

# **Clase 13 - Inferencia estadística**

## **Diplomado en Análisis de Datos con R e Investigación reproducible para Biociencias.**

Dr. José Gallardo Matus | <https://genomics.pucv.cl/>

Pontificia Universidad Católica de Valparaíso

10 October 2022

# PLAN DE LA CLASE

## 1.- Introducción

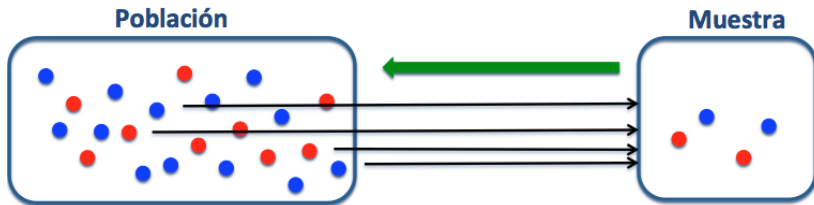
- ▶ ¿Qué es la inferencia estadística?
- ▶ ¿Cómo someter a prueba una hipótesis?
- ▶ Pruebas paramétricas
- ▶ Interpretar resultados de análisis de datos con R.

## 2.- Práctica con R y Rstudio cloud

- ▶ Someter a prueba diferentes hipótesis estadísticas.
- ▶ Realizar gráficas avanzadas con ggplot2.

# ¿QUÉ ES LA INFERENCIA ESTADÍSTICA?

**Inferencia estadística** : Son procedimientos que permiten obtener o extraer conclusiones sobre los parámetros de una población a partir de una muestra de datos tomada de ella.



¿Qué inferencia puede hacer de los datos de esta población?  
¿Qué ocurre si la muestra no es aleatoria?

# INFERENCIA ESTADÍSTICA

¿Par qué es importante la inferencia estadística?

- ▶ **Es más económico que hacer un Censo.**  
¿Cuántas especies hay en una bahía, en una laguna, en un bosque?
- ▶ **Bajo ciertos supuestos permite hacer afirmaciones.**  
Con fertilizante A las plantas crecen más que con fertilizante B.
- ▶ **Bajo ciertos supuestos permite hacer predicciones.**  
Mujeres con genotipos mutante del gen BCRA1 tiene 7 veces más probabilidad de tener cáncer ovárico que mujeres con genotipo normal.

# INFERENCIA ESTADÍSTICA: MÉTODOS

Los métodos de inferencia estadística que revisaremos en este curso son:

- 1. Estimación de parámetros a partir de una muestra.**
- 2. Pruebas de contraste de hipótesis.**
- 3. Modelamiento predictivo.**

# CONCEPTOS IMPORTANTES

## ► **Parámetro**

Constante que caracteriza a todos los elementos de un conjunto de datos de una población. Se representan con letras griegas.

Promedio de una población  $(\mu) = \mu$ .

## ► **Estadístico**

Una función de una muestra aleatoria o subconjunto de datos de una población.

Promedio de una muestra  $(\bar{X}) = \sum \frac{x_i}{n}$

# ESTIMACIÓN DE PARÁMETROS

**Objetivo** Hacer generalizaciones de una población a partir de una muestra.

## Tipos de estimación

- ▶ **Estimación puntual:** Consiste en asumir que el parámetro tiene el mismo valor que el estadístico en la muestra.
- ▶ **Estimación por intervalos:** Se asigna al parámetro un conjunto de posibles valores que están comprendidos en un intervalo asociado a una cierta probabilidad de ocurrencia.

## Ejemplo

A partir del muestreo de 30 individuos estimo que el nivel medio de cortisol en el plasma es de 15  $\mu\text{gramos/decilitro}$ , con un intervalo de confianza del 95% nuestro parámetro estará entre 10 y 20  $\mu\text{gramos/decilitro}$  95 de 100 veces.

