Tema 4: Intervalos de Confianza

COMPARACIONES DE MEDIAS EN POBLACIONES NORMALES

Varianzas Totalmente desconocidas

Ejercicio: Diferencias por sexos

Los datos contenidos en "howell1.csv" son datos censales parciales del área !Kung San compilados a partir de entrevistas realizadas a finales de la década de 1960. ¿Depende la altura de los !Kung adultos del sexo del inviduo? ($\alpha = 0.01$).

Apoya tus resultados con un gráfico y calcula el tamaño del efecto. Emplea los datos en "howell1.csv".

X: " altura hombre adulto "
Y: " altura mujer adulta "
▸ la altura se asume q sigue una distrib. Normal

$X \sim N(mu\_y, \sigma_y)$ & $Y \sim N(mu\_x, \sigma_x)$

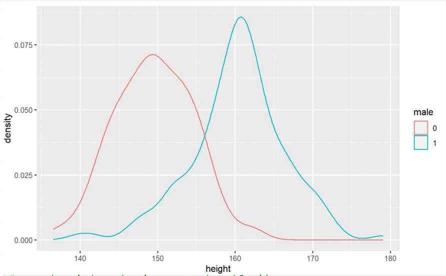
```r
library("tidyverse")
library("effectsize") #necesaria para medias

#Tenemos q cambiar el delimitador --> a ;
library(readr)
howell <- read_delim("data_3(1)/data_3/howell1.csv",
    delim = ";", escape_double = FALSE, trim_ws = TRUE)

library("ggplot2")

#Tenemos que filtrar datos --> SOLO ADULTOS (edad>18)
howell_adult = howell[howell$age >= 18, ]
#Diferenciar  MUJERES y HOMBRES
#Los datos categóricos se deben codificar como factores
  #Columna "male": si es hombre 1, sino 0
  #Tengo q convertir mi columna "male" a tipo factor
howell_adult$male = as.factor(howell_adult$male)
#Nueva geometría: densidad (para representar mejor dos poblaciones)
ggplot(howell_adult, aes(x =height, col=male)) + geom_density()
print("ambos datos tienen una distribución normal")
```

▸ Al hacer import se puede cambiar




```r
#Comprobar independecia para cada población
  #Población infinita o muestreo sin reemplazamiento
nrow(howell_adult) #352 muestras
print("puesto q 350 personas << población objetivo, luego asumimos independencia")
```

¿Depende la altura del sexo?

Ha: $mu\_x \neq mu\_y$; $mu\_x - mu\_y \neq 0$

Ho: $mu\_x = mu\_y$; $mu\_x - mu\_y = 0$

¿Qué estadístico hay para la media si no conocemos la varianza?

↳ 00 Tabla : T de Student
- → Distrib Normal (por separado)
- → Indep. de muestras para X e Y
- → Indep entre X e Y

```r
#Comprobar independecia X entre Y
print("es razonable asumir que la altura de los hombres no afecta a la de las mujeres")
```

```r
#Ha: mu_x != mu_y --> mu_x - mu_y != 0 (dos colas)
  men = howell_adult[howell_adult$male == 1, ]
  women = howell_adult[howell_adult$male == 0, ]


my_test = t.test(men$height, women$height,
       alternative = "two.sided",
       mu = 0) #IMPORTA el orden de los datos
my_test
```

→ Si ponemos 1º women nos saldrán negativos los intervalos de conf.

```
        Welch Two Sample t-test

  data:  men$height and women$height
  t = 18.148, df = 323, p-value < 2.2e-16
  alternative hypothesis: true difference in means is not equal to 0
  95 percent confidence interval:
    9.669319 12.020596
  sample estimates:
  mean of x mean of y
   160.3585  149.5135
```

```r
print("pvalor es muy muy bajo, los datos apoyan Ha")
print("Los datos apoyan que la altura de los hombres es distinta a la de las mujeres")

#el intervalo de confianza; es una estimación  entre las diferencias de alturas
print("La altura de los hombres está entre 9.5 y 12cm más que la de las mujeres")
```

```r
#CALCULAR EL TAMAÑO DEL EFECTO
effectsize(my_test)
#tamaño del efecto: cohens_d = 1,95[tamaño grande según enlace]
print("puesto que el tamaño del efecto es muy grande,
      los resultados son MUY RELEVANTES")
```

```
------------------------
1.95      | [1.69, 2.21]

- Estimated using un-pooled SD.
```

