Tema 4: intervalos de confianza y tests de hipótesis

1 ICs y test para una población normal

1. Una variable aleatoria tiene media μ desconocida y desviación típica poblacional 12.5. La media muestral para una muestra aleatoria de tamaño 50 es $\bar{X} = 72.4$. Halla el intervalo de confianza del 95% para μ.

   Solution: Asumiendo normalidad y muestreo con reemplazamiento, se obtiene $(68.94, 75.87)$.

2. La desviación estándar de las alturas de 16 estudiantes de una escuela de 10000 estudiantes es 2.4 cm. Encuentra el intervalo de confianza al 95% para la varianza de los estudiantes de la escuela.

   Solution: Dado que la muestra es muy pequeña comparada con la población, podemos asumir "población infinita" y, por tanto independencia entre muestras. Como no hay datos al respecto, asumimos normalidad en base a que es un parámetro biológico (la altura es, de hecho, un ejemplo clásico de variable normal). Se obtiene $(3.14, 13.80)$ cm^2 (usando una Chi-cuadrado).

3. Una muestra aleatoria de cinco gasolineras madrileñas da los siguientes precios en euros por litro de gasolina: 124.9, 127.9, 130.9, 128.9, 122.9. Halla el intervalo de confianza del 90% para la media de precios por litro. Haz todas las asunciones que consideres oportunas.

   Solution: Asumimos que el precio por litro se aproxima a una distribución normal. Por otra parte, la muestra es pequeña comparada con la población, por lo que es razonable asumir independencia. Se obtiene, usando una distribución t-Student: $\mu \in (124.06, 130.145)$ euros.

4. Una muestra aleatoria de 25 puntuaciones medias de una universidad tiene como media muestral $\bar{X} = 2.68$ y una varianza muestral $\hat{S}^2 = 0.32^2$. Asumiendo que las medias de puntuaciones están distribuidas aproximadamente como una normal, hallar el intervalo de confianza del 95% para la media de las puntuaciones.

   Solution: Asumiendo que la muestra aleatoria proviene de una universidad suficientemente grande (de forma que las 25 muestras sean pocas comparadas con la población de la universidad), se puede suponer que las muestras son independientes. Por otra parte, la normalidad está asegurada por el enunciado. Usando la t-Student se obtiene $(2.53, 2.83)$

5. Los datos en *data_test/normal.csv* contienen medidas independientes procedentes de una población normal. (1) Testea H0: $\mu > 0$. (2) Testea ahora $\mu < 0$. (3) ¿No es esto una contradicción?

Solution: Usando un T-test, se obtienen los p-valores (1) 0.4446 y (2) 0.5554 por lo que en (1) nos "quedaríamos" con H0 $\mu > 0$ y en (2) $\mu <= 0$. Esto no es una contradicción e ilustra algo muy importante: los datos solo aportan evidencia a favor de la hipótesis alternativa. "Quedarse con H0" no significa necesariamente que el investigador acepte la H0 como verdadera, sino más bien que actualmente no hay evidencia suficiente como para concluir que es falsa. En este sentido, es mejor decir "no hay suficiente evidencia apoyando H1" en lugar de "aceptamos la H0".

6. Una máquina solía producir una pieza mecánica con un grosor de 0.05 cms. Sin embargo, hay sospechas de que la máquina ha dejado de funcionar adecuadamente. Para determinar si es cierto, se toma una muestra al azar de 10 piezas, obteniéndose una media de 0.053 cm y una varianza de 10^{-5} cm^2. Testea si la máquina ha dejado de funcionar correctamente al (a) 95% y (b) 99% de confianza. (c) Encuentra el valor del p-valor y el tamaño del efecto. Datos en "pieces.csv".

Solution: En el ámbito industrial, es razonable suponer que una máquina produce muchísimas piezas a lo largo de su vida útil, por lo que las 10 muestras son pocas en comparación con la potencial población. Esto permite asumir independencia. Para verificar la normalidad, es necesario dibujar un histograma de los datos. Dado que tiene forma razonable, podemos asumir normalidad. Dado que queremos comprobar si la máquina ha dejado de funcionar adecuadamente, $H_1 : \mu \neq 0.05$. El estadístico es $t = 3.00$. Como el test es de dos colas y la distribución es una T-Student con 9 grados de libertad, p-valor≈ 0.015 (con Cohens'D $= 0.95$, lo que sugiere que el tamaño del efecto es grande). Por tanto, (a) hay suficiente evidencia a favor de H_1 y (b) no hay suficiente evidencia que apoye H_1. Como (a) y (b) están parecen estar en contradicción sería recomendable realizar un estudio más exhaustivo (con más muestras) o bien revisar directamente la máquina (si es que hacer un estudio estadístico exhaustivo es más caro que revisar la máquina).

```
pieces <- read_csv("data_exercises/pieces.csv", col_names = FALSE)
colnames(pieces) <- "size_cms"

nbins <- nclass.FD(pieces$size_cms)
ggplot(pieces, aes(x=size_cms)) + geom_histogram(bins = nbins)

my_test <- t.test(pieces$size_cms, mu = 0.05)
print(my_test)
print(
  effectsize(my_test)
)
```

7. Respecto al ejercicio anterior, ¿Cuál es el tamaño de la muestra que se necesitaría para detectar una variación en la media de 0.003 cm, con una significación del 1% y una potencia del 90%.

Solution: Usando la desviación típica muestral $\hat{s} = \sqrt{10^{-5}}$ como estimación de σ:

```
power.t.test(sd=sqrt(1e-5), power=0.9, alternative="two.sided",
             delta=0.003, sig.level=0.01, type="one.sample")
```

Se obtiene que $n \approx 20$.