# ERROR EN LA ESTIMACIÓN DE PARÁMETROS

## ¿Puedo estimar erroneamente un parámetro?

Por supuesto, los errores se producen por violar algunas premisas.

- ▶ **Las muestras deben tomarse de forma aleatoria.**

Si muestreo la diversidad de bacterias en una bahía justo en el efluente de una industria, esta muestra no representa la realidad de la bahía. Si los peces grandes son más fáciles de capturar que peces pequeños, la biomasa de una laguna será menor que la predicha.

- ▶ **Ley de los grandes números.**

¿Mis variables están correlacionadas?, ¿Cual estimación del parámetro es mejor?.

Compare experimento de 3 muestras v/s 300 muestras.



# ERROR EN LA ESTIMACIÓN DE PARÁMETROS 2

## ¿Puedo estimar erroneamente un parámetro?

### ▶ **Sesgo del investigador**

Ej. “ $H_0$  = el aditivo mejora el crecimiento”. Deseo aceptar la hipótesis, repito el ensayo hasta que por azar funciona. No considerando las veces que no funcionó.

### ▶ **Otros**

Errores.

Equipos descalibrados.

Fraude.

# PRUEBAS DE HIPÓTESIS

## Objetivo

Realizar una afirmación acerca del valor de un parámetro, usualmente contrastando con alguna hipótesis.

## Hipótesis estadísticas

*Hipótesis nula* ( $H_0$ ) es una afirmación, usualmente de igualdad.

*Hipótesis alternativa* ( $H_A$ ) es una afirmación que se deduce de la observación previa o de los antecedentes de literatura y que el investigador cree que es verdadera.

## Ejemplo

$H_0$ : El nivel medio de cortisol es  $= 15$  microgramos por decilitro.

$H_A$ : El nivel medio de cortisol es  $> 15$  microgramos por decilitro.

# ¿POR QUÉ DOS HIPÓTESIS?

- ▶ Las pruebas estadísticas tienen como propósito someter a prueba una hipótesis nula con la intención de **rechazarla** o **refutarla** (Falsacionismo de Karl Popper).
- ▶ Por lo tanto, los datos nos dirán si **existen o no** evidencias para **rechazar la hipótesis nula**.

## ¿Por qué no simplemente aceptar la hipótesis alternativa?

- ▶ Porque pueden existir otros fenómenos no conocidos o no considerados en nuestro experimento que posteriormente permitan a otro investigador rechazar nuestra hipótesis alternativa.

Lectura complementaria: El método científico según Juan José Ibáñez, 2006

# ETAPAS DE UNA PRUEBA DE HIPÓTESIS

Para cualquier prueba de hipótesis necesitas lo siguiente:

- ▶ Los ***Datos*** (1).
- ▶ Una ***hipótesis nula*** (2).
- ▶ Una ***prueba estadística*** (3) que se aplicará.
- ▶ El ***nivel de significancia*** (4) para rechazar la hipótesis.
- ▶ La ***distribución*** (5) de la **prueba estadística** respecto de la cual se evaluará la ***hipótesis nula*** con el estadístico que estimas de tus *datos*.

# DISTRIBUCIÓN DEL ESTIMADOR

- ▶ **Distribución muestral del estimador**

Dado que un estimador puntual ( $\bar{X}$ ) también es una variable aleatoria, entonces también tiene una distribución de probabilidad asociada.

- ▶ **¿Cómo distribuye?**

Si  $X \sim Normal(\mu_x, \sigma_x)$

Entonces el estimador de la media tiene  $\bar{X} \sim Normal(\mu_x, \frac{\sigma_x}{\sqrt{n}})$

- ▶ **¿Por qué es importante?**

Conocer la distribución de  $\bar{X}$  nos permitirá hacer pruebas de hipótesis.

# ¿CUÁNDO RECHAZAR $H_0$ ?

## Regla de decisión

Rechazo  $H_0$  cuando la evidencia observada es poco probable que ocurra bajo el supuesto de que la hipótesis sea verdadera.