→ desv. típica no agrupada

coge datos X → sd(X) ▸ uso de n_x datos → estimación

coge datos Y → sd(Y) ▸ uso de n_y datos → Estimación

SOLO si sd(X) = sd(Y) ▸ uso de n_x y n_y → estimación

(100% seguro) (↑ num muestras = mejor)

[varianzas desconocidas pero iguales]

Repite el ejercicio relativo a los !Kung adultos si se puede asumir que la desviación típica poblacional para hombres y mujeres es la misma ($\sigma h = \sigma m$).

```r
library("tidyverse")
library("effectsize") #necesaria para medias

#Tenemos q cambiar el delimitador --> a ;
library(readr)
howell <- read_delim("data_3(1)/data_3/howell1.csv",
    delim = ";", escape_double = FALSE, trim_ws = TRUE)

library("ggplot2")

my_test2 = t.test(men$height, women$height, alternativa = "two.sided", var.equal = TRUE)
      # var.equal = asumo q las desviaciones típicas son las mismas
my_test2

effectsize(my_test2)
#Estimated using pooled SD
```

```
## [1] 4.068028e-53
## Cohen's d |        95% CI
## ------------------------
## 1.96      | [1.70, 2.21]
##
## - Estimated using pooled SD.
```

¿Hay evidencia de q los ♂ miden 10cm más q las ♀?

Ha: $mu\_x - mu\_y \geq 10$

Ho: $mu\_x - mu\_y < 10$

Los pasos son los mismos pero cambiamos:

→ Ha: de una cola

```r
my_test = t.test(men$height, women$height,
       alternative = "greater",
       mu = 10)
```

```
        Welch Two Sample t-test

  data:  men$height and women$height
  t = 1.414, df = 323, p-value = 0.07917
  alternative hypothesis: true difference in means is greater than 10
  95 percent confidence interval:
    9.8592      Inf
  sample estimates:
  mean of x mean of y
   160.3585  149.5135
```

```r
#el intervalo de confianza; es una estimación  entre las diferencias de alturas
print("La altura de los hombres está entre 9.66 y 12cm más que la de las mujeres")
print("Luego rechazamos la Ha ")
```

Tema 4: Intervalos de Confianza

COMPARACIONES DE MEDIAS EN POBLACIONES NORMALES

Datos apareados

Medimos algo $\xrightarrow[\text{operación}]{\text{Hacemos algo}}$ Repetimos

gbef - - - - - - - - - - - - → gbef

Ejercicio: Datos apareados

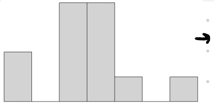
Unos científicos examinaron la función de la vesícula biliar antes y después de una cirugía para detener el reflujo. Los autores midieron la funcionalidad de la vesícula biliar calculando la fracción de eyección de la vesícula biliar (GBEF) antes y después de la operación, cuyo objetivo es aumentar la GBEF. ¿Hay evidencia para concluir que la operación aumenta el GBEF? Datos en "gbef_long.txt" (o "gbef.txt", para un reto).

X: "antes de operarse"
Y: "Después de operarse"
} son DEPENDIENTES

```
library("tidyverse")
library("effectsize")
#renombramos cambiando su extension a csv + delimitador a whitespace
library(readr)
gbef_long <- read_table("C:/Users/ppere/Desktop/2ndo_carrera/estadistica_I/archivos_estadistica/
tema_4/data_3(1)/data_3/gbef_long.csv")
colnames (gbef_long)= c("ID", "class", "gbef") #Quitar dobles comillas

#tengo dos medidas para el individuo--> vamos a crear otra variable
gbef_long_preop = gbef_long[gbef_long$class =="\"Preop\"", ]
gbef_long_postop = gbef_long[gbef_long$class =="\"Postop\"", ]

dif_despues_antes = gbef_long_postop$gbef - gbef_long_preop$gbef #resta de gbef
hist(dif_despues_antes)
#comparar incremento y decremento para cada uno de los individuos
```