8. La vida media de una muestra de 100 lámparas fluorescentes producidas por cierta compañía es de 1570 horas. Si μ es la vida media de las lámparas producidas por la compañía, testea si hay suficiente

evidencia como para concluir que $\mu \neq 1600$ horas con un nivel de significación de (a) 0.05 y (b) 0.01. (c) Calcula el p-valor del test. Puedes asumir que la desviación estándar poblacional es de 120 horas.

Solution: Asumimos normalidad y que el número de muestras es pequeño comparada con la producción de lámparas de la compañía (que sería la población objetivo). Por tanto, podemos considerar las muestras como independientes y, dado que la desviación poblacional es conocida, emplear como estadístico $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$. El test es de dos colas, siendo el valor del estadístico -2.50. (a) hay evidencia suficiente para aceptar H_1. (b) No hay evidencia suficiente para aceptar H_1. (c) p-valor= 0.0124. De nuevo, en una situación realista, lo más lógico sería repetir el experimento con más muestras para salir de dudas.

9. Respecto al ejercicio anterior, ¿Cuál es el tamaño de la muestra que se necesitaría para detectar una variación de 30 horas en la vida media de las lámparas, usando una significación del 1% y una potencia del 90%.

Solution: A pesar de que el ejercicio anterior se basa en la distribución Normal, podemos emplear la función *power.t.test* como aproximación para calcular el número de muestras. Se obtiene que $n = 241.4 \approx 242$.

10. Una máquina está diseñada para añadir de forma automática patatas a sacos de 10 Kg con una varianza de 0.25^2 Kg^2. Sin embargo, una muestra aleatoria de 20 sacos muestra una varianza muestral de 0.32^2 Kg^2. ¿Podemos concluir que el incremento es significativo al (a) 5%, (b) 1%? Datos en "potatoes.csv".

Solution: En el ámbito industrial, es razonable suponer que una máquina produce muchísimos sacos a lo largo de su vida útil, por lo que las 20 muestras son pocas en comparación con la potencial población. Por otra parte, el histograma de los datos sugiere que los datos son normales, por lo se puede usar el test de Chi-cuadrado. Como queremos ver si hay evidencia como para concluir que hay un incremento: $H_1 : \sigma^2 > 0.25^2$.El estadístico toma el valor 31.13. Dado que el test es de una cola, con distribución Chi-cuadrado con 19 grados de libertad, el p-valor es de 0.03907 y (a) rechazamos H_0; (b) no rechazamos H_0. Deberíamos hacer el estudio con más muestras.

2 ICs y test para dos poblaciones normales

11. Una muestra de 150 bombillas de la marca A mostraron una duración media de 1400 horas y desviación típica 120. Por otra parte, 100 bombillas de la marca B mostraron una duración media de 1200 horas y desviación típica 80. Halla el intervalo de confianza del 99% para la diferencia de medias de las poblaciones A y B. Si tuvieses que comprar unas bombillas, ¿Cuáles comprarías? Datos en "bulbs.csv".

Solution: Asumimos que los tamaños de las muestras son pequeños comparados con la producción de las bombillas de ambos fabricantes, y que ambas poblaciones son independientes (la fabricación en A no influye en B). Los histogramas para ambas poblaciones sugieren que ambas son normales. Por tanto, usando la distribución T-Student, obtenemos que $\mu_a - \mu_b \in (167.2, 232.8)$ horas. Del IC se deduce que las bombillas A tienen una vida media superior a las B. Compraríamos A.

El código del análisis se reproduce a continuación:

```r
bulbs <- read_delim("data_exercises/bulbs.csv",
    delim = ";", escape_double = FALSE, trim_ws = TRUE)

ggplot(bulbs, aes(x = lifetime_hours, fill = brand)) +
  geom_density(alpha=0.2)

A <- bulbs[bulbs$brand == "A", ]
B <- bulbs[bulbs$brand == "B", ]
t.test(A$lifetime_hours, B$lifetime_hours, conf.level = 0.99)
```

12. Una muestra de 100 bombillas de la marca A mostraron una duración media de 1300 horas y desviación típica de 400 horas. Por otra parte, 150 bombillas de la marca B mostraron una duración media de 1200 horas y desviación típica de 200. Halla el intervalo de confianza del 99% para la diferencia de medias de las poblaciones A y B. Si tuvieses que comprar unas bombillas, ¿Cuáles comprarías?

Solution: Asumimos que los tamaños de las muestras son pequeños comparados con la producción de las bombillas de ambos fabricantes, y que ambas poblaciones son independientes (la fabricación en A no influye en B). Los histogramas para ambas poblaciones sugieren que ambas son normales. Por tanto, usando la distribución T-Student, obtenemos $\mu_a - \mu_b \in (-12.92, 212.91)$ horas. En este caso, el IC contiene el 0, por lo que no hay suficiente evidencia en los datos como para concluir que hay una marca mejor. De ahí diríamos que es indiferente comprar una marca u otra. Si tuviésemos que elegir una, elegiríamos la marca A, ya que el IC para $\bar{X}_A - \bar{X}_B$ tiene mayor soporte positivo que negativo.

13. En relación al problema anterior. Si quisiésemos hacer un estudio con más muestras para poder hablar de resultados significativos, ¿cuántas muestras serían necesarias si queremos un nivel de significación del 1% y una potencia del 85%?.

