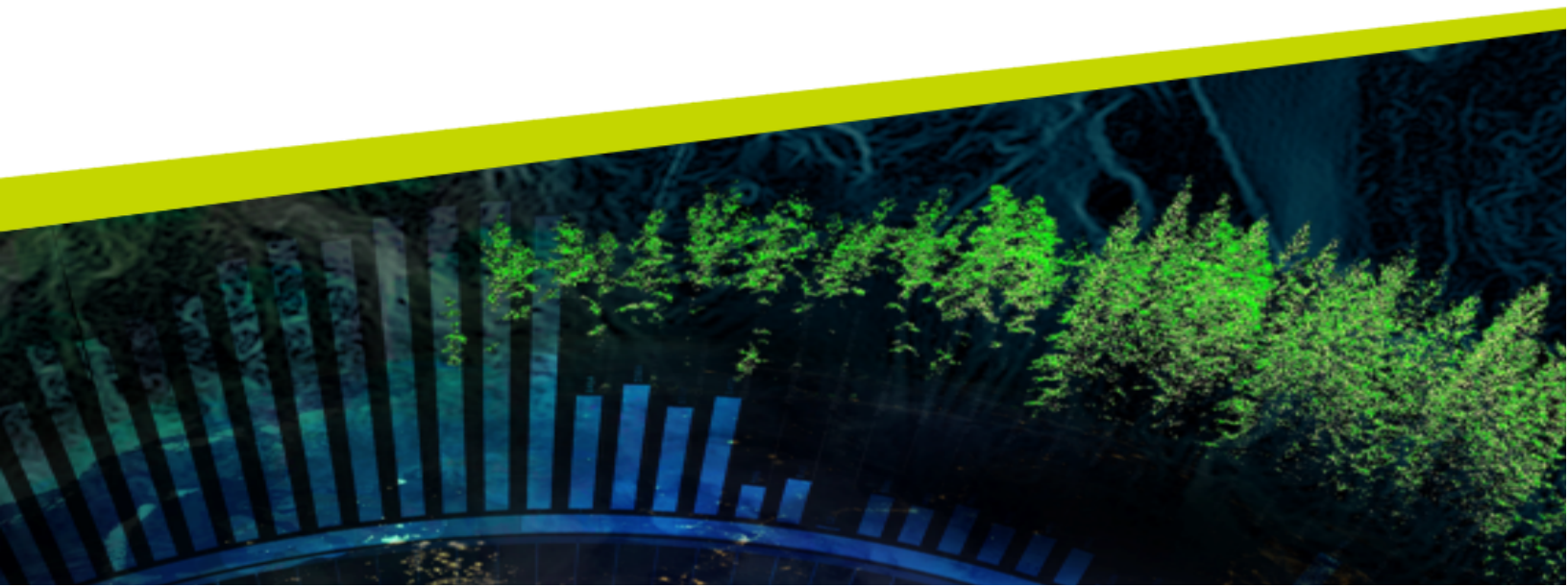
INFERENCIA Y MODELOS ESTADÍSTICOS

Jacqueline Köhler C. y José Luis Jara V.



CAPÍTULO 7. 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 $\boldsymbol{\beta}$ como **la probabilidad de cometer errores de tipo II**. α y β están relacionados: para una misma prueba de hipótesis (mismas poblaciones de origen y tamaños de las muestras) **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** o **potencia** de una prueba de hipótesis, 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 potencia de una prueba es como su **capacidad para distinguir un efecto real de una simple casualidad**.

Con esta última idea aparece otro concepto importante: el **tamaño del efecto**, que corresponde a una cuantificación de la **magnitud de la asociación o diferencia entre dos grupos o variables**. Existen diferentes medidas del tamaño del efecto dependiendo del tipo de análisis estadístico que se esté realizando. Algunas de las medidas se basan en correlación para cuantificar la fuerza de la asociación entre variables continuas, otras en distancias estandarizadas entre medias de grupos, y otras en la disparidad entre frecuencias. Como ha sido habitual, el tamaño del efecto ocurre en las poblaciones estudiadas, pero en la literatura muchas veces se refiere a su estimación a partir de muestras.

El tamaño del efecto es importante porque puede ayudar a interpretar la **relevancia práctica** de los resultados de un estudio o intervención. Una diferencia estadísticamente significativa puede no tener valor desde un punto de vista práctico si el tamaño del efecto es pequeño. Por otro lado, una diferencia no significativa en los resultados estadísticos puede ser importante si el tamaño del efecto es grande.

Por ejemplo, si nos ofrecen una "pastilla" que, si se toma todos los días, asegura (i.e. tiene un efecto estadísticamente significativo) un mejor promedio al completar una carrera de ingeniería. ¿Estaríamos dispuesto a comprarla si el aumento en el promedio asegurado es de 0,5 décimas? Probablemente no. ¿Y estaríamos dispuesto si tal diferencia fuera 8,0 décimas? Es probable que estudiantes que planean seguir estudios de postgrado o trabajar en compañías muy competitivas encontrarían que los costos del tratamiento valen la pena.

En este capítulo, y en algunos de los siguientes, iremos viendo unas pocas alternativas para medir el tamaño del efecto en las pruebas de hipótesis estudiadas. 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), Freund y Wilson (2003, pp. 123-138) y Ellis (2010).

7.1 POTENCIA DE LA PRUEBA Z

Comencemos revisando estos conceptos aplicados a la prueba Z estudiada en el capítulo 5. Como esta prueba se usa para inferir sobre la media de una población, la opción obvia para cuantificar el tamaño del efecto es la **diferencia entre medias**, muchas veces denotada como δ (delta), que tiene la ventaja que los seres humanos podemos dimensionarla fácilmente porque se encuentra en la misma escala que la variable estudiada. Para entender esta idea, pensemos en un ejemplo muy cotidiano: una adulta mayor preocupada de cuidar su presupuesto no estaría dispuesta a caminar 10 cuadras hasta la carnicería más económica del barrio si es que en promedio se va a ahorrar 100 pesos por kilo de carne, mientras que sí estaría dispuesta por un ahorro promedio de mil pesos por kilo.

Consideremos entonces un ejemplo más "técnico": en su trabajo de titulación, Lola Drones, estudiante de informática, ha diseñado un nuevo algoritmo que resuelve instancias grandes del problema de la mochila. Analizando la complejidad teórica de su algoritmo para el caso promedio y la implementación que ha hecho en lenguaje C, Lola predice que, en su máquina, los tiempos de ejecución para resolver problemas con 20.000 objetos deberían ser en promedio 1 minuto con una desviación estándar de 12 segundos. Para tener una confirmación empírica, Lola ha decidido realizar una prueba z, con un nivel de significación $\alpha = 0{,}05$, usando para ello una muestra de 36 instancias para docimar las siguientes hipótesis:

H_0: la media del tiempo de ejecución requerido por la implementación del nuevo algoritmo para resolver instancias del problema de la mochila con 20.000 objetos (μ) es de 1 minuto, es decir: $\mu = 60$ [s].

H_A: el tiempo de ejecución requerido para resolver instancias del problema de la mochila con 20.000 objetos por el programa en estudio es, en promedio, distinto de 1 minuto, es decir: $\mu \neq 60$ [s].

La figura 7.1 grafica cómo se ve la distribución muestral de los tiempos de ejecución para muestras de tamaño 36 según las hipótesis formuladas por Lola. Las áreas coloreadas corresponden a las regiones de rechazo de H_0 en favor de H_A, que se componen de valores menores al cuantil 2,5 % y los mayores al cuantil 97,5 %, por tratarse de una prueba bilateral. En este caso, considerando $\mu = 60$ y $\sigma_{\overline{x}} = 12/\sqrt{36} = 2$, estos cuantiles críticos corresponden respectivamente a $z^*_{inf} = 56{,}08$ y $z^*_{sup} = 63{,}92$.

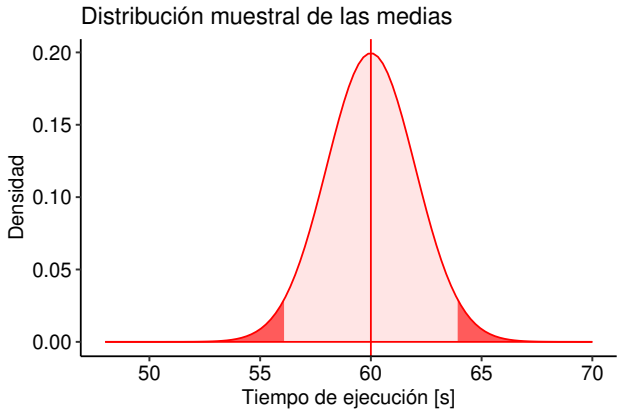


Figura 7.1: representación de las hipótesis para el tiempo de ejecución en muestras de 36 instancias con 20.000 objetos.

Supongamos que, en realidad, el programa de Lola tarda en promedio 55,8 segundos en resolver instancias con 20.000 objetos. Esta situación es representada en color azul en la figura 7.2, junto a las hipótesis de la prueba en color rojo. En este caso tendríamos que la diferencia entre el valor nulo supuesto por Lola y el verdadero valor de la media es $\delta = -4{,}2$ [s], correspondiente al tamaño del efecto.