→ Es Normal

1) Formular hipótesis:

Ha: mu-postop > mu-preop --→ mu_postop - mupreop > 0

: dif- despues - antes > 0

Ho: dif- despues - antes < 0

2) Discutir asunciones

↳ Indep. entre sí [NO se discute]

↳ ¿Población ∞ o Muestreo? ✓✓

num_pacientes <<< Pacientes potenciales

↳ Normalidad de Poblaciones ✓✓

histograma

3) Aplico t.test

```
#Ha: : dif_despues_antes > 0
t.test(dif_despues_antes, alternative = "greater", mu = 0)

#Ha: mu_postop > mu_preop --> mu_postop - mu_preop > 0
 t.test(gbef_long_postop$gbef,  gbef_long_preop$gbef,
       alternative = "greater", mu = 0,
       paired = TRUE #Datos apareados
       )
```

```
        Paired t-test

data:  gbef_long_postop$gbef and gbef_long_preop$gbef
t = 1.9159, df = 11, p-value = 0.04086
alternative hypothesis: true mean difference is greater than 0
95 percent confidence interval:
 1.131919      Inf
sample estimates:
mean difference
         18.075
```

18.075 → incremento medio

0.01 < pvalor → nivel significación

4) Calculo Tamaño del efecto ($\alpha = 0.01$)

```
my_test =  t.test(dif_despues_antes,
                  alternative = "greater", mu = 0,
                  conf.level = 0.99)
  effectsize(my_test)
```

```
## [1] "Cohens'd : 0.553061950304492"
```

```
print("los datos no apoyan suficiente evidencia que la operación
      aumenta el GBEF (nivel significación < pvalor)
      Aunque no lo apoyen el tamaño del efecto es medio
      (incremento medio: 18.075" )

      #es cierto que aumenta el gbef con una media de 18.075,
      #pero no podemos descartar que la operacion no aumente
      #el GBEF, es posible que la operación no aumente el GBEF

print ("Si quisiesemos estar al 95% seguros (alpha=0.5) el
       estudio saldría válido")

#------RECOMENDAMOS MÁS MUESTRAS -----------
#alpha = 0.01 y potencia del 0.9
  power.t.test(delta = 18.075,
               sd = sd(dif_despues_antes),
               sig.level = 0.01,
               power = 0.9,
               type = "one.sample",
               alternative = "one.sided")
print("Deberiamos repetir la prueba con unos 46 individuos")
```

Comparación de Varianzas

Varianza dos Poblaciones Normales

$X_1 \sim N(\mu_1, \sigma_1^2)$ y $X_2 \sim N(\mu_2, \sigma_2^2)$ → $\dfrac{\sigma_1^2}{\sigma_2^2} \begin{cases} < 1 & \sigma_1^2 < \sigma_2^2 \\ \simeq 1 & \sigma_1^2 \simeq \sigma_2^2 \\ > 1 & \sigma_1^2 > \sigma_2^2 \end{cases}$

$$\frac{\hat{S}_1^2/\sigma_1^2}{\hat{S}_2^2/\sigma_2^2}$$ ▶ F de Snedecor o Fisher-Snedecor → $$\frac{\hat{S}_1^2/\sigma_1^2}{\hat{S}_2^2/\sigma_2^2} \sim F_{\underbrace{n_1-1,\, n_2-1}}$$

con n_1-1, n_2-1 → grados de libertad

Ejercicio:

En una empresa, se están comparando dos métodos de producción de cierto chip (A, mucho más barato, y B). La potencia media consumida por ambos chips es idéntica, si bien los dos métodos tienen distinta variabilidad. Se obtienen dos muestras de tamaño 16 y 10 y sus varianzas muestrales son 24 y 18 (en Watts^2). Usando un nivel de confianza del 98 %, ¿qué método es preferible? Usa la función *var.test*.

A: "Potencia Chips A" ▷ $A \sim N(\mu, \sigma_A)$
B: "Potencia Chips B" ▷ $B \sim N(\mu, \sigma_B)$
▶ Si tenemos poca varianza, el resultado de los Chips es predecible (mejor!!)