Generalmente  $\alpha = 0,05$  o  $0,01$ .

Es decir, rechazamos cuando el valor del estadístico está en el 5% inferior de la función de distribución muestral.

## Corrección de Bonferroni para comparaciones múltiples

Pero a veces  $\alpha = 10^{-8}$

Ejemplo: Evaluó 50.000 genotipos diferentes para investigar cual está asociado a ser resistente al Coronavirus. Solo por azar 2.500 estarán asociados con  $P < 0,05$

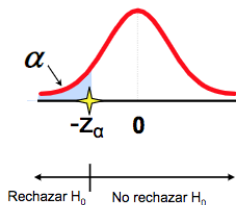
# PRUEBA DE HIPÓTESIS: UNA COLA O DOS COLAS

## Prueba unilateral izquierda

Ejemplo:

$$H_0: \mu \geq 3$$

$$H_A: \mu < 3$$

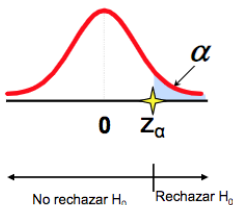


## Prueba unilateral derecha

Ejemplo:

$$H_0: \mu \leq 3$$

$$H_A: \mu > 3$$

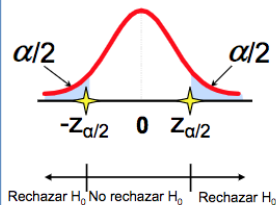


## Prueba bilateral

Ejemplo:

$$H_0: \mu = 3$$

$$H_A: \mu \neq 3$$



# ¿PUEDO COMETER UN ERROR EN LAS PRUEBAS DE HIPÓTESIS?

Por supuesto, siempre es posible llegar a una conclusión incorrecta.

## Tipos de errores

Tipo I ( $\alpha$ ) y tipo II ( $\beta$ ), ambos están inversamente relacionados.

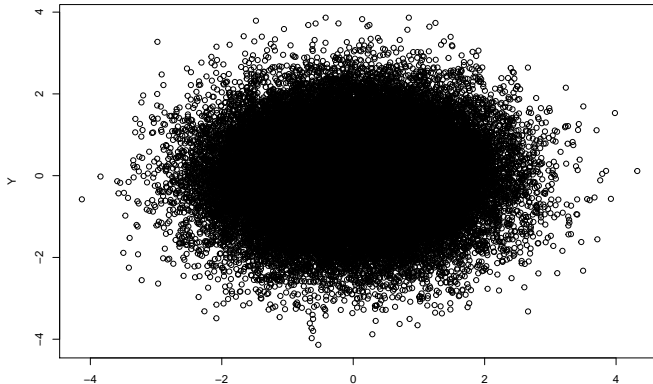
Decisión	$H_0$ es cierta	$H_0$ es falsa
<i>Aceptamos <math>H_0</math></i>	Decisión correcta	Error tipo II
<i>Rechazamos <math>H_0</math></i>	Error tipo I	Decisión correcta



# SIGNIFICANCIA ESTADÍSTICA v/s PRÁCTICA 2

**Problema 1** Si aumento  $n$  siempre lograré rechazar la hipótesis nula, cada vez para diferencias más pequeñas. ¿Esto tiene significancia práctica?

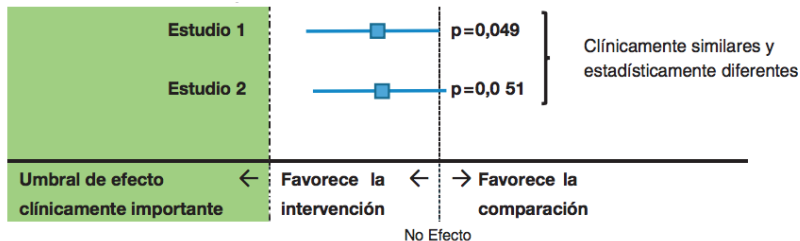
X e Y están significativamente correlacionados  $\rho = 0,01$  (p-value = 0.01901)



# SIGNIFICANCIA ESTADÍSTICA v/s PRÁCTICA 3

## Problema 2

Significancia basada en un punto de corte arbitrario. Abajo dos estudios son clínicamente similares, pero estadísticamente diferentes.