Por supuesto, Lola no conoce las verdaderas condiciones de la prueba. Entonces, ¿qué probabilidad existe de que ella detecte que su hipótesis nula es incorrecta? Sabemos que ella piensa rechazar H_0 si la muestra exhibe un promedio que caiga en las regiones de rechazo que ha definido, las zonas coloreadas en rojo en la figura 7.2. La probabilidad de que esto ocurra corresponde a las áreas análogas definida por los cuantiles críticos, pero en la verdadera distribución: $P(\overline{x} < z^*_{inf} | \mu = 55{,}8) + P(\overline{x} > z^*_{sup} | \mu = 55{,}8) = 55{,}57\,\%$.

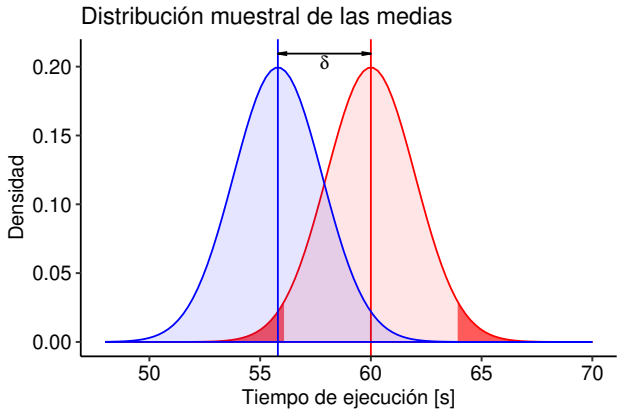


Figura 7.2: representación del tamaño del efecto δ entre la verdadera distribución muestral de las medias (azul) y las hipótesis contrastadas (rojo).

Esta es la situación mostrada en la figura 7.3, donde la probabilidad de detectar que la hipótesis nula es incorrecta está coloreada en azul. En estricto rigor hay dos áreas, una inferior y otra superior, porque Lola definió una hipótesis alternativa bilateral, pero esta última es despreciable en comparación a la primera. Debe llamarnos la atención que esta área, que corresponde al poder de la prueba de Lola, no es mucho más de la mitad de la curva. ¡Lola no sería capaz de detectar una diferencia de -4,2 [s] en casi la mitad de las muestras de tamaño 36 que utilice! Así, la probabilidad de cometer un error de tipo II corresponde al área complemento del poder, marcado con líneas azules en la figura, y que asciende a $\beta = 44{,}43\,\%$.

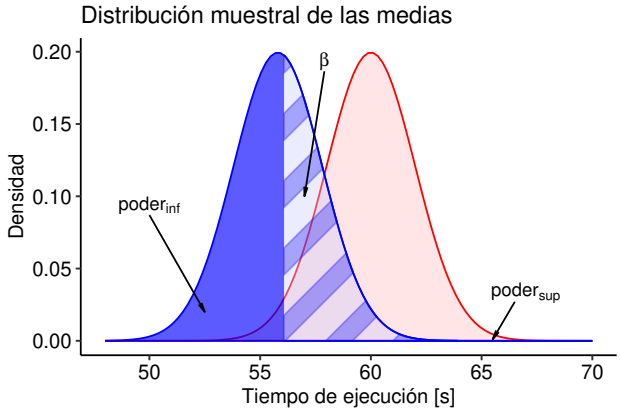


Figura 7.3: probabilidad de cometer un error de tipo II y potencia de la prueba Z del ejemplo.

Estos gráficos, así como el cálculo de las áreas de interés, se obtuvieron con el script 7.1.

Script 7.1: poder estadístico para una prueba Z bilateral.

```
library(ggpattern)
library(ggplot2)
library(ggpubr)

# Valores conocidos.
alfa <- 0.05
n <- 36

# Valores supuestos por Lola.
media_nula <- 60
sigma <- 12

```

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

# Gráficar la distribución muestral de las medias si la hipótesis
# nula fuera verdadera.

# Primero, el gráfico base
g_x_limites <- media_nula + c(-6, 5) * SE
g <- ggplot() + xlim(g_x_limites)
g <- g + labs(x = "Tiempo de ejecución [s]", y = "Densidad")
g <- g + labs(title = "Distribución muestral de las medias")
g <- g + theme_pubr()

# Agregamos la hipótesis nula
dist_0 <- stat_function(fun = dnorm,
                        args = list(mean = media_nula, sd = SE),
                        geom = "area",
                        colour = "red", fill = "red", alpha = 0.1)
g1 <- g + dist_0
g1 <- g1 + geom_vline(xintercept = media_nula, colour = "red")

# Calcular las regiones críticas de la hipótesis nula.
z_critico_inferior <- qnorm(alfa/2, mean = media_nula, sd = SE, lower.tail = TRUE)
z_critico_superior <- qnorm(alfa/2, mean = media_nula, sd = SE, lower.tail = FALSE)

# Colorear regiones de rechazo en el gráfico y el valor nulo.
g2 <- g1 + stat_function(fun = dnorm,
                         args = list(mean = media_nula, sd = SE),
                         geom = "area",
                         xlim = c(g_x_limites[1], z_critico_inferior),
                         fill = "red", alpha = 0.6)
g2 <- g2 + stat_function(fun = dnorm,
                         args = list(mean = media_nula, sd = SE),
                         geom = "area",
                         xlim = c(z_critico_superior, g_x_limites[2]),
                         fill = "red", alpha = 0.6)
print(g2)

# Valores vervaderos desconocidos por Lola.
media_verdadera <- 55.8
delta <- media_nula - media_verdadera

# Agregar la verdadera distribución muestral de las medias.
dist_v <- stat_function(fun = dnorm,
                        args = list(mean = media_verdadera, sd = SE),
                        geom = "area",
                        colour = "blue", fill = "blue", alpha = 0.1)
g3 <- g2 + dist_v + geom_vline(xintercept = media_verdadera, colour = "blue")

# Agrega anotación del tamaño del efecto
x_ann <- c(media_verdadera, media_nula)
y_ann <- c(dnorm(media_verdadera, mean = media_verdadera, sd = SE),
           dnorm(media_nula, mean = media_nula, sd = SE))
y_ann <- y_ann + 0.01
g3 <- g3 + annotate("segment", x = x_ann[1], y = y_ann[1],
                    xend = x_ann[2], yend = y_ann[2],
                    arrow = arrow(angle = 10, length = unit(0.03, "npc"),
                                  ends = "both", type = "open"))
g3 <- g3 + annotate("text", x = sum(x_ann) / 2, y = y_ann[1] - 0.001,
                    label = "delta", vjust = "top", parse = TRUE)
```

```r
print(g3)

# Traspasar las regiones críticas a la verdadera distribución muestral
# de las medias.
g4 <- g + dist_0 + dist_v
g4 <- g4 + stat_function(fun = dnorm,
                         args = list(mean = media_verdadera, sd = SE),
                         geom = "area",
                         xlim = c(g_x_limites[1], z_critico_inferior),
                         fill = "blue", alpha = 0.6)
g4 <- g4 + stat_function(fun = dnorm,
                         args = list(mean = media_verdadera, sd = SE),
                         geom = "area",
                         xlim = c(z_critico_superior, g_x_limites[2]),
                         fill = "blue", alpha = 0.6)
g4 <- g4 + stat_function(fun = dnorm,
                         args = list(mean = media_verdadera, sd = SE),
                         geom = "area_pattern",
                         xlim = c(z_critico_inferior, z_critico_superior),
                         fill = "white", colour = "blue", alpha = 0.3,
                         pattern_spacing = 0.15, pattern_density = 0.4,
                         pattern_fill = "blue", pattern_colour = "blue",
                         pattern_angle = 45, pattern_alpha = 0.3)
# Agrega anotación del poder
g4 <- g4 + annotate("text", x = 50, y = 0.1, label = "poder[inf]",
                    vjust = "top", parse = TRUE)
g4 <- g4 + annotate("text", x = 67, y = 0.04, label = "poder[sup]",
                    vjust = "top", parse = TRUE)
g4 <- g4 + annotate("text", x = sum(x_ann) / 2, y = y_ann[1] - 0.01,
                    label = "beta", vjust = "top", parse = TRUE)
g4 <- g4 + annotate("segment", x = 50, y = 0.087, xend = 52.5, yend = 0.02,
                    arrow = arrow(angle = 10, length = unit(0.03, "npc"),
                                  ends = "last", type = "open"))
g4 <- g4 + annotate("segment", x = 66.5, y = 0.027, xend = 65.5, yend = 0.001,
                    arrow = arrow(angle = 10, length = unit(0.03, "npc"),
                                  ends = "last", type = "open"))
g4 <- g4 + annotate("segment", x = sum(x_ann) / 2, y = y_ann[1] - 0.023,
                    xend = 57, yend = 0.10,
                    arrow = arrow(angle = 10, length = unit(0.03, "npc"),
                                  ends = "last", type = "open"))
print(g4)

# Calcular el poder.
poder_inf <- pnorm(z_critico_inferior, mean = media_verdadera, sd = SE,
                   lower.tail = TRUE)
poder_sup <- pnorm(z_critico_superior, mean = media_verdadera, sd = SE,
                   lower.tail =  FALSE)
poder <- poder_inf + poder_sup
cat("Poder = ", poder, "\n")

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

El procedimiento anterior es muy similar si la hipótesis alternativa es unilateral. La única diferencia es que existe una única área asociada al poder de la prueba, según si se piensa en medias mayores o menores que el valor nulo considerado. Por ejemplo, si Lola pensara que, en promedio, su programa para resolver instancias del problema de la mochila con 20.000 objetos requiere a lo más 1 minuto de ejecución, solo debemos considerar el área de la cola inferior de la distribución muestral.

