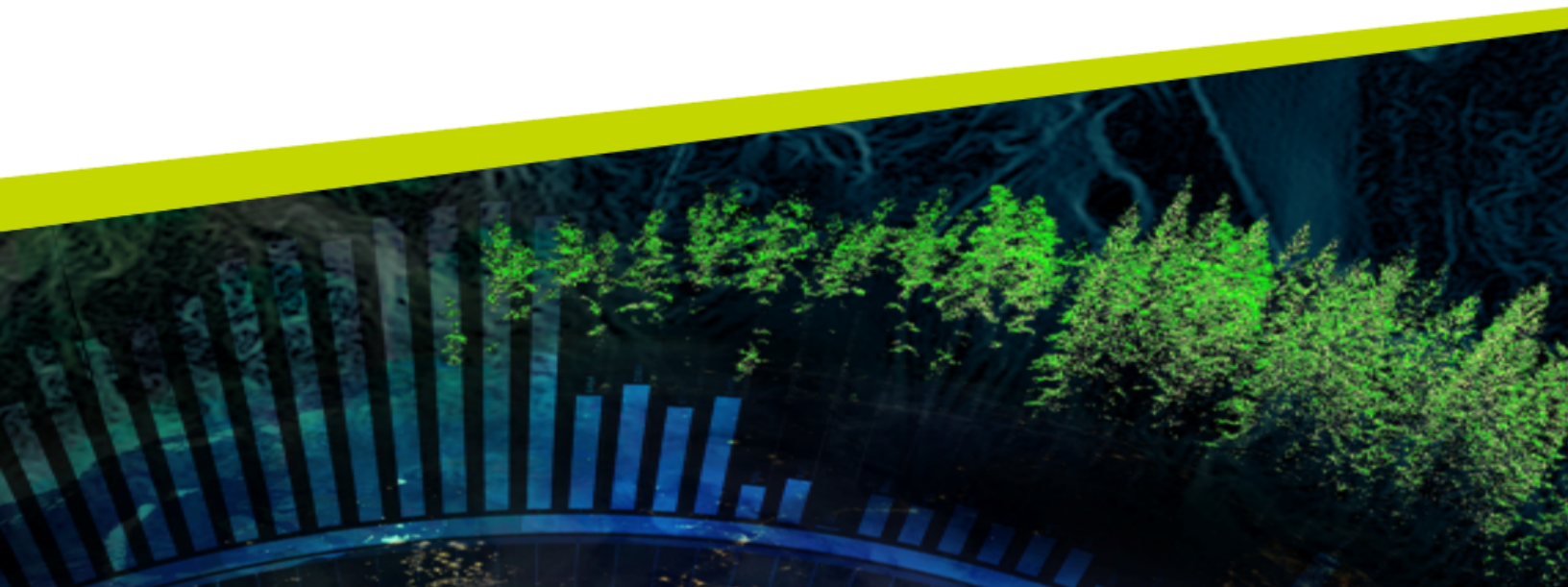
INFERENCIA Y MODELOS ESTADÍSTICOS

Jacqueline Köhler C. y José Luis Jara V.



CAPÍTULO 6. PODER ESTADÍSTICO

En el capítulo 4 estudiamos el procedimiento para someter hipótesis a prueba, junto con los errores de decisión que podríamos cometer:

- Error tipo I: rechazar H_0 en favor de H_A cuando H_0 es en realidad verdadera.
- Error tipo II: no rechazar H_0 en favor de H_A cuando H_A es en realidad verdadera.

Allí conocimos el nivel de significación, α, como herramienta para representar y, de alguna manera, controlar la probabilidad de cometer un error de tipo I, con lo que la preocupación se centra en controlar la ocurrencia de esta clase de errores, desviando la atención de los errores de tipo II. Esto se debe a que la hipótesis nula representa el *status quo*, es decir, mantener las cosas y creencias tal como están y, por ende, cuando no se rechaza H_0, no suele requerirse tomar ninguna acción. En contraste, la hipótesis alternativa describe un cambio de condiciones, por lo que rechazar H_0 en favor de H_A usualmente conlleva un esfuerzo, mayor costo, para adaptarse o aprovechar las nuevas condiciones.

Sin embargo, en el capítulo 4 también vimos que el valor de α debe ser acorde con las consecuencias de cometer errores tanto de tipo I como de tipo II, ¡pero no sabemos cómo se relaciona el nivel de significación con los errores de tipo II!

Así como el nivel de significación α corresponde a la probabilidad de cometer errores de tipo I, definimos ahora β como la probabilidad de cometer errores de tipo II. α y β están relacionados: **para un tamaño fijo de la muestra: al reducir $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ aumenta, y viceversa**. Este fenómeno se evidencia con mayor fuerza mientras más pequeña sea la muestra. No obstante, en la práctica resulta más interesante conocer la probabilidad de **no** cometer errores de tipo II. Esto nos lleva a un nuevo concepto: el **poder estadístico** de una prueba de hipótesis, también llamado "potencia estadística", dado por $1 - \beta$, que se define como **la probabilidad de correctamente rechazar H_0 cuando es falsa**.

Otra forma de entender la noción de poder de una prueba es qué tan propensa es esta para distinguir un efecto real de una simple casualidad, lo que nos lleva a la noción de **tamaño del efecto**, que corresponde a una cuantificación de la diferencia entre dos grupos, o del valor observado con respecto al valor nulo.

En el capítulo 5 conocimos la prueba t y la prueba de proporciones para inferir acerca de dos medias y dos probabilidades de éxito, respectivamente. En este contexto, el tamaño del efecto corresponde a qué tan grande es la diferencia real entre cada par de parámetros. Si quieres aprender más sobre estos conceptos, puedes consultar las fuentes en las que se basa este capítulo: Diez y col. (2017, pp. 239-245) y Freund y Wilson (2003, pp. 123-138).