Solution: En el ejercicio anterior, la diferencia de medias muestrales es de 100 horas. Asumiendo que la diferencia real entre medias poblaciones es aproximadamente 100 horas, vamos a hallar el número de muestras que necesitaríamos para detectar, de forma significativa, este efecto. Para ello, podemos usar como primera aproximación, *power.t.test*. El problema es que esta función solo admite un parámetro *sd*. ¿Cuál elegir? Lo mejor es ponerse en el caso pesimista y elegir la desviación típica mayor. En tal caso:

```r
power.t.test(delta = 100, sd = 400, sig.level=0.01, power = 0.85,
             type = "two.sample", alternative = "two.sided")
```

¡Obtenemos que necesitamos unas 420 muestras (es decir, 420 muestras de A y otras 420 de B)! Parecen muchas muestras. Quizás nuestro escenario es demasiado pesimista. Para obtener una mejor estimación, podemos hacer uso de simulaciones:

```r
my_power_t_test <- function(n, n_sims = 5000) {
  sims <- replicate(n_sims, {
    a <- rnorm(n, mean = 1300, sd = 400)
    b <- rnorm(n, mean = 1400, sd = 200)
    p_value <- t.test(a, b)$p.value
    p_value < 0.01
```

```
  })
  mean(sims)
}
my_power_t_test <- Vectorize(my_power_t_test)
ns <- seq(250, 280, by = 1)
powers <- my_power_t_test(ns)
plot(ns, powers, type = "o")
abline(h = 0.85, col = 2, lty = 2)
```

Las simulaciones revelan que con unas 262 muestras (aprox.) bastaría, con el consiguiente ahorro de tiempo y esfuerzo (nuestro jefe seguramente preferirá este análisis que el anterior).

14. Un ingenierio civil desea medir la resistencia a la compresión de dos tipos de cemento, obteniendo las medidas recogidas en el fichero "cement.csv". Asumiendo que ambos tipos de cemento tienen la misma varianza, ¿hay evidencia suficiente como para afirmar que los cementos no tiene la misma resistencia a la compresión (con $\alpha = 0.05$)? ¿Cuál es el tamaño del efecto? Si te vieses forzado a elegir uno de los cementos, ¿qué cemento elegirías?

Solution:

```
cement <- read_csv("data_exercises/cement.csv")
cement$type <- as.factor(cement$type)
ggplot(cement, aes(x = compressive_strength, col = type)) +
  geom_density()

a <- cement[cement$type == 1, ]
b <- cement[cement$type == 2, ]
my_test <- t.test(a$compressive_strength, b$compressive_strength, var.equal = TRUE)
print(my_test)
print(
  effectsize(my_test)
)
```

El p-valor es 0.06872 por lo que, con un nivel de significación del 5%, no hay suficiente evidencia de que los cementos tengan distinta resistencia. Sin embargo, la diferencia entre las resistencias de los cementos muestra un tamaño del efecto grande (Cohen's D=0.82), siendo el IC al 95% de $\mu_I - \mu_{II} \in (-19.2, 474.6)$.

Por tanto, si no es posible repetir el experimento con más muestras, quizás sea preferible elegir el cemento I ya que, de acuerdo a nuestra estimación actual, parece mostrar mayor resistencia a la compresión (aunque tal y como indica el test, no podemos descartar que ambos cementos sean equivalentes).

15. Dos muestras de tamaños 16 y 10 se toman, usando muestreo aleatorio con reemplazamiento, de dos poblaciones distintas. Si sus varianzas muestrales son 25.6 y 20, respectivamente. Halla los intervalos al (a) 98% y (b) 90% para el ratio de varianzas.

Solution: Debemos asumir normalidad y usar la distribución F de Snedecor-Fisher. Los intervalos resultantes son (a) $(0.258, 4.985)$, (b) $(0.426, 3.312)$.

16. En una productora agrícola se deseaba probar el efecto de un fertilizante en la producción de trigo. Para ello, se eligieron 24 parcelas de tierra aproximadamente iguales y se probó el fertilizante en la mitad de ellas. El rendimiento promedio de trigo en las parcelas no tratadas fue de 4.8 Kg con una desviación estándar muestral de 0.42 Kg, mientras que el rendimiento promedio en las parcelas tratadas fue de 5.1 Kg con desviación estándar muestral de 0.37 Kg. ¿Podemos concluir que hay una mejora significativa en la producción de trigo debido al fertilizante si se usa un nivel de significación de (a) 1%, (b) 5%?. (c) ¿Cuál es el p-valor del test? Puedes asumir que las varianzas poblaciones son iguales.

Solution: Asumimos normalidad y que la varianza de las poblaciones es igual. Obtenemos como estimador de la varianza poblacional $\hat{S}_p^2 = 0.15665$. El estadístico es $t = 1.8567$. Teniendo en cuenta que el test es de una cola p-valor = 0.0384., (a) no hay suficiente evidencia a favor de H_1, (b) hay suficiente evidencia a favor de H_1. Deberíamos realizar un estudio más exhaustivo para verificar si el fertilizante es útil. (c)

17. Se quiere estimar mediante un IC (al 95%) el cociente de las varianzas de dos ríos, para poder construir una presa donde estos desembocan (piensa que si hay mucha variabilidad en los cauces de los ríos, debemos tomar precauciones extra a la hora de diseñar la presa; por eso la variabilidad es importante). Datos en *data_ics/rivers.csv*.

Solution: $\sigma_A^2/\sigma_B^2 \in (0.3287699, 0.7016797)$.

18. Según las especificaciones técnicas, la tensión a la que rompen los cables de un fabricante tiene de media 1800 libras. Usando una nueva técnica de manufacturación, el departamento de I+D afirma que la tensión de ruptura puede incrementarse. Como encargado de las pruebas de verificación, tomas una muestra de 50 cables, encontrándose que la tensión media de ruptura es de 1850 libras (datos en "cables.csv"). ¿Hay evidencia como para apoyar la afirmación del departamento de I+D con un nivel de significación del 1%? (Recuerda calcular el tamaño del efecto?).