En la misma línea, el procedimiento es idéntico para una prueba Z con dos muestras apareadas, considerando para el análisis la distribución muestral para las medias que resulta de las diferencias de los pares de observaciones. Esto aplicaría, por ejemplo, si Lola quisiera comparar dos implementaciones de su algoritmo con la misma muestra de instancias de prueba.

Y si se trata de dos muestras independientes, el análisis sigue siendo el mismo con el cuidado de utilizar el error estándar agrupado SE_p (*pooled standard error*) como muestra la ecuación 7.1. Un ejemplo de este caso sería si Lola quisiera comparar su programa con otro reportado en un artículo científico, que describe la distribución de los tiempos de ejecución (que es normal) pero no entrega detalle de las instancias de prueba utilizadas.

$$SE_p = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{7.1}$$

7.1.1 Cálculo del poder usando R

Como es habitual, R nos entrega facilidades para hacer este tipo de análisis. El paquete `pwr` nos pone a disposición funciones para estimar el poder y sus relaciones, que discutiremos más adelante, para varias de las pruebas estadísticas más comunes. Para la prueba Z, existe `pwr.norm.test(d, n, sig.level, power, alternative)`, donde:

- `d`: tamaño del efecto (normalizado).
- `n`: tamaño de la muestra.
- `sig.level`: nivel de significación de la prueba (probabilidad de cometer un error tipo I).
- `power`: poder de la prueba (probabilidad de cometer un error tipo II).
- `alternative`: tipo de hipótesis alternativa, como de costumbre `"two.sided"` para una prueba bilateral, y `"greater"` o `"less"` para pruebas unilaterales.

Para usar esta función entonces, debemos tener dos consideraciones. Primero, que nos pide el **tamaño del efecto normalizado**, que se consigue dividiendo la diferencia de las medias por la desviación estándar[1]: `d` $= \frac{\mu - \mu_0}{\sigma}$. Y segundo, notaremos que la función tiene un argumento `power` que no parece tener sentido puesto que es lo que queremos estimar. Este argumento existe en todas las funciones de este paquete para utilizarlas con otros fines, como veremos un poco más adelante. Para indicar que queremos que nos calcule el poder de una prueba, debemos hacer la llamada dándole un valor `NULL` u omitiéndolo pues toma este valor por defecto. La figura 7.4 muestra cómo se usa esta función para el ejemplo desarrollado en el script 7.1. En su respuesta, la función indica los valores para todos sus argumentos, incluido para `power` que antes era desconocido y que ha calculado. Vemos que es el mismo valor obtenido en el script.

```
> pwr.norm.test(d = -4.2 / 12, n = 36, sig.level = 0.05, alternative = "two.sided")

     Mean power calculation for normal distribution with known variance

              d = 0.35
              n = 36
      sig.level = 0.05
          power = 0.5557088
    alternative = two.sided
```

Figura 7.4: ejemplo del uso de la función pwr.norm.test() del paquete pwr.

[1] En la práctica se usa `d` $= \frac{\overline{x} - \mu_0}{\sigma}$ donde $\overline{x}$ es el promedio observado en la muestra usada en la prueba Z

7.1.2 Relación entre el poder y el tamaño del efecto

Del análisis que hicimos de las figuras 7.2 y 7.3, resulta más o menos evidente que a medida que δ crece, es decir el centro de las distribuciones hipotética y verdadera se alejan, más va a crecer el área de rechazo en la verdadera distribución, aumentando de esta manera la potencia de la prueba. Por el contrario, entre más pequeña sea δ, más difícil será detectar la diferencia y más alta será la probabilidad de cometer un error de tipo II (menor potencia).

La figura 7.5 (creada mediante el script 7.2), muestra la curva de poder para la prueba Z que estamos usando de ejemplo para diferentes tamaños del efecto. También presenta el caso en que se considerara una prueba unilateral.

En ella podemos observar que el poder de la prueba aumenta mientras mayor es el tamaño del efecto, y que a medida que este disminuye (es decir, el verdadero valor medio se acerca al valor nulo), el poder se aproxima al nivel de significación. Debemos notar eso sí que en esta ecuación no influye directamente la diferencia entre las medias, sino que su importancia en relación a la **variabilidad** en la variable estudiada. Un mismo valor δ será una diferencia importante con baja variabilidad mientras que puede ser magra si la variabilidad es alta. Es por esta razón que se usa el tamaño del efecto normalizado en las llamadas a la función `pwr.norm.test()` que se hacen en el script 7.2.

Por otro lado, en la figura se evidencia claramente la ventaja y la desventaja 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. Pero cuando el este lo hace en el sentido contrario al supuesto en la prueba, las posibilidades que el efecto sea detectado son ínfimas y no suben con el aumento del tamaño del efecto.

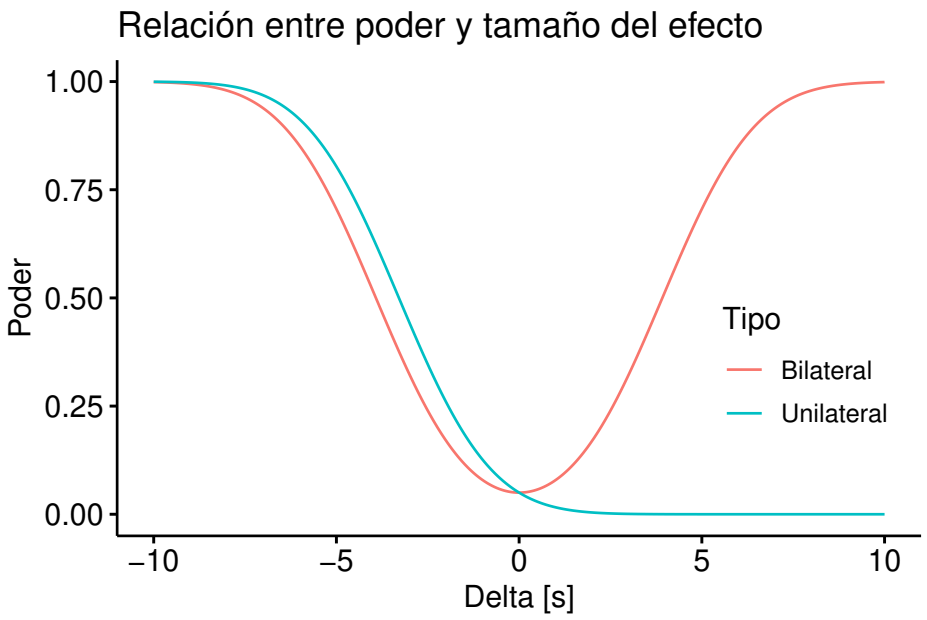


Figura 7.5: relación entre el poder estadístico y el tamaño del efecto en pruebas bilaterales y unilaterales.

Script 7.2: poder estadístico para prueba t bilateral.

```
library(ggpubr)
library(pwr)
library(tidyr)

# Valores hipótesis.
alfa <- 0.05
n <- 36
media_nula <- 60
```

```
sigma <- 12

# Tamaños del efecto.
medias_verdaderas <- seq(50, 70, 0.01)
deltas <- medias_verdaderas - media_nula
deltas_norm <- deltas / sigma

# Calcular poder de la prueba Z bilateral.
f_b <- function(x) pwr.norm.test(x, n = n, sig.level = alfa,
                                 alternative = "two.sided")[["power"]]
poder_bilat <- sapply(deltas_norm, f_b)

# Calcular poder de la prueba Z con hipótesis
# alternativa unilateral tipo "less".
f_u <- function(x) pwr.norm.test(x, n = n, sig.level = alfa,
                                 alternative = "less")[["power"]]
poder_unilat <- sapply(deltas_norm, f_u)

# Graficar estas curvas
datos_anchos <- data.frame(deltas, poder_bilat, poder_unilat)
datos_largos <- datos_anchos %>%
  pivot_longer(-deltas, names_to = "Tipo", values_to = "Poder")
datos_largos[["Tipo"]] <- factor(datos_largos[["Tipo"]],
                                 labels = c("Bilateral", "Unilateral"))
g <- ggline(datos_largos, x = "deltas", y = "Poder",
            color = "Tipo",
            numeric.x.axis = TRUE, plot_type = "l"
)
g <- g + labs(x = "Delta [s]")
g <- g + labs(title = "Relación entre poder y tamaño del efecto")
g <- ggpar(g,legend = c(.85, .35))
print(g)
```

7.1.3 Relación entre el poder y el nivel de significación

En la introducción del capítulo se mencionó que existe una relación inversa entre la probabilidad de cometer un error de tipo I (el nivel de significación α) y la probabilidad de cometer un error de tipo II (β). Si lo pensamos bien, esto es bastante intuitivo. Si nos ponemos demasiado exigentes para rechazar una hipótesis nula, se hace más probable que no rechacemos una que mecería ser rechazada. Al contrario, si bajamos la exigencia para no cometer errores de tipo II, es más probable que se nos pase una hipótesis nula falsa. Como la probabilidad β es el complemento del poder estadístico, es evidente que este último también esta ligado a la probabilidad α

La figura 7.6 (creada mediante el script 7.3), muestra las curvas de poder para la prueba Z de Lola en su versión bilateral y unilateral para diferentes niveles de significación. Puede verse que el poder crece a medida que se relaja la probabilidad de cometer errores de tipo I. Es más, se observa que el poder cae rápidamente para valores $\alpha < 0.05$ y que parece que hay que estar dispuesto a una alta probabilidad de cometer errores de tipo I para conseguir una potencia cercana o mayor a 80 %. Nuevamente se observa que el poder que consigue una prueba unilateral es mayor que el obtenido con una bilateral.