1) Calcular intervalo de confianza [pq no hay hipótesis]

enunciado → nivel de confianza = 98 %

2) Asunciones:

↳ Indep. entre muestras: 16 y 10 <<< Cantidad q fabrican en total ✓✓

↳ Indep entre poblaciones: método fabricación A no influye en B ✓✓

↳ Normalidad de Poblaciones: LO ASUMO

3) Uso del var.test:

→ generar datos • `sim = rnorm(16, mean = 3, sd=sqrt(24))`

10 Muestras
$\sigma = 24$

→ Estandarizar las simulaciones ▶ `sim_z = (sim - mean(sim))/sd(sim)`

```
mean(sim_z) #mas o menos 0
mean(sim_z)+3 #Tengo datos de media 3
var(sim_z * sqrt(24)) #datos para la varianza 24

 #Estandarización para este problema en concreto
 sim_majas = sim_z * sqrt(24) + 3
```

Tema 4: Intervalos de Confianza

Test e IC's para poblaciones no-normales

(Comparación de Varianzas)

```r
sim_a = rnorm(16, sd=sqrt(24))
sims_zA = (sim_a - mean(sim_a))/sd(sim_a)* sqrt(24)

sim_b = rnorm(10, sd=sqrt(18))
sims_zB = (sim_b - mean(sim_b))/sd(sim_b)* sqrt(18)

var.test(sims_zA, sims_zB, conf.level = 0.98)
#var pobl A / var pobl B está en (0.2687046, 5.1930508)
```

```
        F test to compare two variances

data:  sims_zA and sims_zB
F = 1.3333, num df = 15, denom df = 9, p-value = 0.6776
alternative hypothesis: true ratio of variances is not equal to 1
98 percent confidence interval:
 0.2687046 5.1930508
sample estimates:
ratio of variances
          1.333333
```

Teorema Central del límite: $x_1, x_2, \dots x_n$ [todos son indep]

↳ $x_1 + x_2 + \dots x_n$ es aproximadamente normal

$\Rightarrow E[x_1 \dots] = E[x_1] + E[x_2] + \dots E[x_n] = n \cdot \text{lambda}$

$\Rightarrow Var[x_1 + \dots] \xrightarrow{\text{son indep}} = Var[x_1] + \dots Var[x_n] = n \cdot \text{lambda}$

$x_1 + x_2 + \dots x_n \sim Normal(n * \lambda, var = n * \lambda)$

Media muestral: $\frac{x_1 + x_2 + \dots x_n}{n} \sim Normal\left(\lambda, \frac{\lambda}{n}\right)$

las ctes salen elevados al cuadrado!!

Datos no-normales

Ejercicio: Fútbol y Poisson

Bajo ciertas condiciones, el número de goles marcados en un partido de fútbol se puede aproximar por una distribución de Poisson. El fichero "spain_league.csv" contiene datos sobre la liga de fútbol española. Para hacer tus apuestas deportivas, es importante saber el número de goles promedio por equipo en cada partido. Usando los resultados de la liga 21-22, construye un intervalo de confianza del 98 % para el promedio de goles del equipo local en un partido.

X: "num de goles en un partido" → $X \sim Poiss(\lambda)$

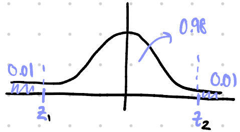
↳ $= E[X] = mean()$

$\bar{X} \sim N(\lambda, var = \lambda/n)$ (aprox.) $\begin{cases} \text{• Sucesos indep.} \\ \text{• num grande de observación } (n >>> 30) \end{cases}$