Solution: Como queremos saber si hay evidencia (en base a los datos) de que se ha incrementado la resistencia, $H_1 : \mu > 1800$. Es de suponer que una muestra de 50 cables representa un número pequeño comparada con la población de cables fabricadas por una empresa. Por tanto, podemos asumir muestras independientes. Dado que la prueba es de una sola cola, el p-valor es de 0.0004597, por lo los datos apoyan que el nuevo proceso incrementa la tensión de ruptura. El IC al 99% sugiere que la nueva tensión de ruptura es superior a 1815 libras. Esta diferencia constituye un tamaño del efecto mediano (Cohen's D = 0.50).

```r
cables <- read_csv("data_exercises/cables.csv")
colnames(cables) <- "tension"

nbins <- nclass.FD(cables$tension)
ggplot(cables, aes(x = tension)) + geom_histogram(bins = nbins)

my_test <- t.test(cables$tension, alternative = "greater", mu = 1800, conf.level = 0.99)
print(my_test)
effectsize(my_test)
```

19. Se quiere estimar mediante un IC al 99%, la diferencia entre los pesos medios de los cerebros de las vacas sanas y vacas enfermas por encefalopatía espongiforme bovina (enfermedad de las vacas locas). Datos en *cows.csv*. ¿Qué conclusión podría sacarse acerca de los pesos de los cerebros?

Solution: Los histogramas para ambas poblaciones sugieren normalidad. Por otra parte, el número de muestras es ridículo comparado con la población mundial de vacas (del orden de 1,000,000,000 de vacas), por lo que podemos asumir independencia en el muestreo. Finalmente, asumimos que los pesos de los cerebros de las vacas sanas son independientes de los pesos de las vacas enfermas. Usando la aproximación de Welch, se obtiene, $\mu_{healthy} - \mu_{sick} = (-1.18, 27.18)$. De acuerdo al intervalo, no parece que existan diferencias entre ambos tipos de cerebros. Ahora bien, el límite inferior está muy cerca de 0, por lo que en una situación realista la decisión lógica sería repetir el experimento pero con más muestras.

20. Repite el ejercicio anterior si pudieses asumir que las desviaciones típicas poblaciones son idénticas.

Solution: En este caso, usamos la varianza agrupada y obtenemos como intervalo $\mu_{healthy} - \mu_{sick} = (0.987, 25.013)$. Ahora se concluiría que los cerebros sanos son más pesados que los enfermos pero, atendiendo de nuevo a las mismas razones que en el ejercicio anterior, lo más sensato sería repetir el experimento con más muestras.

21. En el supuesto del caso anterior (varianzas iguales), si se desea detectar un diferencia de 10 gramos, ¿cuántas muestras son necesarias para obtener una potencia del 90% y un nivel de significación del 1%?

Solution: En el ejercicio anterior se usa la *pooled variance* (o varianza agrupada) para estimar la varianza poblacional σ^2 (que se supone común a ambas poblaciones). Consultando la fórmula en el chuletario de estadística

$$\hat{S}^2_{\text{pooled}} = \frac{(n_1 - 1)\hat{S}_1^2 + (n_2 - 1)\hat{S}_2^2}{n_1 + n_2 - 2}.$$

Usando esta fórmula para especificar *sd* en *power.t.test*:

```
cows <- read.csv("data_exercises/cows.csv")
healthy <- cows[cows$healthy, ]
sick <- cows[!cows$healthy, ]

# Cálculo de pooled SD
n_h <- length(healthy$weight_grams)
v_h <- var(healthy$weight_grams)
n_s <- length(sick$weight_grams)
v_s <- var(sick$weight_grams)
var_pooled <- ((n_h - 1) * v_h + (n_s - 1) * v_s) / (n_h + n_s - 2)
sd_pooled <- sqrt(var_pooled)

power.t.test(
  delta = 10, sig.level = 0.01, power = 0.9,
  sd = sd_pooled, type = "two.sample",
  alternative = "two.sided"
)
```

Se obtienen 427 muestras para cada grupo.

22. A un inversor le preocupa que el *stock 1* sea una inversión más arriesgada que el *stock 2* porque su variabilidad en los precios diarios es mayor. Los datos en "stock.csv" muestran los precios en 100 días

aleatorios de los últimos 5 años para ambos *stocks*. ¿Hay evidencia suficiente como para concluir que la variabilidad del *stock 1* es mayor que la variabilidad del *stock 2*?

Solution: Para proceder con el análisis debemos asumir la independencia de los datos. En este caso podemos asumir independencia porque se han muestreado 100 días aleatorios de un total de $5 \cdot 365$. Sin embargo, si por ejemplo los datos se correspondiesen a 100 días consecutivos, esta asunción sería incorrecta y deberíamos detener el análisis (en la bolsa, los precios del día actual dependen fuertemente del día anterior).

A continuación, debemos comprobar si los datos pueden aproximarse por una distribución normal. Dado que no hay desviaciones evidentes con respecto a la normalidad, podemos continuar el análisis.

Una vez verificadas las asunciones, se trata de un test de varianzas de una sola cola. Usando R, obtenemos que el p-valor es 0.2515, de donde concluiríamos que los datos no aportan evidencia suficiente para concluir que la variabilidad del *stock 1* es mayor que la variabilidad del *stock 2*.

El código R sería:

```
stock_df = read_csv("data_exercises/stock.csv")

ggplot(stock_df, aes(x = prices, fill = stock)) +
    geom_density(alpha = 0.2)

prices <-  split(stock_df$prices, stock_df$stock)
print(
    var.test(
        prices$stock_1, prices$stock_2,
        alternative = "greater"
    )
)
```

23. Se hace el mismo examen a dos cursos de informática. Se toma una muestra con reemplazamiento de la clase A de 16 estudiantes y una muestra de la clase B de 25 estudiantes. Aunque no hay diferencias entre las puntuaciones medias de ambas clases, la muestra de la clase A tiene una varianza de 86.4 mientras que la muestra de la clase B tiene una varianza de 150 (las puntuaciones son sobre 100). ¿Podemos concluir que la varianza de B es mayor que la de A al 5% de significación?