Script 7.3: poder estadístico para prueba t bilateral.

```
library(ggpubr)
library(pwr)
library(tidyr)

# Valores hipótesis.
n <- 36
```

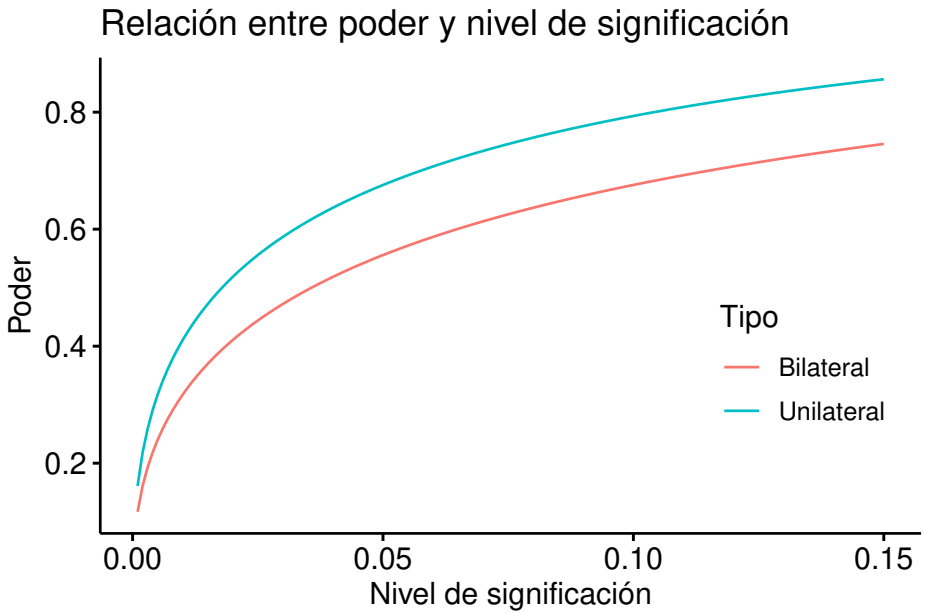


Figura 7.6: relación entre el poder estadístico y el nivel de significación en pruebas bilaterales y unilaterales.

```r
media_nula <- 60
sigma <- 12

# Tamaño del efecto.
media_verdadera <- 55.8
delta <- media_verdadera - media_nula
delta_norm <- delta / sigma

# Niveles de significación
alfas <- seq(0.001, 0.15, 0.001)

# Calcular poder de la prueba Z bilateral.
f_b <- function(x) pwr.norm.test(delta_norm, n = n, sig.level = x,
                                 alternative = "two.sided")[["power"]]
poder_bilat <- sapply(alfas, f_b)

# Calcular poder de la prueba Z con hipótesis
# alternativa unilateral tipo "less".
f_u <- function(x) pwr.norm.test(delta_norm, n = n, sig.level = x,
                                 alternative = "less")[["power"]]
poder_unilat <- sapply(alfas, f_u)

# Graficar estas curvas
datos_anchos <- data.frame(alfas, poder_bilat, poder_unilat)
datos_largos <- datos_anchos %>%
  pivot_longer(-alfas, names_to = "Tipo", values_to = "Poder")
datos_largos[["Tipo"]] <- factor(datos_largos[["Tipo"]],
                                 labels = c("Bilateral", "Unilateral"))
g <- ggline(datos_largos, x = "alfas", y = "Poder",
            color = "Tipo",
            numeric.x.axis = TRUE, plot_type = "l"
)
g <- g + labs(x = "Nivel de significación")
g <- g + labs(title = "Relación entre poder y nivel de significación")
g <- ggpar(g,legend = c(.85, .35))
```

```
print(g)
```

7.1.4 Relación entre el poder y el tamaño de la muestra

Debemos recordar que el poder estadístico está determinado por las distribuciones muestrales involucradas, hipotética y verdadera, y que la forma de estas distribuciones están influenciadas, a consecuencia del teorema del límite central, por el tamaño de las muestras utilizadas.

La figura 7.7 (creada mediante el script 7.4), muestra las curvas de poder para la prueba Z ejemplo cuando la hipótesis alternativa es bilateral y unilateral a medida que aumenta el tamaño de la muestra considerada por Lola en su experimento.

Como es de esperar, la figura muestra que el poder de la prueba va aumentando con el tamaño de la muestra. Una vez más podemos ver que la curva de poder de una prueba unilateral está por encima de la curva que describe una bilateral.

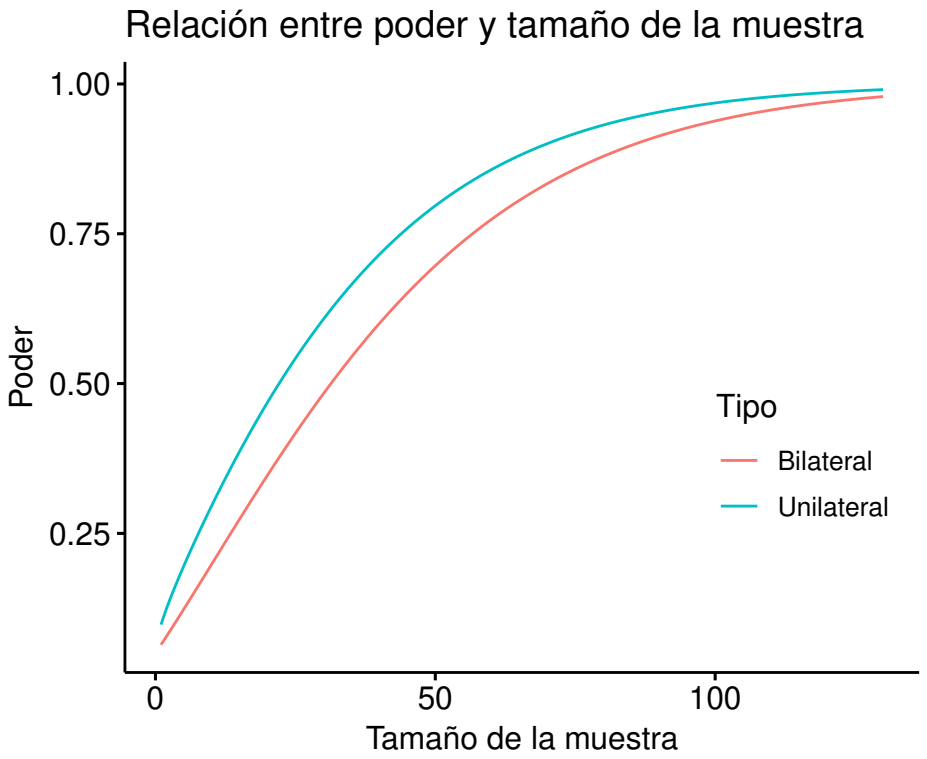


Figura 7.7: relación entre el poder estadístico y el el tamaño de la muestra en pruebas bilaterales y unilaterales.

Script 7.4: poder estadístico para prueba t bilateral.

```
library(ggpubr)
library(pwr)
library(tidyr)

# Valores hipótesis.
alfa <- 0.05
media_nula <- 60
sigma <- 12

# Tamaño del efecto.
```

```r
media_verdadera <- 55.8
delta <- media_verdadera - media_nula
delta_norm <- delta / sigma

# Tamaños de la muestra.
ns <- seq(1, 130, 0.1)

# Calcular poder de la prueba Z bilateral.
f_b <- function(x) pwr.norm.test(delta_norm, n = x, sig.level = alfa,
                                 alternative = "two.sided")[["power"]]
poder_bilat <- sapply(ns, f_b)

# Calcular poder de la prueba Z con hipótesis
# alternativa unilateral tipo "less".
f_u <- function(x) pwr.norm.test(delta_norm, n = x, sig.level = alfa,
                                 alternative = "less")[["power"]]
poder_unilat <- sapply(ns, f_u)

# Graficar estas curvas
datos_anchos <- data.frame(ns, poder_bilat, poder_unilat)
datos_largos <- datos_anchos %>%
  pivot_longer(-ns, names_to = "Tipo", values_to = "Poder")
datos_largos[["Tipo"]] <- factor(datos_largos[["Tipo"]],
                                 labels = c("Bilateral", "Unilateral"))
g <- ggline(datos_largos, x = "ns", y = "Poder",
            color = "Tipo",
            numeric.x.axis = TRUE, plot_type = "l"
)
g <- g + labs(x = "Tamaño de la muestra")
g <- g + labs(title = "Relación entre poder y tamaño de la muestra")
g <- ggpar(g,legend = c(.85, .35))
print(g)
```

7.1.5 Ajustando los factores de una prueba de hipótesis

Si bien en las secciones previas hemos revisado en forma separada las relaciones entre poder y tamaño del efecto, el poder y nivel de significación, y el poder y tamaño de la muestra, en la práctica estas interacciones no se tratan individualmente pues se **combinan en el diseño** de una prueba estadística.