6.1 PODER, NIVEL DE SIGNIFICACIÓN Y TAMAÑO DE LA MUESTRA

En la introducción de este capítulo vimos que el poder estadístico corresponde a la probabilidad de **no** cometer un error de tipo II, y que está muy relacionado con el tamaño de la muestra. También mencionamos que existe una relación entre el poder y el nivel de significación, la cual exploraremos en esta sección.

La figura 6.1 muestra cuatro curvas de poder para la prueba t de Student de una muestra con desviación estándar $s = 1$ y valor nulo $\mu_0 = 0$. En ella, el tamaño del efecto está representada en la misma escala de la variable, aunque en la sección siguiente veremos otra alternativa. La curva roja considera $\alpha = 0,05$ y $n = 6$; la azul, $\alpha = 0,01$ y $n = 6$; la verde, $\alpha = 0,05$ y $n = 10$, y la naranja, $\alpha = 0,01$ y $n = 10$. En ella podemos observar que:

- El poder de la prueba aumenta mientras mayor es el tamaño del efecto (en este caso, la distancia entre el valor nulo y la media de la muestra).
- A medida que el tamaño del efecto disminuye (es decir, el estimador se acerca al valor nulo), el poder se aproxima al nivel de significación.
- Usar un valor de α más exigente (menor), manteniendo constante el tamaño de la muestra, hace que la curva de poder sea más baja para cualquier tamaño del efecto (lo que verifica la relación entre α y β).
- Usar una muestra más grande aumenta el poder de la prueba para cualquier tamaño del efecto distinto de 0.

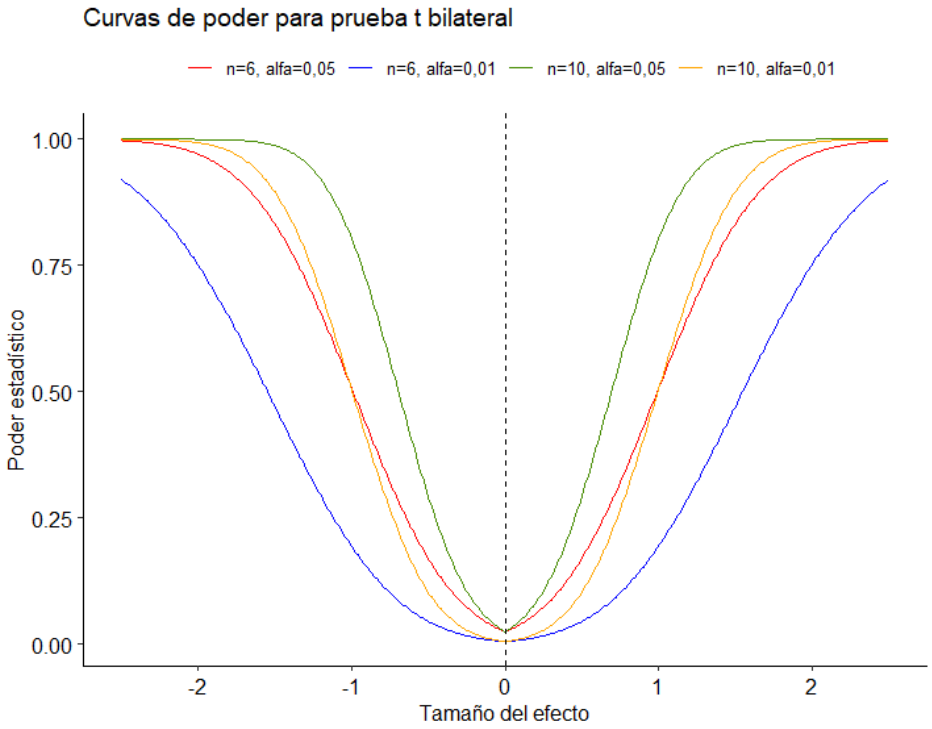


Figura 6.1: poder estadístico para prueba t bilateral.

La figura 6.1 fue creada mediante el script 6.1, que hace uso de la función `power.t.test()` y que es explicada en detalle más adelante en este capítulo.

De manera similar, la figura 6.2 considera las mismas muestras y los mismos niveles de significación que la figura 6.1, pero ahora para una prueba t unilateral. En ella se aprecia que la gran desventaja de las pruebas unilaterales es que el poder tiende a cero a medida que el tamaño del efecto aumenta en sentido contrario a la hipótesis alternativa, por lo que no sería posible detectar una diferencia en el sentido opuesto aunque fuese muy grande (pues no hay una región de rechazo en dicho sentido). El script empleado para la construcción de la figura 6.2 es idéntico al script 6.1, excepto porque el argumento `alternative` toma como valor `"one.sided"` en las llamadas a `power.t.test()`.

Script 6.1: poder estadístico para prueba t bilateral.

```
library(ggpubr)
library(tidyverse)

# Generar un vector con un rango de valores para la efecto
# de medias.
efecto <- seq(-2.5, 2.5, 0.01)

# Calcular el poder para una prueba t bilareral, para cada tamaño
# del efecto, asumiendo una muestra con desviación estándar igual a 1.
```

