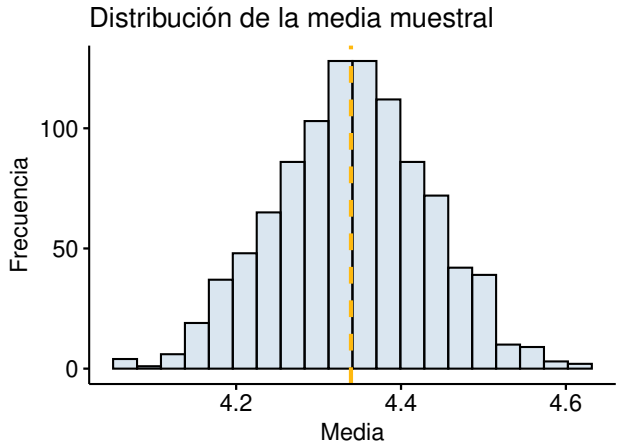


Figura 4.2: distribución muestral de la media para muestras con 100 observaciones.

Script 4.2: distribución muestral de la media.

```r
library(ggpubr)

# Fijar la semilla para generar números aleatorios
```

```r
set.seed(94)

# Generar aleatoriamente una población de tamaño 1500
# (en este caso, siguiendo una distribución 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")

# 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, fill = "steelblue", alpha = 0.2,
                 title = "Distribución de la media muestral",
                 xlab = "Media", ylab = "Frecuencia",
                 )

# Agregar línea vertical con la media de la población.
g <- g + geom_vline(aes(xintercept = media_poblacion),
                    color = "darkgoldenrod1", linetype = 2,
                    linewidth = 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.

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

Donde σ es la desviación estándar de la población 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.
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

[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.

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 et al., 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 sería 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: $\mu_N < \mu_A$

Notemos que estas hipótesis dan por descartado la posibilidad de que $\mu_N > \mu_A$. Esto es útil por razones que veremos más adelante. Obviamente en otros casos podría interesar solamente si valor alternativo es mayor que el valor nulo, descartando la posibilidad de que sea menor.

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:

$$\overline{x}_N - z^* \cdot SE_{\overline{x}} = 527{,}9 - 1{,}64 \cdot 1{,}2 = 525{,}932$$

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: $[-\infty; 525{,}932]$.

Ahora el valor $\mu_A = 530$ [ms] cae 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 et al.

(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 $n = 1.600$ observaciones. Esta distribución se vería como muestra la figura 4.3, creada mediante el script 4.3.

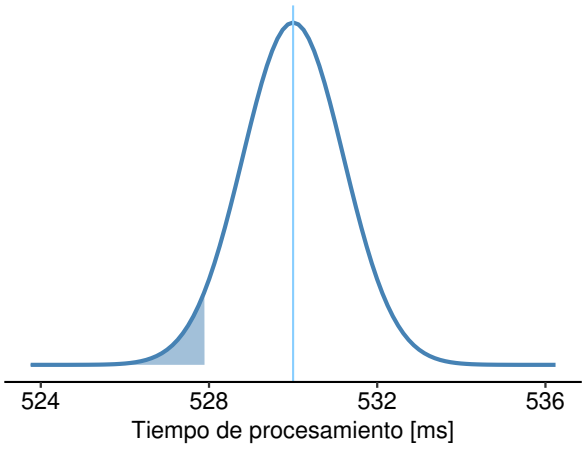


Figura 4.3: probabilidad de encontrar una media igual o menor que $\overline{x} = 527{,}9$ [ms] en la distribución muestral con $\mu_{\overline{x}} = 530$ y $\sigma_{\overline{x}} = 1{,}2$.

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)
muestra <- data.frame(x, y)

# Graficar la muestra
```

```
19 g <- ggplot(data = muestra, aes(x))
20 g <- g + stat_function(fun = dnorm,
21                         args = list(mean = media_poblacion_antiguo,
22                                     sd = error_est),
23                         colour = "steelblue", linewidth = 1)
24 g <- g + ylab("")
25 g <- g + scale_y_continuous(breaks = NULL)
26 g <- g + scale_x_continuous(name = "Tiempo de procesamiento [ms]")
27 g <- g + theme_pubr()
28
29 # Colorear el área igual o menor que la media observada
30 g <- g + geom_area(data = subset(muestra, x < media_muestra_nuevo),
31                    aes(y = y), colour = "steelblue", fill = "steelblue",
32                    alpha = 0.5)
33
34 # Agregar una línea vertical para el valor nulo
35 g <- g + geom_vline(aes(xintercept = media_poblacion_antiguo),
36                     color = "skyblue1", linetype = 1)
37
38 print(g)
39
40 # Calcular el valor Z para la muestra
41 Z <- (media_muestra_nuevo - media_poblacion_antiguo) / error_est
42
43 # Calcular y mostrar el valor p
44 p_1 <- pnorm(Z, lower.tail = TRUE)
45 cat("Valor p: ", p_1, "\n")
46
47 # También se puede calcular el valor p directamente a partir de la
48 # distribución muestral definida por el valor nulo y el error estándar.
49 p_2 <- pnorm(media_muestra_nuevo, mean = media_poblacion_antiguo,
50              sd = error_est)
51 cat("Valor p: ", p_2, "\n")
```