Estimar el poder de una prueba realizada no es la única aplicación de este tipo de análisis. Conocer este valor podría tener importancia, por ejemplo, para un científico que está publicando un artículo con los resultados de su proyecto de investigación. Pero en otras ocasiones, un investigador busca la probabilidad α que le asegura un cierto nivel de potencia en su prueba, o cuál sería el tamaño del efecto que podría detectar en su estudio. Pero la aplicación más frecuente es determinar el **tamaño de la muestra** que se **ajusta al presupuesto** del proyecto y que permite detectar el **efecto** que se espera observar con **probabilidades razonables** de cometer errores de tipo I y II. Cada uno de estos factores puede estimarse usando las relaciones representadas en la figura 7.3, y ¡una buena cuota de álgebra!

Afortunadamente, R nos facilita las cosas enormemente ya que las funciones de poder, como las disponibles en el paquete `pwr`, permiten estimar cualquiera de los factores (tamaño mínimo de la muestra necesitada, tamaño mínimo del efecto detectable, potencia de la prueba o probabilidad de cometer errores de tipo I) simplemente dándole valor `NULL` al argumento que le corresponde. Notemos eso sí, que solamente uno de estos factores puede calcularse a la vez y es obligación **definir todos los otros**. También debemos tener presente que usualmente se usan **soluciones numéricas** (¡y no solo álgebra!), por lo que estas funciones fallan en estimar el factor desconocido si los valores otorgados a los otros factores son descabellados (y no permiten convergen a una solución).

Volvamos al ejemplo del capítulo. Supongamos que Lola quiere comparar su algoritmo con el estado del arte,

que corresponde a una implementación en `python` de una heurística llamada M^*. En el último artículo se reporta un análisis de la complejidad de esta implementación, encontrándose que para instancias de prueba con N objetos, $N > 50$, los tiempos de ejecución se distribuyeron según $\mathcal{N}(3{,}15^{\ln N} - 8{,}07\frac{N}{10},\ 81{,}9 \cdot 2^{2\ln(2N)})$ milisegundos. Recordemos que el análisis que Lola hizo de su programa optimizado para resolver instancias del problema de la mochila con 20.000 objetos, que bautizó L^*, le llevó a la conclusión que los tiempos de ejecución requeridos siguen $\mathcal{N}(60; 144)$ segundos. Al hacer los cálculos, los tiempos de ejecución que la heurística M^* requiere para instancias de este tamaño seguirían una distribución $\mathcal{N}(69.961{,}83; 196.378{,}40)$ [ms] $\approx \mathcal{N}(70; 196)$ [s]. Con estos números, Lola formula las siguientes hipótesis:

H_0: la media del tiempo de ejecución requerido por la implementación de L^* (μ_{L^*}) para resolver instancias del problema de la mochila con 20.000 objetos es similar a los observados para la implementación de M^* (μ_{M^*}), es decir: $\mu_{L^*} - \mu_{M^*} = 0$ [s].

H_A: la implementación de L^* para resolver instancias del problema de la mochila con 20.000 objetos es, en promedio, más eficiente que la implementación de M^*, es decir: $\mu_{L*} - \mu_{M^*} < 0$ [s].

El script 7.5 muestra cómo usar la función `pwr.norm.test()` para buscar el tamaño de la muestra requerido para contrastar estas nuevas hipótesis de Lola. La figura 7.8 muestra la respuesta que se obtiene de este script. Así, el tamaño (total) de las muestras requeridas en la prueba Z de dos muestras independientes (para los valores δ, α y $1 - \beta$ especificados) es $\lceil 58{,}23417 \rceil =$ `ceiling(58,23417)` $= 59$. Debemos notar que en este cálculo hemos utilizado la ecuación 7.1 para estimar el tamaño del efecto normalizado, asumiendo, como hace la función `pwr.norm.test()`, que ambas muestras tienen igual tamaño. Bajo este supuesto, Lola debería asegurar muestras independientes (i.e. que no existan instancias que aparezcan en ambas) con 30 observaciones de los tiempos de ejecución requeridos por cada programa para poder contrastar su prueba de hipótesis.

Script 7.5: cálculo del tamaño (total) de las muestras requeridas en una prueba Z para dos muestras independientes.

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

# Valores hipótesis.
alfa <- 0.05
poder <- 0.90

# Valores L*.
media_L <- 60
sigma_L <- sqrt(144)

# Valores M*.
media_M <- 70
sigma_M <- sqrt(196)

# Tamaño del efecto.
delta <- media_L - media_M
sigma <- sqrt(2 * (sigma_L^2 + sigma_M^2))
delta_norm <- delta / sigma

# Tamaño total de la muestra
factores <- pwr.norm.test(d = delta_norm, sig.level = alfa,
                          power = poder, alternative = "less")
print(factores)

cat("Número total de observaciones:", ceiling(factores[["n"]]), "\n")
```

Pero no siempre es posible o conveniente tener muestras independientes de igual tamaño. En el área de la medicina, notablemente, los grupos de control suelen ser más grandes que el grupo de pacientes que recibe el tratamiento experimental, ya que existen menos riesgos (comparado a someterse a un tratamiento no

```
     Mean power calculation for normal distribution with known variance

              d = -0.3834825
              n = 58.23417
      sig.level = 0.05
          power = 0.9
    alternative = less