Solution: Debemos asumir que las distribuciones de las puntuaciones son normales (algo que es bastante razonable). Por otra parte, el enunciado menciona que el muestreo ha sido con reemplazamiento, por lo que podemos asumir independencia de muestras (¡si no lo hiciese, deberíamos detener el análisis porque una muestra de 25 estudiantes no parece despreciable con respecto al tamaño total de una clase habitual!). Finalmente, asumimos que las puntuaciones entre clases son independientes.

Usamos un test basado en la F-Snedecor con 24,15 grados de libertad. El valor del estadístico es 1.74. Como el test es de una sola cola, el p-valor asociado es 0.135, por lo que al 5% de significación no rechazamos H_0 (No hay evidencia para afirmar que la varianza de B es mayor que la de A).

24. El *data.frame births* de *openintro* representa una muestra aleatoria de 150 casos de madres y sus recién nacidos en Carolina del Norte durante un año. Estamos particularmente interesados en dos variables: peso y humo. ¿Existe evidencia de que los recién nacidos tienen un peso medio distinto en función de que la madre fume o no?

Solution: Dado que 150 nacimientos son pocos con el número de nacimientos en un estado, podemos asumir la independencia de las muestras. Por otra parte, es razonable asumir que las poblaciones son independientes (el peso de un nacimiento de madre fumadora no influye en los pesos de los bebés de madres no-fumadoras). Para verificar la normalidad, dibujamos las densidades de ambas poblaciones. Ambas poblaciones muestran forma acampanada, aunque también muestran algunas eventos sospechosos en la cola izquierda (debidos a nacimientos prematuros). En general, no son valores muy extremos, por lo que podemos seguir con el análisis. (Por si fuera poco, dado que tenemos bastantes muestras ($n \geq 30$) en ambas poblaciones, también podríamos invocar al TCL para asumir que la distribución de muestreo de $\bar{X}$ es normal).

En base a estas observaciones podemos emplear la aproximación de Welch, que devuelve p-value = 0.138. De aquí se concluye que estos datos no apoyan que existan diferencias entre los pesos de los bebés.

¡Ojo! Esto no implica que podéis recomendar a las futuras madres que fumen. Este problema ilustra como hemos de interpretar con cuidado los resultados de un test estadístico. En primer lugar, recuerda que el hecho de que un test se quede con H0 no implica que el investigador acepte dicha hipótesis como cierta. En segundo lugar, es posible que estemos cometiendo un error tipo II, por lo que sería interesante conocer la potencia del test. Con respecto a este punto, una forma sencilla de mejorar la potencia es recolectar más datos. De hecho, estudios con más muestras que nuestro ejemplo así como estudios de meta-análisis (un análisis estadístico que combina los datos de varios estudios previos para tener muchíiiiiisimas muestras) indican que las madres que fuman tienen hijos más pequeños (y esto no es algo bueno para la salud del niño). En definitiva, ojo con los estudios estadísticos, en especial si no han rechazado la H0.

El código R para llevar a cabo el análisis se muestra a continuación:

```r
library("openintro")
data("births14")  # you may find help in ?births
births <- births

ggplot(births, aes(x=weight, fill=smoke)) +
  geom_density(alpha=0.2)

births[births$weight< 3, ]

# split weights depending on the smoke column
weights = split(births$weight, births$smoke)
print(
    t.test(weights$smoker, weights$nonsmoker)
)
```

3 ICs y test para poblaciones no-normales y datos categóricos

25. Como parte de un estudio de cierta liga de fútbol para determinar si jugar como local influye en el resultado, se contabilizan el número de partidos ganados como local. Del total de 420 partidos, 280 los ha ganado el equipo local. ¿Cuál es el intervalo de confianza de la proporción de partidos ganados como local? Usa un nivel de confianza del 99%. En base, a este intervalo de confianza, ¿Hay evidencia suficiente como para afirmar que la probabilidad de ganar como local es diferente a 50%?

Solution: $p \in (0.6045627, 0.7249134)$. De aquí, sí hay evidencia suficiente (con $alpha = 0.01$) como para afirmar que $p \neq 0.5$ ya que 0.5 no pertenece al IC.

26. En el ejercicio anterior, ¿hay evidencia suficiente como para afirmar que es más probable ganar como local que como visitante?

Solution: Siendo estrictos, no deberíamos responder a esta pregunta empleando el IC anterior, ya que la $Ha : p > 0.5$. Para responder a esta pregunta, deberíamos emplear un IC de una sola cola. Si calculamos este intervalo, obtenemos $p > 0.6106$, por lo que sí hay evidencia de que es más probable ganar como local que como visitante.

27. Demuestra mediante simulaciones en R que si n es grande, la población normal y el muestreo es aleatorio con reemplazamiento; entonces la distribución de la desviación típica muestral puede aproximarse por $\mathcal{N}(\sigma, \sigma^2/2n)$.

28. La desviación estándar muestral de 200 bombillas eléctricas es de 100 horas. Encuentra el intervalo de confianza al 99% de la varianza de la población de bombillas. ¿Cuál sería el intervalo de confianza para la desviación típica?

Solution: Asumiendo normalidad y muestreo con reemplazamiento/población infinita podemos usar la distribución Chi-cuadrado, de donde obtenemos $(7830.479, 13146.600)$ [horas^2] Para el IC de la desviación típica tenemos dos opciones: (1) aproximar el IC como la raíz cuadrada del IC de la varianza (aunque esto es solo una aproximación sin mucha justificación) y obtener $(88.490, 114.656)$ [horas] o (2) usar la aproximación normal del anterior ejercicio y usar $\hat{s} \pm z_c \hat{s}/\sqrt{2n} = (87.121, 112.879)$ [horas] (preferible).

29. En una universidad de una ciudad grande hay 5000 alumnos. En una muestra aleatoria de 200 alumnos de primero de esta universidad, el 35% dijeron que pretendían trabajar de 16 a 20 horas a la semana para ganar dinero. Hallar el intervalo de confianza del 95% para el número esperado de alumnos de primero de dicha universidad que pretenden trabajar entre 16 y 20 horas a la semana.