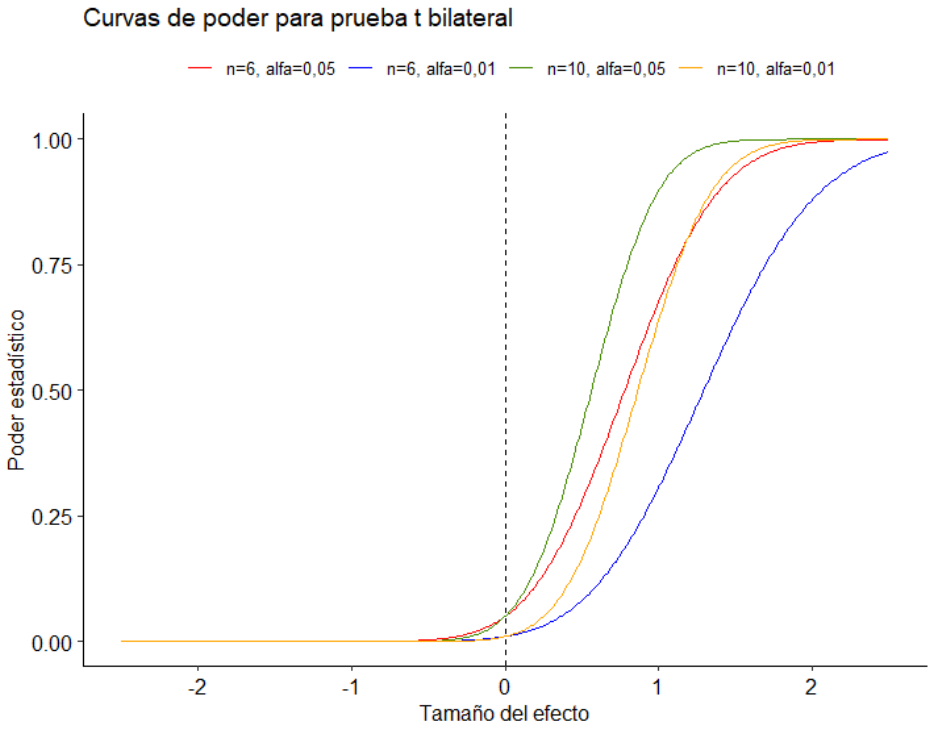


Figura 6.2: poder estadístico para prueba t unilateral.

```r
# Se consideran 4 escenarios para calcular el poder:
# 1. Una muestra de tamaño 6 y nivel de significación 0.05.
# 2. Una muestra de tamaño 6 y nivel de significación 0.01.
# 3. Una muestra de tamaño 10 y nivel de significación 0.05.
# 4. Una muestra de tamaño 10 y nivel de significación 0.01.
n_6_alfa_05 <- power.t.test(n = 6,
                        delta = efecto,
                        sd = 1,
                        sig.level = 0.05,
                        type = "one.sample",
                        alternative = "two.sided")$power

n_6_alfa_01 <- power.t.test(n = 6,
                        delta = efecto,
                        sd = 1,
                        sig.level = 0.01,
                        type = "one.sample",
                        alternative = "two.sided")$power

n_10_alfa_05 <- power.t.test(n = 10,
                         delta = efecto,
                         sd = 1,
                         sig.level = 0.05,
                         type = "one.sample",
                         alternative = "two.sided")$power

n_10_alfa_01 <- power.t.test(n = 10,
                             delta = efecto,
                             sd = 1,
                             sig.level = 0.01,
```

```
                                type = "one.sample",
                                alternative = "two.sided")$power

# Construir matriz de datos en formato ancho.
datos <- data.frame(efecto, n_6_alfa_05, n_6_alfa_01,
                    n_10_alfa_05, n_10_alfa_01)

# Llevar a formato largo.
datos <- datos %>% pivot_longer(!"efecto",
                                names_to = "fuente",
                                values_to = "poder")

# Formatear fuente como variable categórica.
niveles <- c("n_6_alfa_05", "n_6_alfa_01", "n_10_alfa_05",
             "n_10_alfa_01")

etiquetas <- c("n=6, alfa=0,05", "n=6, alfa=0,01", "n=10, alfa=0,05",
               "n=10, alfa=0,01")

datos[["fuente"]] <- factor(datos[["fuente"]], levels = niveles,
                            labels = etiquetas)

# Graficar las curvas de poder.
g <- ggplot(datos, aes(efecto, poder,  colour = factor(fuente)))
g <- g + geom_line()
g <- g + labs(colour = "")
g <- g + ylab("Poder estadístico")
g <- g + xlab("Tamaño del efecto")

g <- g + scale_color_manual(values=c("red", "blue", "chartreuse4",
                                     "orange"))

g <- g + theme_pubr()
g <- g + ggtitle("Curvas de poder para prueba t bilateral")
g <- g + geom_vline(xintercept = 0, linetype = "dashed")

print(g)
```

La figura 6.3 muestra las curvas de poder para una prueba t unilateral y otra bilateral, ambas para una muestra de tamaño 6, desviación estándar $s = 1$ y $\alpha = 0,05$. En ella se evidencia claramente la ventaja de las pruebas unilaterales: cuando el tamaño del efecto aumenta en el sentido de la hipótesis alternativa, el poder es mayor que para una prueba bilateral.

Debemos notar que, si bien esta discusión se ha hecho con la prueba t de Student, las ideas expuestas aquí aplican a cualquier prueba estadísticas y los estadísticos que estas consideran.

Es deseable que las pruebas que se empleen para docimar hipótesis tengan un alto poder y, si hay más de una prueba disponible, se debe escoger la más poderosa. No obstante, los cálculos del poder suelen ser altamente complejos. Afortunadamente, la teoría permite en muchos casos conocer la prueba con mayor poder posible ante cualquier hipótesis alternativa, nivel de significación y tamaño de muestra (siempre que se cumplan las condiciones de base). Estas pruebas reciben el nombre de **uniformemente más poderosas**.