Número total de observaciones: 59
```

Figura 7.8: salida producida por el script 7.5.

autorizado todavía), por lo que es más fácil conseguir voluntarios, y es más barato (el desarrollo de un tratamiento suele ser muy caro y deben considerarse posibles gastos por hacerse cargo de consecuencias negativa que el nuevo tratamiento pueda causar a los participantes del estudio).

En nuestro ejemplo, ¿qué se podría hacer si el artículo científico de referencia solamente reporta tiempos de ejecución de M^* para 25 instancias con 20.000 objetos? Con esta limitación, Lola **no podría hacer** la prueba que desea y tendría que ajustar los valores (de al menos uno) de los otros factores para hacerlos compatibles con muestras de tamaño 25.

La verdad es se han propuesto algunos procedimientos para determinar tamaños distintos para muestras independientes, pero no hay un método ampliamente aceptado como para presentarlo aquí[2].

7.2 POTENCIA DE LA PRUEBA T DE STUDENT

Los conceptos estudiados en la sección anterior no son exclusivos a la prueba Z y se extienden a **todas las pruebas de hipótesis**, y en particular a la prueba t de Student para inferir sobre medias cuando además se desconoce las varianzas.

El paquete `pwr` implementa la función `pwr.t.test(n, d, sig.level, power, type, alternative)` para ajustar los factores de una prueba t de Student, donde:

- `n`: tamaño de la(s) muestra(s) (por cada grupo).
- `d`: tamaño del efecto (d de Cohen).
- `sig.level`: nivel de significación.
- `power`: poder de la prueba.
- `type`: tipo de prueba (`"two.sample"` para diferencia de medias, `.°ne.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).

Esta función trabaja de forma análoga a la función `pwr.norm.test()` que revisamos en detalle en la sección anterior. Sin ir más lejos, tienen los mismos argumentos y con los mismos significados. La diferencia es que, internamente, se consideran distribuciones t en vez de distribuciones Z para calcular el valor del factor especificado (que tiene valor `NULL`). Sin ir más lejos, reemplazando el nombre de la función en las llamadas a `pwr.norm.test()` que aparecen en los scripts 7.3, 7.2 y 7.4, más algunos ajuste en los textos, se obtienen gráficos equivalentes de las relaciones entre los factores vistos para la prueba Z, pero ahora para la prueba t, que se muestran en la figura 7.9.

Notemos que la función `pwr.t.test()` es adecuada para una muestra, dos muestras pareadas o dos muestras independientes de 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)`.

[2] Existe un sitio web que asegura tener a disposición de los cibernautas calculadoras de poder y tamaño de muestra "gratis y fáciles de usar" (HyLown Consulting LLC, 2022), pero estos autores no han revisado la validez de esta aseveración. Estudiantes entusiastas podrían ayudar en esta tarea.

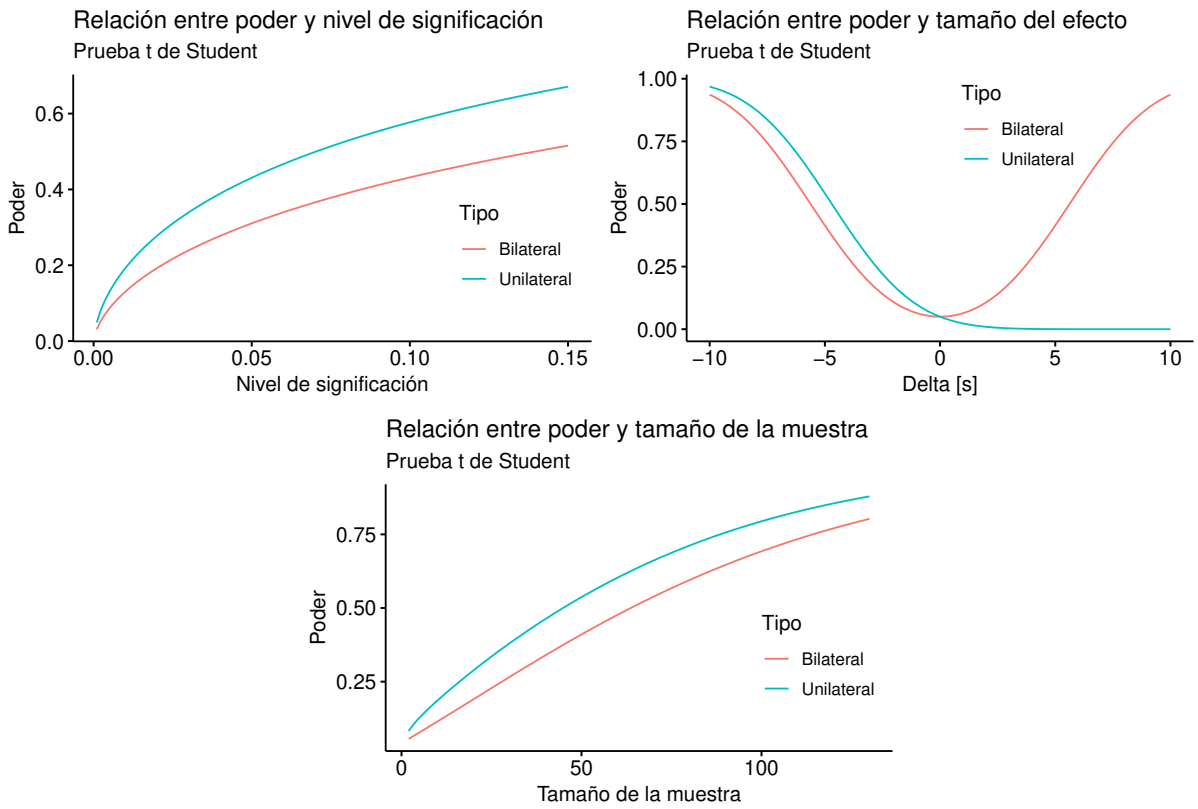


Figura 7.9: relaciones entre el poder y los otros factores en una prueba t para una muestra.

Solo falta que nos hagamos cargo del parámetro `d` de estas funciones. El lector observador habrá notado que en su definición, además de indicar que se trata del tamaño del efecto, se comenta que corresponde a la "d de Cohen", medida que revisamos a continuación.

7.2.1 Tamaño del efecto de la prueba t

Vimos que la diferencia $\delta = (\mu_1 - \mu_2)$ como medida del tamaño del efecto, en la misma escala de la variable estudiada, tiene la ventaja de que puede ser entendida por el común de la gente. El problema con esta medida es que **su escala varía** de variable en variable. Para poder hacer comparaciones con mayor libertad, como es usual en estadística, es que existen las **medidas estandarizadas del efecto**.

Para la prueba t de Student existen varias, pero aquí describiremos la llamada ***d* de Cohen** (Kassambara, 2019), que es empleada ampliamente en la comparación de medias.

En el caso de la prueba t de una muestra, la d de Cohen se calcula como muestra la ecuación 7.2, donde:

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

$$d = \frac{\overline{x} - \mu_0}{s} \tag{7.2}$$

Podemos ver que esta ecuación es similar a como normalizamos el tamaño del efecto para la prueba Z, con la diferencia que se usa la estimación muestral de la desviación estándar, el estadístico s, en reemplazo de la desconocida desviación estándar de la población, el parámetro σ.

Para la prueba t de diferencia de dos medias (también llamada prueba t para dos muestras independientes o, simplemente, prueba t independiente), si el tamaño de la muestra es mayor a 50 elementos, se calcula como muestra la ecuación 7.3, y para muestras pequeñas se aplica un factor de corrección, como indica la ecuación 7.4, 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 7.5[3].

$$d = \frac{\overline{x}_1 - \overline{x}_2}{s_p} \tag{7.3}$$

$$d = \frac{\overline{x}_1 - \overline{x}_2}{s_p} \cdot \frac{n_1 + n_2 - 3}{n_1 + n_2 - 2{,}25} \tag{7.4}$$

$$s_p = \sqrt{\frac{\sum(x - \overline{x}_1)^2 + \sum(x - \overline{x}_2)^2}{n_1 + n_2 - 2}} \tag{7.5}$$

En el caso de la variante de Welch para la prueba t independiente, la fórmula para el cálculo de la d de Cohen es ligeramente diferente, como puede apreciarse en la ecuación 7.6.

$$d = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2 + s_2^2}{2}}} \tag{7.6}$$

Por último, las ecuaciones 7.7 y 7.8 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.

$$d = \frac{\overline{x}_D}{s_D} \tag{7.7}$$

$$d = \frac{\overline{x}_D}{s_D} \cdot \frac{n_1 - 2}{n_1 - 1{,}25} \tag{7.8}$$

Existen varios paquetes en R que implementan funciones para calcular las diferentes versiones de la d de Cohen, como la función `cohens_d` del paquete `rstatix`, la función `cohensD()` del paquete `lsr` y la función `cohen.d()` del paquete `effsize`, entre otras.

Retomemos el ejemplo del análisis propuesto por Lola, quien obtuvo un registro de 25 tiempos de ejecución de M^* reportados en la literatura para instancias de prueba con 20.000 objetos. También pudo ejecutar su algoritmo L^*, en condiciones similares, con 30 instancias creadas aleatoriamente. Estos conjuntos de datos se ven en las líneas 5 y 14 del script 7.6. Con estos datos, Lola condujo una prueba t de Student con 95 % de confianza, pero le interesaría saber cuál es la probabilidad de que esté cometiendo un error tipo II al verificar la hipótesis nula que $\mu_{L^*} - \mu_{M^*} = -10$ versus la alternativa $\mu_{L^*} - \mu_{M^*} < -10$. Para obtener esta probabilidad (líneas 21-36 del script) hizo uso de la función `pwr.t2n.test()`, pues tiene dos muestras independientes de distinto tamaño. La salida del script puede verse en la figura 7.10, resultando que en la prueba de Lola existe una probabilidad cercana al 30 % de cometer un error tipo II.

Script 7.6: cálculo del poder en una prueba t unilateral para dos muestras independientes con diferentes tamaños.

```
1 library(effsize)
2 library(pwr)
3
```

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

```r
# Valores L*.
muestra_L <- c(50916.01, 68274.39, 60212.33, 57973.14, 74787.28,
               61396.89, 72907.14, 55807.43, 61142.34, 61986.08,
               69704.93, 73718.12, 70488.12, 61836.25, 71255.53,
               61133.57, 57702.44, 79472.14, 69546.98, 56296.91,
               79657.66, 52530.76, 64012.86, 75995.01, 53014.13,
               69883.13, 62638.55, 87312.34, 47351.77, 66807.14)
n_L <- length(muestra_L)

# Valores M*.
muestra_M <- c(95075.86, 64758.71, 80269.73, 74365.69, 86104.68,
               41772.91, 116915.74, 33103.66, 61553.61, 55498.1,
               73996.43, 101619.51, 61037.45, 53973.06, 65523.67,
               69378.84, 80254.29,  84242.37, 91978.80, 73853.76,
               98258.72, 61785.34, 59753.93, 66855.87, 101783.46)
n_M <- length(muestra_M)

# Obtener tamaño del efecto.
tdf <- cohen.d(muestra_L, muestra_M)
cat("Tamaño del efecto:\n")
print(tdf)

# Obtener poder de la prueba realizada.
d <- tdf[["estimate"]]
alfa <- 0.05
valor_nulo <- 10
factores <- pwr.t2n.test(n1 = n_L, n2 = n_M, d = d, sig.level = alfa,
                         alternative = "less")
cat("Factores::\n")
print(factores)

# Mostar beta
cat("Beta:", 1 - factores[["power"]], "\n")
```

```
Tamaño del efecto:

Cohen's d

d estimate: -0.5949732 (medium)
95 percent confidence interval:
      lower       upper
-1.14992166 -0.04002474

Factores::

     t test power calculation

             n1 = 30
             n2 = 25
              d = -0.5949732
      sig.level = 0.05
          power = 0.6998763
    alternative = less

Beta: 0.3001237
```

Figura 7.10: salida producida por el script 7.6.

Supongamos que Lola encuentra inaceptable esta probabilidad β, que ella desea que esté en el orden del 10 %. Una forma de incrementar la potencia de su prueba es aumentar el tamaño de las muestras, o de una de las muestras en este caso. Pero eso significa que no puede calcular el tamaño del efecto usando las funciones que R proporciona para la d de Cohen, puesto que estas necesitan **ambas muestras**.

En estos casos, cuando se quiere determinar el tamaño de la muestra que se necesita, hay dos opciones: **suponer** el tamaño del efecto que se va a encontrar en las muestras, o **definir** el tamaño del efecto mínimo que se quiere detectar. En ambos casos, es clave usar **información previa** que se tenga, ya sea revisando el estado del arte en la literatura o a través de estudios pilotos. Por supuesto, algún papel juega la experiencia y intuición de la investigadora o del investigador en lograr una buena estimación.

Por supuesto, esta estimación no necesita ser exacta. De hecho existe una recomendación general de valores d de Cohen para estimar el tamaño de muestras cuando no hay mucha información previa[4]: un valor $d = 0{,}2$ está asociado a un efecto pequeño (imperceptible a simple vista), $d = 0{,}5$ a un efecto mediano (probablemente perceptible a simple vista) y $d = 0{,}8$ a un efecto grande (definitivamente perceptible a simple vista) (Cohen, 1992). Evidentemente es más fácil (¿o no?) decidir si uno piensa encontrar un efecto pequeño, mediano o grande en su experimento. Obviamente, usando estos valores de referencia, uno puede definir valores para efectos "entre pequeño y mediano" (e.g. $d = 0{,}4$) o "no tan grande" (e.g. $d = 0{,}7$).

Volvamos al caso ejemplo. Lola tiene información previa: $L^* \sim \mathcal{N}(60; 144)$ [s] y $M^* \sim \mathcal{N}(70; 196)$ [s]. Así, muestras de estas poblaciones deberían mostrar promedios cercanos a 60 y 70 [s], respectivamente, así como desviaciones estándares en el orden de los 12 y 14 [s]. Usando la ecuación 7.6, Lola podría determinar que, en teoría, espera observar un tamaño del efecto $d \approx -0{,}767$, y usar este valor para determinar el tamaño de la muestra que le falta.

Pero Lola ya hizo una prueba con una muestra de 30 instancias (figura 7.10), en una especie de estudio piloto, donde observó un tamaño del efecto empírico $d = -0.595$, bastante menor que el teórico, y una potencia de aproximadamente 70 %. Por prudencia, y queriendo asegurar la potencia de su prueba, Lola decide estimar el tamaño de la muestra de instancias que necesita considerando $d = 0{,}6$ y $\beta = 0{,}1$. La figura 7.11 presenta este cálculo, resultando que ella necesita ¡518 instancias! en su muestra. Probablemente Lola tendrá que reconsiderar los niveles de significación y, especialmente, de potencia estadística para su prueba.