Solution: Dado que $200 \ll 5000$, podemos asumir que las muestras son independientes. Usando la distribución binomial se obtiene $5000 \cdot (0.284, 0.420) \approx (1420, 2102)$ alumnos.

30. Trabajas para una farmacéutica que desea probar la efectividad de un tratamiento en personas enfermas. Para ello, se prueba el tratamiento en 100 personas enfermas, mientras que a otro grupo de 100 personas se le da un tratamiento placebo (este será el grupo de control). El fichero "recovered.csv" recoge cuantos de los pacientes de cada uno de los grupos se ha curado de la enfermedad. Testea la hipótesis de que el tratamiento es efectivo al (a) 0.01, (b) 0.05.

Solution: Dado que trabajamos para la farmacéutica, $H_1 : p_t > p_c$. Asumiendo que la probabilidad de recuperación de una persona no influye en la probabilidad de recuperación de otra persona (independencia), el test es binomial de una cola, con p-valor = 0.0825, y (a) no rechazamos H_0 (No hay suficiente evidencia de que el tratamiento sea efectivo en comparación con el placebo); (b) no rechazamos H_0 (No hay suficiente evidencia de que el tratamiento sea efectivo en comparación con el placebo).

```
df <- read_csv("data_exercises/recovered.csv")
counts <- table(df$group, df$has_recovered)
prop.test(counts[, "Recovered"], rowSums(counts), alternative = "less")
```

31. El número de fraudes en compañías de seguros de un país sigue una distribución de Poisson de parámetro λ. En una muestra de 100 compañías, se obtuvo una media de 0.85 fraudes. Determina el IC al 95% para λ.

Solution: Sea X la VA: número de fraudes en compañías de seguros. Para estimar λ, usamos que $[E][X] = \lambda$, por lo que un estimador de λ es $\bar{X}$. Si asumimos que el muestreo es con reemplazamiento, las muestras son independientes y podemos usar el teorema central del límite (dado que el número de muestras es grande). Por tanto, $\frac{\bar{X}-\lambda}{\sqrt{\lambda/n}} \overset{asintóticamente}{\sim} \mathcal{N}(0,1)$. De aquí, $\lambda \in \bar{X} \pm z_{1-\alpha/2}\sqrt{\lambda/n}$, pero como no conocemos λ, aproximamos el IC por $\lambda \in \bar{X} \pm z_{1-\alpha/2}\sqrt{\bar{X}/n}$. Finalmente se obtiene $(0.67, 1.03)$.

32. El fabricante de una patente médica afirma que es efectiva en al menos un 90% de los casos en aliviar una alergia por un período de 8 horas. En una muestra de 200 personas con alergia, la medicina alivió los síntomas de 160 personas.
    (a) Trabajas para una agencia de control de medicamentos y sospechas que la afirmación del fabricante no es cierta. Determina si la afirmación del fabricante es legítima usando un 1% como nivel de significación.
    (b) Encuentra el p-valor del test.

Solution: Asumimos la afirmación del fabricante como cierta (por "presunción de inocencia': asumimos que dice la verdad): $H_0 : p > 0.9$). Por tanto, $H_1 : p < 0.9$. El P-valor es $\approx 1.7 \cdot 10^{-5}$, por lo que existe evidencia suficiente de que $p < 0.9$.

4 Popurrí

33. A una muestra de 10 estudiantes se les hace un examen antes de estudiar una lección en particular. Una vez completada la lección, se les vuelve a hacer la misma prueba. Queremos descubrir si, en general, el método de enseñanza empleado conduce a mejoras en las habilidades de los estudiantes (es decir, a mejores en los resultados del examen). Los datos recogidos se encuentran en "lesson.tsv".

    ¿Existe evidencia estadística (al 5% de significación) de que el método de enseñanza es efectivo?

Solution: Asumimos que el número de estudiantes participantes en el estudio (10) es muy pequeño comparado con los estudiantes susceptibles de emplear el nuevo método de enseñanza. Por otra parte, observamos que se trata de un experimento con datos apareados. El histograma de las diferencias revela que estas son normales, por lo que podemos emplear un t-test de datos apareados. Dado que queremos saber si el método de enseñanza es efectivo, la H_1 es $H_1 : \mu_{despues} > \mu_{antes}$. El estadístico es $t = -2.2616$ con p-valor 0.02502. Podemos afirmar, a un nivel del 5% de significación, que el método de enseñanza es efectivo.

```r
lesson <- read_table("data_exercises/lesson.tsv")

# Para testear normalidad en test apareados, usamos las diferencias
d <- lesson$después - lesson$antes
hist(d)

t.test(lesson$después, lesson$antes, paired = TRUE, alternative = "greater")
```

34. [OS] ¿Es efectivo el tratamiento con células madre embrionarias para mejorar la función cardíaca tras un ataque cardíaco? Los datos en el *data.frame stem_cells* de *openintro* contienen los resultados de un estudio con ovejas. Las ovejas fueron asignadas aleatoriamente al grupo de control (sin tratamiento) o al grupo de tratamiento. Además, se les midió la capacidad de bombeo de sus corazones al principio y final del estudio. Los investigadores desean

    (a) que calcules la mejora en la capacidad de bombeo de las ovejas ($\text{funcionamiento}_{\text{final}} - \text{funcionamiento}_{\text{principio}}$), creando una nueva columna en el *data.frame*.

    (b) que calcules la media de estas diferencias por grupo (tratamiento y control) usando funciones (y sin repetir código).

    (c) que compruebes si hay evidencia a favor de que el tratamiento mejora la función cardíaca. No hace falta que verifiques la normalidad de los datos, ya que ya ha sido verificada por los demás investigadores.