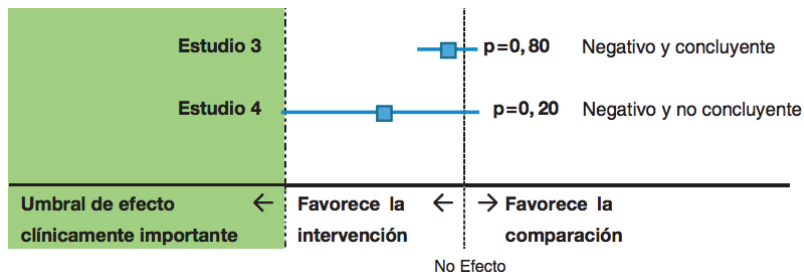


Lectura recomendada: Ciapponi, 2013

# SIGNIFICANCIA ESTADÍSTICA v/s PRÁCTICA 4

## Problema 3

Resultados “estadísticamente no significativos” pueden ser o no ser concluyentes.



# TIPOS DE PRUEBAS ESTADÍSTICAS

Según la forma de la distribución de la variable aleatoria.

## 1. Métodos paramétricos

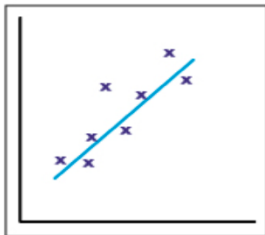
- ▶ Las pruebas de hipótesis usualmente asumen una distribución normal de la variable aleatoria.
- ▶ Útil para la mayoría de las variables cuantitativas continuas.

## 2. Métodos NO paramétricos

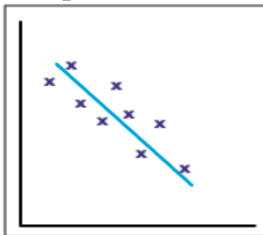
- ▶ Las pruebas de hipótesis no asumen una distribución normal de la variable aleatoria.
- ▶ Útil para todas las variables, incluyendo cuantitativas discretas y cualitativas.

# CORRELACIÓN ENTRE VARIABLES

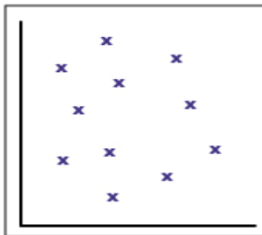
**Positive correlation**



**Negative correlation**



**No correlation**



# HIPÓTESIS PRUEBA DE CORRELACIÓN

## Hipótesis

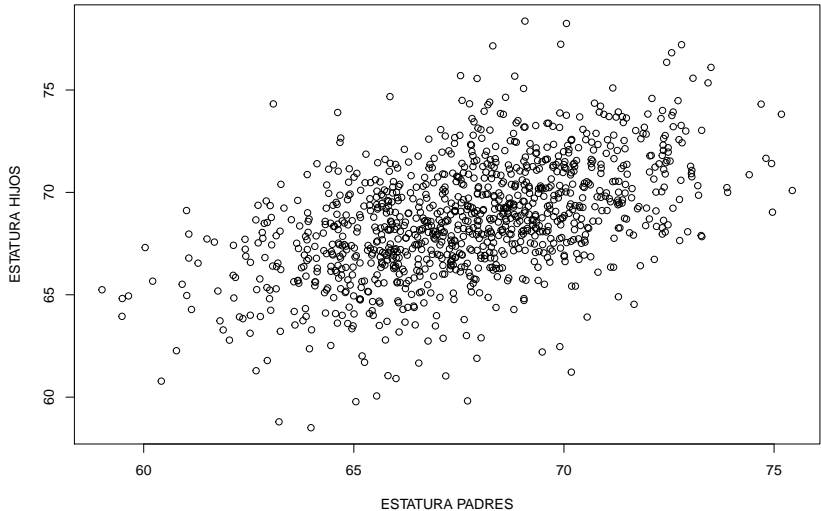
$H_0 : \rho = 0$  ausencia de correlación.