```
> pwr.t2n.test(n2 = 25, d = -0.6, sig.level = 0.05,
+              power = 0.90, alternative = "less")

     t test power calculation

             n1 = 517.6086
             n2 = 25
              d = -0.6
      sig.level = 0.05
          power = 0.9
    alternative = less
```

Figura 7.11: determinación del tamaño de una de las muestras para una prueba t de diferencia de medias unilateral.

7.3 POTENCIA DE LA PRUEBA DE LA DIFERENCIA DE DOS PROPORCIONES

Consideramos ahora la potencia estadística de la prueba de la diferencia de proporciones estudiada en el capítulo 6. Una vez más, el paquete `pwr` de R nos ofrece varias funciones que podemos usar al trabajar con esta prueba:

- `pwr.p.test(h, n, sig.level, power, alternative)`: para pruebas con una única proporción.

[4] Reglas similares existen para la mayoría de las medidas normalizadas del tamaño del efecto.

- `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.

Los argumentos de estas funciones siguen las convenciones de las que ya hemos visto de este paquete:

- `h`: tamaño de efecto (h de Cohen).
- `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 en las secciones anteriores: se especifica el parámetro `alternative` (o se deja el valor por omisión `"two.sided"`) y todos los otros factores excepto uno, el que debe tener valor `NULL` para indicar que es desconocido. Como resultado, la función devuelve un objeto con los factores de la prueba incluyendo una estimación del que se desconoce. Observemos que para la función `pwr.2p2n.test()` esto significa que debemos conocer el tamaño de al menos una de las muestras para poder estimar el tamaño que necesita la otra.

Notemos que el tamaño del efecto en estas funciones se mide con la medida h de Cohen, que puede calcularse como muestra la ecuación 7.9, implementada en R en la función `ES.h(p1, p2)` del paquete `pwr`. En el caso de una única proporción, los autores del paquete sugieren usar $p_2 = 0{,}5$ (Champely y col., 2020).

$$h = 2\arcsin(\sqrt{\hat{p}_1}) - 2\arcsin(\sqrt{\hat{p}_2}) \tag{7.9}$$

Las razones para transformar las proporciones antes de compararlas tiene relación con que sus valores, al estar limitados entre 0 y 1 de acuerdo a una distribución binomial de n experimentos de Bernoulli independientes con probabilidad de éxito p, pueden estar sesgados (y no se distribuyen normalmente) porque la varianza ($np(1-p)$) depende fuertemente de la media (np), lo que puede introducir distorsiones, especialmente cuando p se acerca a los valores extremos.

En la h de Cohen se utiliza la función **arcoseno** para estabilizar la varianza de las proporciones, que se vuelve (aproximadamente) constante después de la transformación, obteniendo datos transformados que siguen una distribución más parecida a una distribución normal[5] (Warton & Hui, 2011).

Veamos un ejemplo. Supongamos que Lola está desconcertada por la diferencia entre el tamaño del efecto teórico y el empírico en su estudio piloto de los tiempos de ejecución de L^* y M^* (visto en la sección anterior). Revisando los algoritmos que producen aleatoriamente las instancias de prueba usadas en las muestras, sospecha que la desviación podría deberse a que hay una mayor posibilidad de incluir objetos de gran tamaño en las instancias para M^* que en las instancias de prueba para L^*. Esto es relevante pues para los algoritmos es "más fácil" descartar combinaciones con muchos objetos grandes, reduciendo así el espacio de búsqueda. Ella estima que la probabilidad de generar instancias con al menos 3.000 objetos grandes (con más del 15 % del tamaño de la mochila) es aproximadamente $p_L = 0.13$ para L^* y $p_M = 0.18$ M^*. Luego, formula las siguientes hipótesis:

H_0: la probabilidad de generar instancias con al menos 3.000 objetos grandes para probar M^* (p_{M^*}) es similar a la probabilidad de generarlas para L^* (p_{L^*}), es decir: $p_{M^*} = p_{L^*}$.
H_A: la probabilidad de generar instancias con al menos 3.000 objetos grandes para probar M^* es mayor que la probabilidad de generarlas para L^*, es decir: $p_{M^*} > p_{L^*}$.

Lola piensa hacer una prueba piloto basada en la hecha para las medias, es decir con muestras de 25 y 30 observaciones para L^* y M^*, respectivamente,y necesita saber conocer la potencia que tendría una prueba para docimar sus hipótesis. La figura 7.12 muestra cómo pudo obtener este valor usando la función `pwr.2p2n.test()`.

[5] La transformación basada en la función arcoseno ha sido criticada en algunos contextos. Veremos otra transformación estabilizadora de la varianza para datos binomiales en capítulos posteriores.

```
> pwr.2p2n.test(h = ES.h(0.18, 0.13), n1 = 25, n2 = 30,
+               alternative = "greater", sig.level = 0.05)

     difference of proportion power calculation for binomial distribution (arcsine transformation)

              h = 0.1385721
             n1 = 25
             n2 = 30
      sig.level = 0.05
          power = 0.1285773
    alternative = greater

NOTE: different sample sizes
```

Figura 7.12: determinación del tamaño de una de las muestras para una prueba t de diferencia de medias unilateral.

Podemos ver que el poder de la prueba es muy bajo: solo tiene alrededor de 13 % de probabilidades de detectar la diferencia hipotetizada.

Por supuesto Lola necesita una mayor potencia para contrastar sus hipótesis. Como tiene los algoritmos para generar instancias de prueba aleatorias para cada caso, no está limitada a un tamaño fijo y tiene la libertad de usar muestras tan grandes como necesite. Para hacer este análisis, considera comparar muestras de igual tamaño por lo que utiliza el siguiente código para obtener la figura 7.13:

```
> factores <- pwr.2p.test(h = ES.h(0.18, 0.13), alternative = "greater",
+                         power = 0.80, sig.level = 0.05)
> plot(factores) + theme_pubr()
```

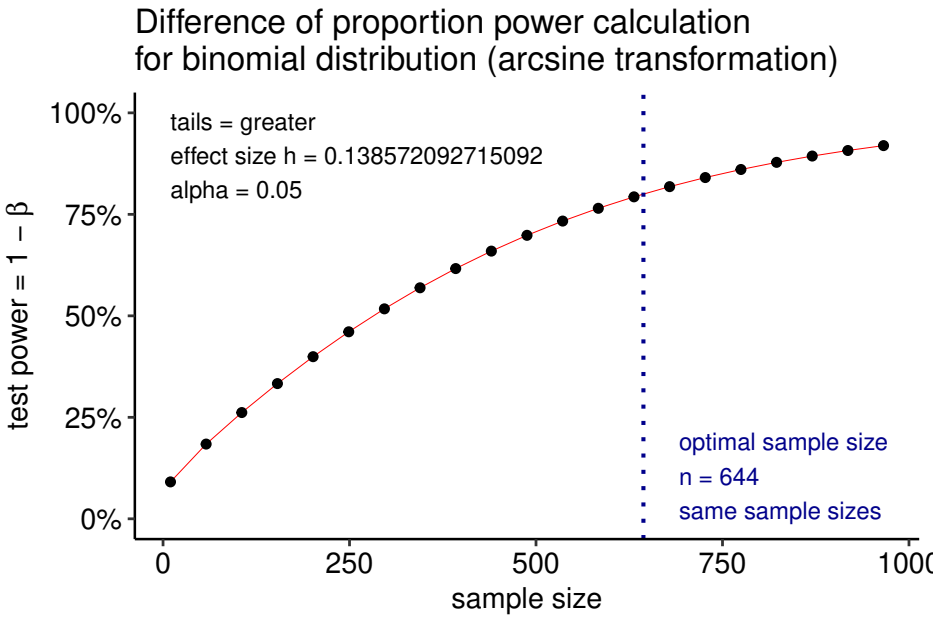


Figura 7.13: gráfico del tamaño de las muestras para una prueba de la diferencia de dos proporciones unilateral.

La figura muestra el resultado que se obtiene al "plotear" el objeto devuelto por las funciones del paquete `pwr`. En ella, se ve cómo va subiendo el poder a medida que aumenta el tamaño de la muestra, marcando el valor del tamaño solicitado que asegura los niveles de significación y poder solicitados. Luego, Lola tendrá que generar muestras con al menos 644 instancias en cada una para validar sus hipótesis.

7.4 CONSIDERACIONES FINALES

Por lo visto en este capítulo, la estimación de los factores que determina la calidad de una prueba de hipótesis no es una tarea simple, y sigue siendo un área de investigación. Por ejemplo, las funciones que manejan muestras independientes de tamaños distintos en el paquete `pwr` tienen varias limitaciones. Funciones que entregan más libertades pueden encontrarse en otros paquetes de R, aunque están basadas en procedimientos más bien nuevos y pueden tener algún grado de inestabilidad todavía.

Por ejemplo, el paquete `PASSED` (Li y col., 2021) implementa funciones para buscar los factores adecuados para realizar inferencia con muestras independientes de diferente tamaño. En el caso de la prueba t, además, considera la posibilidad de tener muestras con varianza distinta (generalizando la corrección de Welch) como se describe en Jan y Shieh (2011) (para el caso bilateral) y Jan y Shieh (2017) (para los casos unilaterales). Las mismas funciones están disponibles en el `MESS`, con la ventaja que estas usan argumentos muy parecidos a los que hemos conocido en el paquete `pwr`.

Esta duplicidad, o más bien multiplicidad, de implementación de funciones y procedimientos con el mismo objetivo parece ser una característica constante (y molesta) de R. Y no siempre es fácil determinar si funciones cuya descripción parece indicar que son lo mismo en realidad implementan exactamente el mismo procedimiento, pues, obviamente, estos están llenos de detalles. Por ejemplo, es fácil encontrarse con las funciones nativas `power.t.test()` y `power.prop.test()` que también sirven para buscar factores adecuados para estas pruebas, pero que reciben el tamaño del efecto sin normalizar. Entonces, ¿internamente normalizan el tamaño del efecto?. Hay que navegar por los manuales o, muchas veces, revisar el código para poder estar seguros.

Otro ejemplo son las funciones `power_t_test()` y `power_prop_test()` del paquete `MESS` que agregan un nuevo argumento llamado `ratio` cuyo valor corresponde a la razón n_2/n_1, que son los tamaños de la muestra más grande y más pequeña, respectivamente, y que permite indicar a las funciones que busquen un par óptimo de valores para tamaños distintos de las muestras en función de esta relación. Esto permite enfrentar situaciones en que no es fácil conseguir muestras con el mismo tamaño. Pero, si uno quiere muestras de igual tamaño (`ratio = 1`), ¿realizan el mismo procedimiento que las funciones del paquete `pwr`? Las funciones del paquete `PASSED` no tienen esta opción, mientras que en otros paquetes existen alternativas que usan variaciones (e.g. `fraction` $= n_1/(n_1 + n_2)$ en el paquete `Hmisc`).

Un último punto importante es que, como estas funciones usan métodos numéricos, la búsqueda de factores adecuados puede fallar si se solicitan cosas imposibles. Por ejemplo, si se trabaja con muestras pequeñas y un nivel de significación muy alto, al mismo tiempo que se quiere detectar un tamaño del efecto muy pequeño, va a ser muy difícil que se encuentre un valor β que satisfaga estas condiciones. En estos casos, las funciones suelen dar un error críptico, que menciona la función `uniroot()`, que intenta decirnos que no pudo encontrar una solución a nuestra solicitud.

Esto suele con especial frecuencia cuando utilizamos pruebas unilaterales. Cuando analizamos la figura 7.5, observamos que en el sentido contrario a la hipótesis alternativa el poder es prácticamente nulo. Luego, si nos equivocamos y la forma en que calculamos el tamaño del efecto no es **consistente** con la dirección de la hipótesis alternativa, es probable que tengamos problemas para encontrar una solución, ya que nunca se podría conseguir una potencia razonable. Si nos fijamos en la figura 7.12, tuvimos el cuidado de especificar `ES.h(0.18, 0.13)` que es consistente con la hipótesis alternativa especificada `alternative = "greater"` (H_A : $0.18 > 0.13$).

7.5 EJERCICIOS PROPUESTOS

1. 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".

Considera el siguiente escenario:

> Una familia emprendedora está considerando concursar por la concesión de una de las cafeterías del campus. Las bases aseguran que el 95 % de las transacciones de venta están en el rango $1.000-$10.000. La familia estima que el negocio es rentable si el valor medio de las transacciones de venta es de $3.000 o mayor, y que el verdadero valor medio debe estar alrededor de los $3.500. Para contrastar las hipótesis de que el valor medio es $3.000 versus que es mayor a $3.000 usando una prueba t con un nivel de significación de 0,05, han "espiado" de manera aleatoria 100 transacciones de venta a lo largo de una semana.

y responde las siguientes preguntas:

2. ¿Qué potencia tiene la prueba planeada?
3. ¿Cuántas transacciones deberían registrar para asegurar un poder de 0.9?

Considera el siguiente escenario:

> La nueva concesión de una de las cafeterías del campus recibió la recomendación de que oriente su decoración para atraer alumnas, porque estas gastan más que los alumnos. Antes de tomar esta decisión, planean observar 30 alumnos y 30 alumnas, elegidos al azar, al minuto de pagar en caja y calcular el monto promedio gastado por cada grupo. Luego van a aplicar una prueba t para la docimar la igualdad o diferencia de estas medias con un nivel de significación de 0,05.

y responde las siguientes preguntas:

4. ¿Qué potencia tiene la prueba si se quiere detectar un efecto mediano?
5. ¿Qué tamaño deben tener dos muestras con el mismo número de observaciones si se apunta a tener 80 % de potencia?
6. ¿Cuánto varía este tamaño si la hipótesis alternativa se cambia a que el gasto medio de las alumnas es mayor al de los alumnos?

> Se encontró que las nuevas y los nuevos estudiantes de Educación Física no están llegando con la preparación requerida. Se ha iniciado un programa piloto con 24 estudiantes que están comenzando esta carrera en que se someten a prolongados ejercicios de saltar la cuerda. En teoría, este programa debería mejorar su resistencia y bajar sus tiempos en completar 1.500 metros planos, que exhiben una distribución aproximadamente normal con desviación estándar de 14 [s]. El estudio tiene planificado emplear una prueba t con 99 % confianza para determinar si hay una diferencia significativa en los tiempos registrados antes y después del programa.

y responde las siguientes preguntas:

7. ¿Qué potencia tiene la prueba si se quiere detectar una reducción de 8 [s]?
8. ¿Qué tamaño debe tener el conjunto de estudiantes adscritos al programa si se desea obtener 80 % de potencia para este resultado del programa?
9. ¿Cuánto varía este tamaño si la hipótesis alternativa se cambia a que los tiempos se redujeron en más de 8 [s]?
10. 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.
11. 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.
12. Reconstruye el gráfico de la figura 7.9, pero ahora para la prueba de la diferencia de dos proporciones.

Considera el siguiente escenario:

> Luego de un ensayo clínico exitoso, se ha decidido aumentar la proporción de pacientes que se van con una receta de aspirina cuando son dados de alta tras un infarto de miocardio, que históricamente llega el 80 % de las veces. Un proyecto de mejora de la calidad de enfermería en los departamentos de cardiología de los hospitales de la zona occidente ha intentado aumentar

> esta tasa al 95 %. El servicio de salud quiere poder hacer su estudio con 99 % de confianza y solo 10 % de probabilidades de cometer un error de tipo II.

y responde las siguientes preguntas:

13. ¿A cuántos pacientes hay que hacer un seguimiento después de la intervención para determinar si se consiguió llegar a la proporción objetivo?
14. ¿Cómo varía este número si la hipótesis alternativa cambia a que la proporción aumentó?

Considera el siguiente escenario:

> CENDA quiere tomar una muestra aleatoria de estudiantes de la Facultad de ambos sexos y preguntarles si consumen alcohol al menos una vez a la semana. Su hipótesis nula es que no hay diferencia en la proporción que responde afirmativamente y la hipótesis alternativa es que sí hay diferencia. A CENDA le gustaría detectar una diferencia pequeña con $alpha = 0{,}05$ y $\beta = 0.20$.

y responde las siguientes preguntas:

15. Si se piensa que la proporción de uno de los grupo es de 55 % y la del otro 50 %: ¿cuántos estudiantes de cada sexo se deben encuestar si se quiere tener la misma cantidad de cada uno?
16. ¿Cómo varía este número si la hipótesis alternativa es unilateral?

Considera el siguiente escenario:

> El servicio de salud local quiere reducir la mortalidad de pacientes en diálisis renal, que históricamente ronda por un 20 %. Asignará aleatoriamente a las y los pacientes a un cóctel de betabloqueadores (tratamiento) frente a que sigan la atención habitual (control), y espera reducir la mortalidad anual al 10 %. El servicio quiere detectar esta diferencia con una potencia del 80 % para un nivel de significación bilateral de 0,05.

y responde las siguientes preguntas:

17. ¿Qué número de pacientes debe tener cada grupo si son de igual tamaño?
18. ¿Cómo varía este número si la hipótesis alternativa se cambia a que es menor que el 20 % histórico?
19. Suponiendo que el presupuesto asignado al programa alcanza solamente para financiar 120 tratamientos ¿qué tamaño debe tener el grupo de control si volvemos a considerar la hipótesis alternativa bilateral?

7.6 BIBLIOGRAFÍA DEL CAPÍTULO

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. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155-159.

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

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge university press.

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

HyLown Consulting LLC. (2022). *Power and Sample Size .com*. Consultado el 10 de marzo de 2024, desde http://powerandsamplesize.com

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

Jan, S.-L., & Shieh, G. (2017). Optimal sample size determinations for the heteroscedastic two one-sided tests of mean equivalence: Design schemes and software implementations. *Journal of Educational and Behavioral Statistics*, *42*(2), 145-165.

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