Solution: Dado que los datos son normales y es razonable asumir independencia entre poblaciones y muestras, podemos usar la aproximación de Welch. Teniendo en cuenta que el test es de una sola cola, obtenemos que el p-valor es 0.0008388, de donde concluimos que existe evidencia a favor de que el tratamiento mejora el funcionamiento del corazón.

```r
library("openintro")
data("stem_cell")
# a) Let's use with. Can you remember what with does?
stem_cell$diff = with(stem_cell, after - before)

# b)
# Method 1)
trmts = unique(stem_cell$trmt)
summary = sapply(
    trmts,
    function(trmt) mean(stem_cell[stem_cell$trmt == trmt, ]$diff)
)
names(summary) = trmts
print(summary)
# Method 2) Let's introduce a new function, tapply
tapply(stem_cell$diff, stem_cell$trmt, mean)

# c)
diffs = split(data$diff, data$trmt)
print(
  t.test(
    diffs$esc, diffs$ctrl,
    alternative = "greater"
```

```
  )
)
```

35. El procesador insignia de cierto fabricante es capaz de trabajar a 4.5 GHz. Para actualizar su diseño, han contratado un equipo de ingenieros informáticos del CEU. Usando nuevas técnicas, el equipo de diseño afirma que ha logrado incrementar la frecuencia de trabajo del procesador. Para testear esta afirmación, se encarga a una planta de fabricación 50 procesadores manufacturados con la nueva técnica, encontrándose que la media muestral de la frecuencia de trabajo es de 4.7 GHz y la desviación típica muestral es de 0.4 GHz. ¿Podemos apoyar la afirmación del equipo de diseño?, es decir, ¿tenemos evidencia para afirmar que el procesador funciona a más de 4.5 GHz? Usa un nivel de significación de 0.01. Asume que es razonable suponer normalidad en la distribución de las frecuencias de trabajo.

Solution: Asumimos que la fabricación de los 50 procesadores son procesos independientes. Para que esto sea razonable, el experimento debería planearse minuciosamente. Por ejemplo, si las máquinas que fabrican los procesadores tardan en calentarse, los procesadores fabricados al principio del proceso (mientras la máquina calienta) tendrán características similares entre sí (no serán independientes). Para prevenir este tipo de problemas, la fabricación de las muestras no debería comenzar hasta que las máquinas estuviesen calientes. Asimismo, deberíamos permitir que haya cierto tiempo entre la fabricación de los distintos procesadores. Asumimos que todo esto se cumple y que los procesadores son independientes.

Dado que se nos pregunta si existe evidencia apoyando que el procesador funciona a más de 4.5 GHz, $H_1 : \mu > 4.5\, GHz$. El estadístico es $t = 3.5355$ y se distribuye como una t-Student con 49 grados de libertad. Al ser el test de una cola el p-valor es 0.00045 por lo que hay una gran evidencia a favor de que el procesador funciona a más de 4.5 GHz.

36. Un historiador está investigando sobre la intoxicación por plomo a lo largo de la historia. Los datos de "pb.txt" se obtuvieron a partir de una comparación del contenido de plomo del cabello humano extraído de individuos adultos muertos entre 1880 y 1920 con el contenido de plomo de los adultos actuales.

Al historiador le gustaría demostrar que los niveles de plomo en la antigüedad son superiores en al menos $10\mu g$ a los niveles de plomo de hoy en día. Haz el análisis estadístico adecuado teniendo en cuenta que hay evidencia que sugiere que las desviaciones estándar de ambas poblaciones son iguales.

Solution: El número de muestras es pequeño comparado con la cantidad de gente que ha vivido/vive en la antigüedad/hoy en día, por lo que podemos asumir independencia entre muestras. Además, ambas poblaciones son independientes (al ser períodos distintos, no se influyen entre sí). Los histogramas revelan que ambas muestras tienen distribución normal. En base a todo esto, podemos emplear el t.test.

La Ha es $\mu_{antig} > \mu_{hoy} + 10$. Dado que se puede asumir que las varianzas de las poblaciones son iguales, el estadístico toma el valor 4.4551. El p-valor asociado es muy pequeño ($\approx 9 \cdot 10^{-6}$; prueba de una cola), por lo que hay evidencia que apoya que los niveles de plomo en la antigüedad son mayores que los de hoy en día en, al menos, $10\mu g$. La diferencia entre las cantidades de plomo es superior a 17.47 μg (usando un nivel de confianza del 95%). El tamaño del efecto correspondiente a esta diferencia es grande (Cohen's D=0.79),

37. [OS] ¿Son los libros de texto realmente más baratos si se compran *online*? En este ejercicio comparamos el precio de los libros de texto en la librería de la Universidad de California (UCLA) y los precios en *Amazon.com*. Setenta y tres cursos de UCLA fueron muestreados al azar en la primavera de 2010 (lo

que representa menos del 10% de todos los cursos de UCLA) obteniendo los precios disponibles en el *data.frame textbooks* de *openintro*. ¿Hay diferencias estadísticamente significativas en los precios?

Solution: En este caso, tenemos datos apareados. Como primer paso, analizamos si podemos asumir normalidad inspeccionando el histograma de las diferencias. Los datos son muy asimétricos, pero podemos seguir adelante ya que es posible usar el teorema central del límite para asumir la normalidad de la media muestral. Esto es debido a que n es bastante grande y, por otra parte, la muestra emplea solo el 10% de la población, por lo que podemos aproximar el muestreo como un muestreo de población infinita (equivalente a con reemplazamiento).