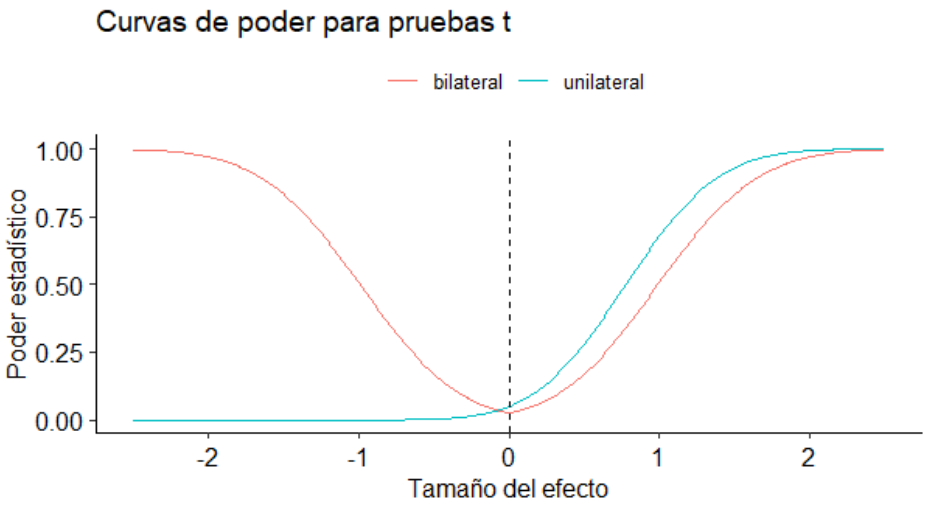


Figura 6.3: poder estadístico para pruebas t.

6.2 TAMAÑO DEL EFECTO

El problema que podríamos tener al considerar el tamaño del efecto en la misma escala de la variable estudiada, como hemos hecho hasta ahora, es que esta escala varía de variable en variable. Para poder hacer comparaciones con mayor libertad, existen diferentes **medidas estandarizadas de efecto** que podemos usar.

Al comparar dos medias, existe la llamada ***d* de Cohen** (Kassambara, 2019). En términos generales, se considera que $d = 0,2$ es un efecto pequeño (imperceptible a simple vista), $d = 0,5$ es un efecto mediano (probablemente perceptible a simple vista) y $d = 0,8$, un efecto grande (definitivamente perceptible a simple vista).

Cuando se trabaja con una muestra, la d de Cohen se calcula como en la ecuación 6.1, donde:

- $\overline{x}$: media muestral.
- μ_0: media teórica para el contraste (valor nulo).
- s: desviación estándar de la muestra con $n - 1$ grados de libertad.

$$d = \frac{\overline{x} - \mu_0}{s} \tag{6.1}$$

Cuando se trabaja con la diferencia de dos medias no pareadas, si el tamaño de la muestra es mayor a 50 elementos, se calcula como muestra la ecuación 6.2, y para muestras pequeñas se aplica un factor de corrección, como indica la ecuación 6.3, donde:

- $\overline{x}_1$, $\overline{x}_2$: medias muestrales de cada grupo.
- n_1 y n_2 son los tamaños de ambas muestras.
- s_p: desviación estándar agrupada, dada por la ecuación 6.4[1].

$$d = \frac{\overline{x}_1 - \overline{x}_2}{s_p} \tag{6.2}$$

$$d = \frac{\overline{x}_1 - \overline{x}_2}{s_p} \cdot \frac{n_1 + n_2 - 3}{n_1 + n_2 - 2,25} \tag{6.3}$$

[1]Note que esta corresponde a la raíz de la varianza agrupada descrita en 5.5

$$s_p = \sqrt{\frac{\sum(x - \overline{x}_1)^2 + \sum(x - \overline{x}_2)^2}{n_1 + n_2 - 2}} \tag{6.4}$$

En el caso de la variante de Welch para la prueba t de Student para muestras independiente, la fórmula para el cálculo de la d de Cohen es ligeramente diferente, como puede apreciarse en la ecuación 6.5.

$$d = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2 + s_2^2}{2}}} \tag{6.5}$$

Por último, las ecuaciones 6.6 y 6.7 muestran cómo se calcula la d de Cohen en el caso de la prueba t con muestras pareadas grandes ($n > 50$) y pequeñas, respectivamente, donde D corresponde a las diferencias entre las observaciones pareadas.

$$d = \frac{\overline{x}_D}{s_D} \tag{6.6}$$

$$d = \frac{\overline{x}_D}{s_D} \cdot \frac{n_1 - 2}{n_1 - 1,25} \tag{6.7}$$

Por otro lado, se tiene la ***h* de Cohen** (Cohen, 2013) que es la medida análoga para usarse cuando se trabaja con probabilidades o proporciones, manteniendo la idea que valores $h = 0,2$, $h = 0,5$ y $h = 0,8$ son indicadores de un efecto pequeño, mediano y grande respectivamente.

Suponiendo que se tienen dos probabilidades, o proporciones, digamos p_1 y p_2, que en el caso de trabajar con una muestra p_2 podría referirse al valor nulo, la h de Cohen se calcula como indica la ecuación 6.8, donde:

$$h = \varphi(p_1) - \varphi(p_2) \tag{6.8}$$

con

$$\varphi(p) = 2 \arcsin \sqrt{p}$$

La transformación en la ecuación 6.8 es necesaria porque, en una distribución binomial, la varianza no es constante, sino que una función de la media que tiene su valor máximo cuando la probabilidad de éxito es 0,5, y disminuyendo a cero cuando esta alcanza valores cero y uno. Por esto, se utilizan **transformaciones estabilizadoras de la varianza** al trabajar con frecuencias (datos binomiales), y dos de las más comunes son las basadas en las funciones *logit* y *arcoseno*. Estas transformaciones también se utilizan para datos porcentuales, aún cuando estos pueden no seguir una distribución binomial (Warton & Hui, 2011).

Existen muchas otras medidas estandarizadas del tamaño del efecto que trabajan con otros estadísticos (varianza, coeficientes de regresión, etc.) y otras pruebas de hipótesis. Las hiremos mencionando a medida que vayamos estudiándolas más adelante.

6.3 PODER, TAMAÑO DEL EFECTO Y TAMAÑO DE LA MUESTRA

Mencionamos en páginas anteriores que el poder puede también entenderse como qué tan propensa es una prueba estadística para distinguir un efecto real de una simple casualidad, y que podemos cuantificar este efecto.