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 rechaza 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.

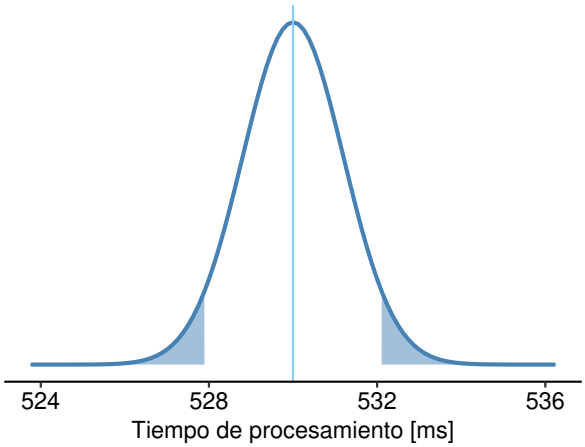


Figura 4.4: cuando la prueba de hipótesis es bilateral, se deben colorear ambas colas.

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

```r
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)
muestra <- data.frame(x, y)

# Graficar la muestra
g <- ggplot(data = muestra, 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(muestra, 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
```

```r
corte_x <- qnorm(1 - area_inferior,
                 mean = media_poblacion_antiguo,
                 sd = desv_est)
g <- g + geom_area(data = subset(muestra, 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 = "skyblue1", linetype = 1)

print(g)

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

# Calcular y mostrar 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 et al. (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

4.1 ¿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.

4.2 ¿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.

4.3 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?

4.4 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 _____ _____.

4.5 Se sabe que una prueba para medir el coeficiente intelectual de jóvenes de 18 años produce puntuaciones que siguen una distribución $\mathcal{N}(100, 100)$ (normal con $\mu = 100$ y $\sigma = 10$).

(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?

4.6 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.

4.7 Considerando el enunciado de la pregunta anterior:

(a) 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?

(b) Para el intervalo de confianza anterior, ¿cuál sería un error de tipo I?

(c) Conociendo el intervalo de confianza anterior, ¿es posible cometer un error de tipo II? Explica.

4.8 Si una hipótesis nula es falsa, aumentar el nivel de significación para un tamaño de muestra dado, ¿reduce la probabilidad de rechazarla?

4.9 ¿Qué significa que un estadístico tenga un valor p de 0,025?

4.10 Si una hipótesis nula es rechazada a un nivel de significación de 0,01, ¿será rechazada a un nivel de significación 0,05? Explica.

4.11 Si una hipótesis nula es rechazada por una prueba unilateral (una cola), ¿será también rechazada por una prueba bilateral (dos colas)? Explica.

4.12 Explica por qué se incrementa la probabilidad de cometer errores tipo I al cambiar de una prueba de hipótesis bilateral a otra unilateral.

4.13 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 sobre su validez.

4.14 Da el ejemplo de un estudio en que es más dañino cometer un error tipo II que un error tipo I.

4.15 Si para un estudio de una determinada variable aleatoria categórica es igualmente dañino cometer errores de tipo I como errores tipo II con hipótesis sobre la probabilidad de un evento:

(a) Dibuja la distribución de una muestra (un gráfico de barras, por ejemplo) para la que el contraste de hipótesis con nivel de significación 0,05 sea confiable.

(b) Dibuja la distribución de una muestra 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.

4.16 Si para un estudio de una determinada variable aleatoria continua es igualmente dañino cometer errores de tipo I como errores tipo II con hipótesis sobre su media:

(a) Dibuja la distribución de una muestra de tamaño 16 (un diagrama de caja o un histograma, 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.

4.17 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?

4.8 BIBLIOGRAFÍA DEL CAPÍTULO

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