1) $P(a < \bar{X} < b) = 0.98$

$z = \frac{\bar{X} - \lambda}{\sqrt{\lambda/n}} \rightsquigarrow \frac{\bar{X} - \lambda}{\sqrt{\bar{X}/n}}$

en estos casos se puede calcular así



$z_2 = qnorm(0.99) = 2.33$

$z_1 = qnorm(0.01) = -2.33$

2) $\left(-2.33 < \frac{\bar{X} - \lambda}{\sqrt{\bar{X}/n}} < 2.33\right) \rightsquigarrow \bar{X} - 2.33\sqrt{\frac{\bar{X}}{n}} < \lambda < \bar{X} + 2.33\sqrt{\frac{\bar{X}}{n}}$

mean()

```r
library("tidyverse")
library("effectsize")

library(readr)
spain_league <- read_csv("data_4/data_4/spain_league.csv")
spain_2021 = spain_league[spain_league$Season==2021, ]

goles_local = spain_2021$hgoal
#Porque no usar los vgoal? para que sean independientes
lamda_est = mean(goles_local)
```

```r
lamda_est + z *sqrt(lamda_est/length(goles_local))
#Calcular cuantiles
z = c(z1 = qnorm(0.01), z2 = qnorm(0.99))
lamda_est + z *sqrt(lamda_est/length(goles_local))
```

→

```
      z1       z2
1.278791 1.563314
```

Datos Tabulares

Ejercicio: Racismo en la selección de jurados

Durante los 60s-70s, se dieron casos de racismo en la elección de jurados populares. Supuestamente, la elección es al azar entre un listado de todos los ciudadanos. Sin embargo, se daban situaciones como que en una preselección de 80 posibles jurados solo 4 fuesen afroamericanos (de una población con un 50 % de afroamericanos). Datos en "juries.csv". Las autoridades se defendían diciendo que era pura casualidad. ¿Es esto creíble? Apoya tus conclusiones con gráficos.

```r
```{r}
prop.test(4,80, alternative="less", p=0.5)

print("El p-valor es casi 0 , son racistas
      los datos apoyan la hipótesis alternativa")

print("la propabilidad de la selección de un afroamericano es menor a 11%")
```

```
        1-sample proportions test with continuity correction

data:  4 out of 80, null probability 0.5
X-squared = 63.013, df = 1, p-value = 1.027e-15
alternative hypothesis: true p is less than 0.5
95 percent confidence interval:
 0.0000000 0.1148949
sample estimates:
   p
0.05
```

Ejercicio: A/B testing

Una página web de venta de productos ha estudiado el número de conversiones de su página web actual (conversión = el usuario hace click en "comprar ahora"). Para aumentar el número de conversiones, rediseña el aspecto de su página web en base a *heatmaps*[1] La nueva página se prueba con un nuevo conjunto de usuarios, midiendo el número de conversiones. Datos en "ab_testing.csv". ¿Se puede concluir que la nueva página incrementa el número de conversiones? Apoya tus conclusiones con un gráfico.

COMPARACIÓN DE BINOMIALES

1) Hipótesis:

Ha: num_clicks_new > num_clicks_old

P_nueva > p_vieja

N: num de clicks de la página nueva ; B~(n_nueva , p_nueva)

V: num de clicks de la página antigua ; B~(n_vieja , p_vieja)
```

Tema 4: Intervalos de Confianza

Test e IC's para poblaciones no-normales

(Datos Tabulares)

Hola

```
library(readr)
ab_testing <- read_csv("data_4/data_4/ab_testing.csv")

ab_testing_nueva = ab_testing[ab_testing$page_design== "new", ]
ab_testing_viej = ab_testing[ab_testing$page_design== "old", ]

ttt=table(ab_testing)
exito_new = ttt[1,2]
exito_old= ttt[2,2]
vec_x = c(exito_new, exito_old)
prop.test(vec_x, rowSums(ttt), alternative = "greater")

print("la propabilidad de conversión en la página nueva es al menos
      un 3% superior a la página vieja (95%). Debemos cambiar el diseño")

print("dado que el pvalor es casi cero, acogemos la hipótesis alternativa")
```

```
        2-sample test for equality of proportions with continuity correction

data:  vec_x out of rowSums(ttt)
X-squared = 14.046, df = 1, p-value = 8.922e-05
alternative hypothesis: greater
95 percent confidence interval:
 0.03847633 1.00000000
sample estimates:
   prop 1    prop 2
0.2587917 0.1897233
```

Paloma Pérez De Madrid

Inés Lorenzo Lallana De la Fuente Garzón