Una gran ventaja del poder estadístico es que nos sirve para determinar el tamaño adecuado de la muestra para detectar un cierto tamaño del efecto. La figura 6.4, elaborada con el script 6.2, muestra el aumento del poder estadístico a medida que el tamaño de la muestra aumenta (para un tamaño del efecto y nivel de significación fijos). En ella se aprecia que, a medida que el tamaño de la muestra crece, el poder estadístico también crece asintóticamente a 1, acercándose a tener la certeza de rechazar la hipótesis nula si esta es falsa.

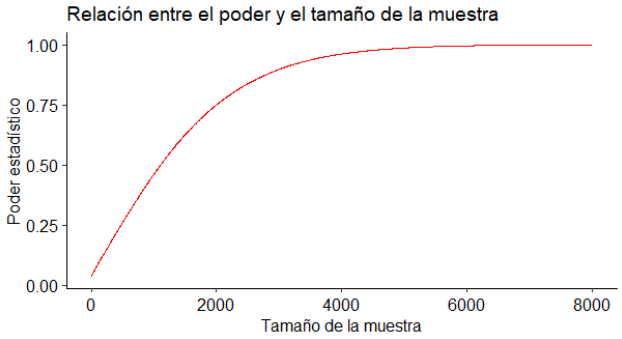


Figura 6.4: aumento del poder estadístico a medida que crece el tamaño de la muestra (manteniendo fijos el tamaño del efecto y el nivel de significación).

Script 6.2: aumento del poder estadístico a medida que crece el tamaño de la muestra.

```r
library(ggpubr)

# Generar un vector con un rango para el tamaño de la muestra.
n <- seq(5, 8000, 5)

# Definir constantes
desv_est <- 6
alfa <- 0.05
tam_efecto <- 0.5

# Se calcula el poder con que se detecta el tamaño del efecto para
# cada tamaño de la muestra, asumiendo una prueba bilateral para
# una sola muestra.
poder <- power.t.test(n = n,
                      delta = tam_efecto,
                      sd = desv_est,
                      sig.level = alfa,
                      type = "two.sample",
                      alternative = "two.sided")$power

# Crear un data frame.
datos <- data.frame(n, poder)

# Graficar la curva de poder.
g <- ggplot(datos, aes(n, poder))
g <- g + geom_line(colour = "red")
g <- g + ylab("Poder estadístico")
g <- g + xlab("Tamaño de la muestra")
g <- g + theme_pubr()
g <- g + ggtitle("Relación entre el poder y el tamaño de la muestra")

print(g)
```

Nuevamente, si bien la figura 6.4 se construyó considerando la prueba t de Student, esta relación entre tamaño de la muestra y poder estadístico se da con todas las pruebas de hipótesis y los estadísticos que estas

consideran.

6.4 CÁLCULO TEÓRICO DEL PODER

Como ya hemos mencionado a lo largo de este capítulo, el poder es la probabilidad de correctamente rechazar H_0 cuando es falsa, lo que equivale a la probabilidad de distinguir un efecto real de una mera casualidad. Ahora veremos algunos ejemplos de cómo podemos usar el poder.

Lola Drones, estudiante de computación, ha diseñado dos nuevos algoritmos (A y B) que resuelven un mismo problema como parte de su trabajo de titulación. Lola desea saber si existe diferencia entre los tiempos de ejecución de ambos algoritmos. Para ello, ha decidido realizar una prueba t con muestras pareadas, con un nivel de significación $\alpha = 0,05$, usando para ello 36 instancias del problema de tamaño fijo que se ejecutan bajo iguales condiciones con cada algoritmo. Además, Lola ya sabe que la diferencia en el tiempo de ejecución sigue una distribución normal con desviación estándar $\sigma = 12$ milisegundos. Así, Lola ha formulado las siguientes hipótesis:

H_0: $\mu_{(A_i - B_i)} = 0$, es decir que la media de las diferencias en el tiempo de ejecución necesitado por los algoritmos A y B, para cada posible instancia i, es cero
H_A: $\mu_{(A_i - B_i)} \neq 0$

La figura 6.5 muestra cómo sería la distribución de la muestra (media de las diferencias en los tiempos de ejecución) si la hipótesis nula (H_0) fuese cierta, con las áreas correspondientes a la región de rechazo de H_0 coloreadas.

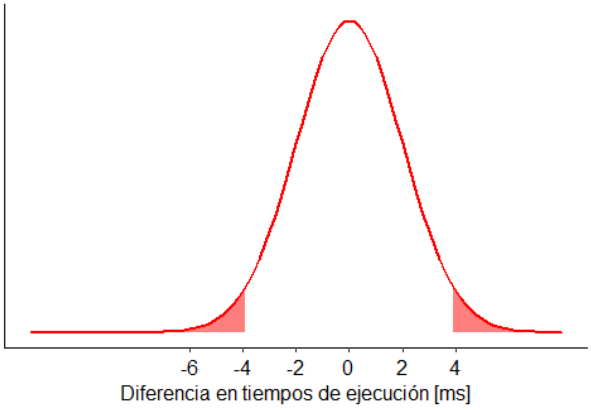


Figura 6.5: distribución de la diferencia de medias muestrales del tiempo de ejecución, señalando zonas de rechazo de la hipótesis nula.