$H_1 : \rho \neq 0$  existencia de correlación.

## Supuestos:

- 1) Las variables X e Y son continuas y su relación es lineal.
- 2) La distribución conjunta de (X,Y) es una distribución Bivariable normal.

# ESTUDIO DE CASO: RELACIÓN ESTATURA PADRES - HIJOS



# PRUEBA DE CORRELACIÓN DE PEARSON

```
cor.test(father.son$fheight, father.son$sheight)
```

```
##
```

```
## Pearson's product-moment correlation
```

```
##
```

```
## data: father.son$fheight and father.son$sheight
```

```
## t = 19.006, df = 1076, p-value < 2.2e-16
```

```
## alternative hypothesis: true correlation is not equal to 0
```

```
## 95 percent confidence interval:
```

```
## 0.4552586 0.5447396
```

```
## sample estimates:
```

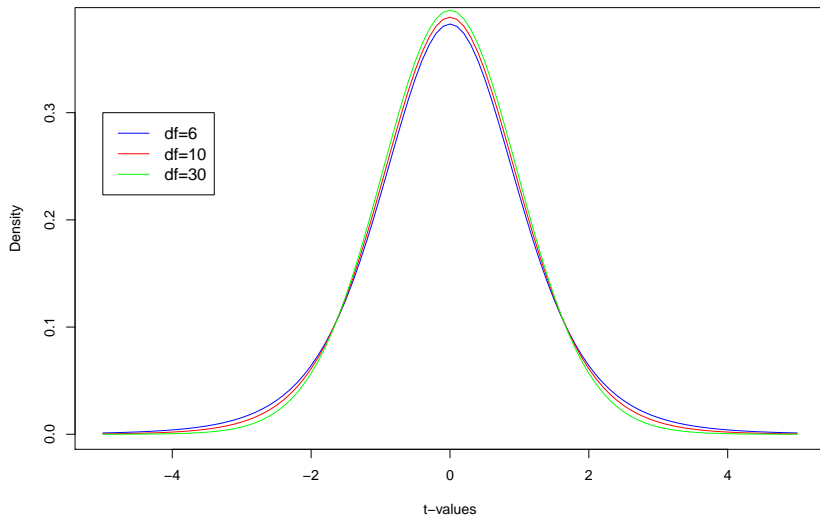
```
## cor
```

```
## 0.5013383
```



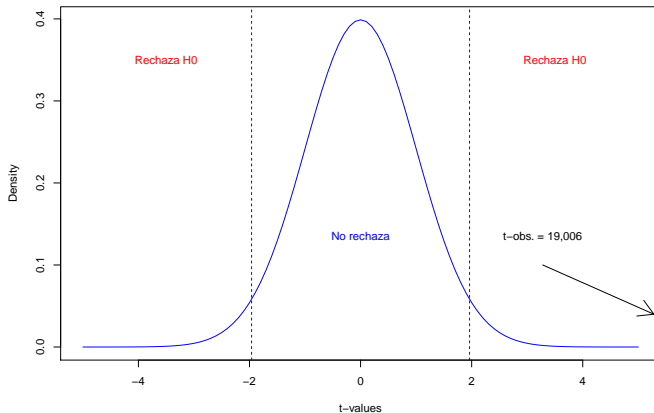
# DISTRIBUCIÓN T STUDENT

Origen: William Sealy Gosset, estadístico de la cervecería Guinness.