Usando el *t.test* para datos apareados encontramos que el p-valor es $6.928e-11$, de donde concluimos que existen diferencias significativas entre los precios de los libros. Observando el IC para las diferencias de medias, podemos aseverar que los libros de amazon son más baratos que en la UCLA (¡12$ más baratos en media!).

```
library("openintro")
data("textbooks")

textbooks$diff_prices = textbooks$ucla_new - textbooks$amaz_new

# Histograma sobre diferencias
hist(textbooks$diff_prices)

t.test(textbooks$ucla_new, textbooks$amaz_new, paired = TRUE)
```

38. [OS] Estamos interesados en saber si el corredor estadounidense medio se está volviendo más rápido o más lento. Para responder a esta pregunta, usaremos datos de la Cherry Blossom Race, que es una carrera de 10 millas en Washington, DC. El tiempo promedio para todos los corredores que terminaron la carrera en 2006 fue de 93 minutos y (aproximadamente) 17.4 segundos. Queremos determinar, utilizando datos de los participantes en el año 2012, si los corredores en esta carrera son cada vez más rápidos o más lentos, frente a la otra posibilidad de que no haya habido cambios. Datos en el paquete *cherrybossom*, *data.frame run12*.

Solution: En primer lugar, es de resaltar que el muestreo es sin reemplazamiento (los participantes en la carrera no pueden aparecer repetidos). Sin embargo, dado que es lógico suponer que el número de participantes en la carrera (la muestra) es mucho menor que el número total de corredores en Estados Unidos, podemos asumir que la población es infinita y que las muestras son, por tanto, independientes. A continuación, debemos comprobar si los datos pueden aproximarse por una distribución normal (con *hist*). Los datos parecen ligeramente asimétricos (la cola a la derecha es ligeramente más pesada) pero, dado que la forma es razonablemente acampanada y dado que n es grande (podemos invocar el teorema central del límite por la independencia de las muestras), podemos seguir adelante con el análisis. Planteamos que H0 es $\mu = 93.29$ minutos (no ha habido cambios) y H1 es $\mu \neq 93.29$. Empleando entonces *t.test* obtenemos que el p-valor es $< 2.2 \cdot 10^{-12}$, de donde concluimos que los datos sugieren que ha habido un cambio en los tiempos de los corredores. Con un nivel de confianza del 95%, el nuevo tiempo medio es $\mu \in (94.3, 94.8)$ minutos (más lentos). Sin embargo, el tamaño del efecto es pequeño (Cohen's D = 0.08). (Es decir, que el test sea significativo significa que estamos seguros de que hay una diferencia en los tiempos, el hecho de que el tamaño del efecto es pequeño significa que esta diferencia es pequeña).

El código R para llevar a cabo el análisis se muestra a continuación:

```r
library("cherryblossom")
data("run12")
# 93 min and 17.4 secs  = 93 + 17/60
mu = 93 + 17.4/60

nbins <- nclass.FD(run12$time)
ggplot(run12, aes(x=time)) + geom_histogram(bins = nbins)

my_test <- t.test(run12$time, mu = mu)

print(my_test)
print(
   effectsize(my_test)
)
```

39. En un experimento de percepción extrasensorial (PES) se le pide a un sujeto que describa el color (rojo o azul) de una carta obtenida al azar de una baraja de 50 cartas (barajadas). La carta se extrae en una habitación contigua, sin que el sujeto que adivina pueda ver la carta. El sujeto identifica 32 cartas correctamente (de las 50 posibles). Supón los siguientes casos:
    (a) Dado que somos escépticos acerca de la existencia de la PES, determina si hay suficiente evidencia como para descartar que el individuo posee PES al 1% de significación.
    (b) Hemos ingresado en una secta de PES, por lo que queremos demostrar que la PES es real. Repite el análisis al 1% de significación.

Solution: (a) Asumimos que acertar/fallar una carta no influye en el resultado de intentar adivinar la siguiente: las muestras son independientes. Por tanto, la distribución es binomial. Queremos saber si hay suficiente evidencia como para afirmar que $p <= 0.5$. Por tanto, $H_1 : p <= 0.5$. Teniendo en cuenta que el test es de una sola cola, el p-valor es 0.9836, por lo no tenemos suficiente evidencia en contra del PES.

(b) En este caso, queremos demostrar que $H_1 : p > 0.5$. El estadístico sigue tomando el mismo valor, pero el p-valor es 0.03245. En este caso, tampoco rechazamos H_0, por lo que no hay evidencia a favor del PES.

Si la aparente contradicción entre (a) y (b) te parece sorprendente, revisa las clases.

40. Una casa farmacéutica es capaz de producir una determinada pastilla de una forma muy precisa. Esto es, el peso de las pastillas producidas tiene una desviación estándar muy pequeña. El equipo de investigación de la empresa ha propuesto un nuevo método de producción del fármaco. Sin embargo, esto conlleva algunos costes y se adoptará solo si hay pruebas sólidas de que la desviación estándar del peso de los artículos caerá por debajo de 0.4 miligramos (mg) con el nuevo método (¡mejor que la precisión actual!). Se produce una muestra de 10 medicamentos con el nuevo método obteniéndose que los medicamentos tienen una media de 5.724 mg y una desviación estándar de 0.35 mg. Con un nivel de significación del 1%, ¿hay evidencia que apoye la adopción del nuevo método?

Solution: Ha es $\sigma^2 < 0.4^2$. Bajo las hipótesis habituales (¿cuáles?) podemos emplear el estadístico $(n-1)\hat{S}^2/\sigma^2$ que, bajo la H0, toma el valor $t = 6.89$ y se distribuye como una χ^2_9. El p-valor es 0.351 y no hay evidencia suficiente a favor de Ha. No adoptaríamos el nuevo método.

Fuentes

- [OS]: Christopher D. Barr et al. OpenIntro Statistics, Third Edition.
- [Ross]: Ross, Sheldon M. Introduction to probability models. Academic press, 2014.
- [Dobrow]: Dobrow, Robert P. Probability: with applications and R. John Wiley & Sons, 2013.
- [Spiegel]: Spiegel et at. Probabilidad y estadística, Schaum, segunda edición, 2003.