Supongamos por un momento que, en realidad, el algoritmo B es en promedio 4 milisegundos más rápido que el algoritmo A. En este caso tendríamos que la media de las diferencias es de -4 [ms], correspondiente al tamaño del efecto. En este caso, la verdadera distribución muestral sería como muestra la curva azul de la figura 6.6 (ver script 6.3). Al superponer esta nueva curva a la que ya teníamos bajo el supuesto de que la hipótesis nula fuera verdadera, vemos que el área de la curva real que se situaría dentro de la región de rechazo de la curva teórica es aquella coloreada en azul. Esta área corresponde al poder de la prueba t, que en este caso es de $0,516$ de acuerdo al análisis teórico (ver script 6.3, líneas 77–86). Puesto que el poder corresponde a la probabilidad de **no** cometer un error de tipo II, de acuerdo al resultado obtenido se tiene que $\beta = 0,484$. ¡Lola no sería capaz de detectar una diferencia de -4 [ms] casi la mitad de las veces!

Script 6.3: cálculo teórico del poder.

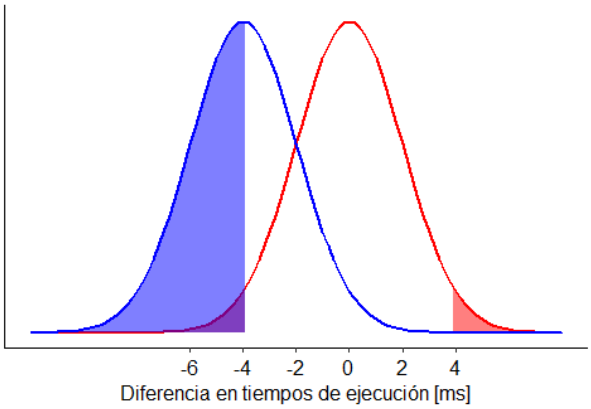


Figura 6.6: región de rechazo de la hipótesis nula en la distribución cuando el programa B es, en promedio, 4 milisegundos más rápido que el programa A.

```r
library(ggpubr)
library(pwr)

# Fijar valores conocidos.
sigma <- 12
alfa <- 0.05
n <- 36

# Calcular el error estándar.
SE <- sigma / sqrt(n)

# Gráficar la distribución muestral de la media de las diferencias si
# la hipótesis nula fuera verdadera.
x <- seq(-6 * SE, 4 * SE, 0.01)
y <- dnorm(x, mean = media_nula, sd = SE)
g <- ggplot(data = data.frame(x, y), aes(x))

g <- g + stat_function(
  fun = dnorm,
  args = list(mean = media_nula, sd = SE),
  colour = "red", size = 1)

g <- g + ylab("")
g <- g + scale_y_continuous(breaks = NULL)
g <- g + scale_x_continuous(name = "Diferencia en tiempos de ejecución [ms]",
                            breaks = seq(-6, 4, 2))

g <- g + theme_pubr()

# Colorear la región de rechazo de la hipótesis nula.
media_nula <- 0
Z_critico <- qnorm(alfa/2, mean = media_nula, sd = SE, lower.tail = FALSE)
q_critico_inferior <- media_nula - Z_critico
q_critico_superior <- media_nula + Z_critico

g <- g + geom_area(data = subset(df, x < q_critico_inferior),
                   aes(y = y),
                   colour = "red",
                   fill = "red",
                   alpha = 0.5)
```

```r

g <- g + geom_area(data = subset(df, x > q_critico_superior),
                   aes(y = y),
                   colour = "red",
                   fill = "red",
                   alpha = 0.5)

print(g)

# Superponer la distribución muestral de la media de las diferencias
# si la la diferencia de medias fuera -4.
g <- g + stat_function(
  fun = dnorm,
  args = list(mean = media_efecto, sd = SE),
  colour = "blue", size = 1)

# Colorear la región de la nueva curva situada en la región de
# rechazo de la curva original.
x1 <- seq(-6 * SE, 4 * SE, 0.01)
y1 <- dnorm(x, mean = media_efecto, sd = SE)
g <- g + geom_area(data = subset(data.frame(x1, y1),
                                 x < q_critico_inferior),
                   aes(x = x1, y = y1),
                   colour = "blue",
                   fill = "blue",
                   alpha = 0.5)

g <- g + geom_area(data = subset(data.frame(x1, y1),
                                 x > q_critico_superior),
                   aes(x = x1, y = y1),
                   colour = "blue",
                   fill = "blue",
                   alpha = 0.5)
print(g)

# Calcular el poder de acuerdo al análisis teórico.
poder <- pnorm(q_critico_inferior,
               mean = media_efecto,
               sd = SE,
               lower.tail = TRUE)
+ pnorm(q_critico_superior,
        mean = media_efecto,
        sd = SE,
        lower.tail =  FALSE)

cat("Poder = ", poder, "\n")

# Calcular la probabilidad de cometer un error tipo II.
beta <- 1 - poder_teorico
cat("Beta = ", beta, "\n")
```

Es necesario reiterar que la discusión hecha en esta sección con medias naturalmente se extienden al contraste de hipótesis con otros estadísticos.

6.5 CÁLCULO DEL PODER EN R

Desde luego, si trabajamos con R, podemos usar funciones para calcular el poder. En scripts anteriores ((6.1 y 6.2) hemos usado una primera alternativa, la función `power.t.test(n, delta, sd, sig.level, power, type, alternative)` que R trae incorporada, donde:

- `n`: tamaño de la muestra (por cada grupo, si corresponde).
- `delta`: diferencia observada entre las medias, o entre la media muestral y el valor nulo, no estandarizada.
- `sd`: desviación estándar observada.
- `sig.level`: nivel de significación.
- `power`: poder de la prueba.
- `type`: tipo de prueba t de Student (`''two.sample''` para diferencia de medias, `''one.sample''` para una sola muestra o `''paired''` para dos muestras pareadas).
- `alternative`: tipo de hipótesis alternativa (`''one.sided''` si es unilateral, `''two.sided''` si es bilateral).

Esta función entrega como resultado un objeto con diversos elementos (que podemos indexar del mismo modo que las columnas de una matriz de datos), entre los que se incluyen los 5 primeros argumentos definidos para la función.

Si revisamos con detenimiento los argumentos de la función `power.t.test()`, veremos que ¡recibe el poder como uno de sus argumentos! Esto no parece tener sentido... ¿o sí? Como ya hemos visto, existe una relación entre: poder, tamaño de la muestra, tamaño del efecto y nivel de significación. A esta combinación de elementos debemos añadir también la desviación estándar, aunque no estudiaremos las matemáticas subyacentes.

En realidad, para usar la función `power.t.test()` siempre debemos señalar el tipo de prueba t con la que estamos trabajando y si la hipótesis alternativa es de una o dos colas. Esta función nos permite calcular cualquiera de los demás argumentos (tamaño de la muestra, tamaño del efecto, desviación estándar, nivel de significación o poder estadístico) para la prueba en cuestión a partir de los 4 argumentos restantes. Así, al argumento que queremos calcular se le asigna el valor `NULL` en la llamada.

Recordemos que en el ejemplo de la sección anterior, Lola Drones desea usar una prueba t bilateral para dos muestras pareadas a fin de determinar si hay diferencia entre los tiempos de ejecución promedio de ambos algoritmos. Para ello, ha considerado $n = 36$ y $\alpha = 0,05$, sabiendo que $\sigma = 12$ [ms]. Las líneas 4 a 14 del script 6.4 muestran cómo calcular el poder para este ejemplo si se desea detectar un tamaño del efecto (δ) de 4 [ms], obteniéndose como resultado que el poder es de 0.494 (y $\beta = 1 - poder = 0,506$), ligeramente diferente al obtenido en forma teórica debido a errores de redondeo.

¿Cuántas instancias debería usar Lola para lograr un poder de 0,9, manteniendo $\alpha = 0,05$, $\sigma = 12$ [ms] y $\delta = 4$ [ms]? Las líneas 17–28 del script 6.4 muestran cómo hacer este cálculo, obteniéndose como resultado $n = 97$. Como el tamaño de la muestra siempre debe ser un entero positivo, la línea 27 aproxima el resultado al entero superior.

Otra alternativa es usar la función `pwr.t.test(n, d, sig.level, power, type, alternative)` (ver script 6.4, líneas 37–63), incluida en el paquete `pwr`, donde:

- `n`: tamaño de la muestra (por cada grupo, si corresponde).
- `d`: tamaño del efecto (d de Cohen).
- `sig.level`: nivel de significación.
- `power`: poder de la prueba.
- `type`: tipo de prueba t de Student (`''two.sample''` para diferencia de medias, `''one.sample''` para una sola muestra o `''paired''` para dos muestras pareadas).
- `alternative`: tipo de hipótesis alternativa (`''greater''` o `''less''` si es unilateral, `''two.sided''` si es bilateral).

Debemos fijarnos en que, si bien esta función opera de manera similar a `power.t.test()`, en este caso la desviación estándar y la diferencia son reemplazadas por el tamaño del efecto que podemos cuantificar, como

ya vimos, mediante la d de Cohen. Sin embargo, debemos tener cuidado, pues la función `pwr.t.test()` solo es adecuada para una muestra, dos muestras pareadas o cuando ambas muestras tienen igual tamaño. En el caso de la prueba t para dos muestras independientes con diferentes tamaños, debemos usar, en cambio, la función `pwr.t2n.test(n1, n2, d, sig.level, power, alternative)`.

Script 6.4: cálculo del poder en R.

```r
library(pwr)

# Fijar valores conocidos.
n <- 36
diferencia <- 4
desv_est <- 12
alfa <- 0.05
poder <- 0.9

# Calcular el poder usando la función power.t.test().
cat("Cálculo del poder con power.t.test()\n")

resultado <- power.t.test(n = n,
                          delta = diferencia,
                          sd = desv_est,
                          sig.level = alfa,
                          power = NULL,
                          type = "paired",
                          alternative = "two.sided")

print(resultado)

# Cálculo del tamaño de la muestra usando la función power.t.test().
cat("Cálculo del tamaño de la muestra con power.t.test()\n")

resultado <- power.t.test(n = NULL,
                          delta = diferencia,
                          sd = desv_est,
                          sig.level = alfa,
                          power = poder,
                          type = "paired",
                          alternative = "two.sided")

n <- ceiling(resultado[["n"]])
cat("n = ", n, "\n")

# Calcular el tamaño del efecto (d de Cohen).
d <- (4 / desv_est) * ((n - 2) / (n - 1.25))

# Calcular el poder usando la función pwr.t.test().
cat("\n\nCálculo del poder con pwr.t.test()\n")

resultado <- pwr.t.test(n = n,
                        d = d,
                        sig.level = alfa,
                        power = NULL,
                        type = "paired",
                        alternative = "two.sided")

print(resultado)

# Cálculo del tamaño de la muestra usando la función pwr.t.test().
cat("\nCálculo del tamaño de la muestra con pwr.t.test()\n")
```

```r

resultado <- pwr.t.test(n = NULL,
                        d = d,
                        sig.level = alfa,
                        power = poder,
                        type = "paired",
                        alternative = "two.sided")

n <- ceiling(resultado[["n"]])
cat("n = ", n, "\n")
```

Para trabajar con proporciones, R base nos ofrece la función `power.prop.test(n, p1, p2, sig.level, power, alternative)`, donde:

- `n`: número de observaciones por cada grupo.
- `p1`: probabilidad de éxito en un grupo.
- `p2`: probabilidad de éxito en otro grupo.
- `sig.level`: nivel de significación.
- `power`: poder de la prueba.
- `alternative`: tipo de hipótesis alternativa (`‘‘one.sided’’` si es unilateral, `‘‘two.sided’’` si es bilateral).

Al igual que la función `power.t.test()`, recibe cuatro de los primeros argumentos y al restante debe asigrnársele el valor `NULL`. Como resultado, retorna un objeto que incluye el valor calculado para el argumento faltante.

Como para la prueba t de Student, el paquete `pwr` también nos ofrece varias funciones que podemos usar como alternativa para las diferentes formas de la prueba de proporciones:

- `pwr.p.test(h, n, sig.level, power, alternative)`: para pruebas con una única proporción.
- `pwr.2p.test(h, n, sig.level, power, alternative)`: para pruebas con dos proporciones donde ambas muestras son de igual tamaño.
- `pwr.2p2n.test(h, n1, n2, sig.level, power, alternative)`: para pruebas con dos proporciones y muestras de diferente tamaño.

Donde:

- `h`: tamaño de efecto.
- `n`, `n1`, `n2`: tamaño(s) de la(s) muestra(s).
- `sig.level`: nivel de significación.
- `power`: poder.
- `alternative`: tipo de hipótesis alternativa (`‘‘two.sided’’`, `‘‘less’’` o `‘‘greater’’`).

El funcionamiento de esta familia de funciones es igual al que ya conocimos para la familia `pwr.t.test`. Se entrega el parámetro `alternative` y todos los demás excepto uno (al cual debe asignarse explícitamente el valor `NULL`). Como resultado, la función calcula dicho valor. El paquete `pwr` también proporciona una implementación de la h de Cohen en la función `ES.h(p1, p2)` para calcular el tamaño del efecto. En el caso de una única proporción, los autores del paquete `pwr` sugieren usar $p_2 = 0,5$ (Champely y col., 2020).

Otra función que nos puede ser de ayuda es `bsamsize(p1, p2, fraction, alpha, power)`, del paquete `Hmisc`, que implementa el método propuesto por Fleiss y col. (1980). En el caso de una prueba de Wilson con dos muestras, calcula los tamaños de cada grupo dados los siguientes argumentos:

- `p1`: probabilidad de la población para el grupo 1.
- `p2`: probabilidad del grupo 2.
- `fraction`: fracción de las observaciones en el grupo 1 ($n1/(n1 + n2)$.
- `alpha`: nivel de significación.
- `power`: poder deseado.

Es decir, asignando valores distintos a 1/2 al argumento `fraction`, podemos obtener tamaños distintos para cada una de las muestras, lo que puede ser conveniente cuando conseguir observaciones desde una de las poblaciones es más difícil o costoso que desde la otra.

Hasta ahora no hemos encontrado una función en R que realice algo similar para la prueba t de Student, aunque una forma potencialmente correcta se propuso en Jan y Shieh (2011). Notemos también que el paquete `pwr` implementa funciones para otros estadísticos y pruebas de hipótesis (Champely y col., 2020).

6.6 EJERCICIOS PROPUESTOS

1. Define con tus propias palabras lo que es el tamaño del efecto.
2. Un estudio sobre en el tiempo que necesitan los estudiantes para resolver una guía de ejercicios de Cálculo I, comparó un grupo de estudiantes que cursaban la asignatura por primera vez con un grupo que la cursaba en segunda ocasión. Sabiendo que este tiempo se distribuye normalmente en ambos casos, con varianza similar, dibuja cómo se verían los datos si el efecto de repetir la asignatura sobre el tiempo requerido para resolver la guía fuera "grande" y si este efecto fuera "pequeño, pero positivo".
3. Investiga cómo se calcula y cómo se interpreta la medida g de Hedges para el tamaño del efecto, e indica en qué casos es adecuada.
4. ¿Por qué se necesita conocer el tamaño del efecto?
5. ¿Cómo se relaciona el tamaño del efecto con la significación estadística?
6. ¿Por qué sería útil determinar un tamaño de muestra apropiado?
7. Explica en tus palabras lo que se muestra en la figura 6.4.
8. Ante algunas acusaciones de colusión, el Tribunal de la Libre Competencia quiere estudiar dos compañías del mercado de los seguros de automóviles. En base a datos del gremio de las aseguradoras, se puede asumir que el precio de las primas estándares para diferentes marcas de vehículos sigue una distribución aproximadamente normal con desviación estándar de $16.000. Fija los otros parámetros del estudio y determina qué tamaño debería tener la muestra de automóviles para detectar una diferencia de $10.000 en el precio medio de las compañías bajo sospecha.
9. Un laboratorio homeopático acaba de lanzar un tónico que asegura que ayuda a prevenir el resfrío durante el periodo invernal, con igual eficacia tanto en mujeres como en hombres. Para comprobar esta promesa, el laboratorio está realizando un estudio de la eficacia del producto en una muestra aleatoria de 100 mujeres y 200 hombres. El estudio encontró que, durante las semanas de prueba, 38 mujeres y 102 hombres presentaron síntomas de resfrío. Si Si se usa un nivel de significación de 0,05, ¿qué poder tendría esta la prueba?
10. En el estudio anterior, ¿qué tamaño deberían tener las muestras aleatorias de mujeres y hombres (manteniendo la proporción del ejemplo) para conseguir un poder de 0,85 con 99 % de confianza?

REFERENCIAS

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

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.

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

Fleiss, J. L., Tytun, A. & Ury, H. K. (1980). A simple approximation for calculating sample sizes for comparing independent proportions. *Biometrics*, 343-346.

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

Jan, S.-L. & Shieh, G. (2011). Optimal sample sizes for Welch's test under various allocation and cost considerations. *Behavior research methods*, *43*(4), 1014-1022.

Kassambara, A. (2019). *T-test Effect Size using Cohen's d Measure*. Consultado el 27 de abril de 2021, desde https://www.datanovia.com/en/lessons/t-test-effect-size-using-cohens-d-measure/#cohens-d-for-paired-samples-t-test

Warton, D. I. & Hui, F. K. (2011). The arcsine is asinine: the analysis of proportions in ecology. *Ecology*, *92*(1), 3-10.