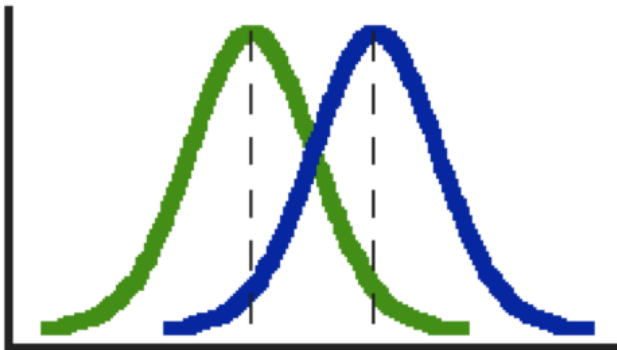


# PRUEBA DE HIPÓTESIS

- ▶  $gl \text{ correlación} = N^{\circ} \text{ datos} - 2 = 1078 - 2$
- ▶ Región de no rechazo =  $t\text{-student} (gl=1076) = -1.96 - 1.96$



# PRUEBA DE COMPARACIÓN DE MEDIAS



# HIPÓTESIS COMPARACIÓN DE MEDIAS

## Hipótesis

$$H_0 : \mu_1 = \mu_2.$$

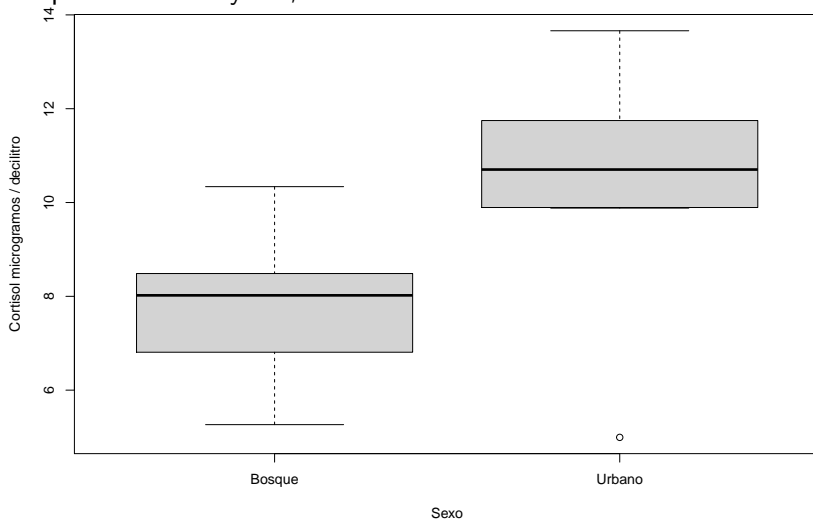
$$H_1 : \mu_1 \neq \mu_2$$

## Supuestos

- 1) Las variables X es continua.
- 2) Distribución normal.

# ESTUDIO DE CASO: CORTISOL

Adaptado de Kobayashi, et al 2019



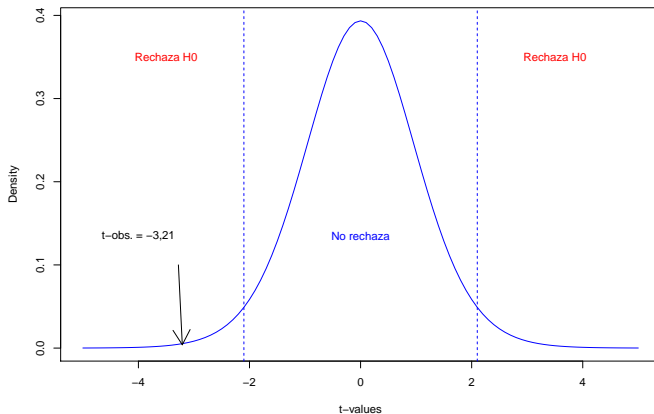
# PRUEBA DE T PARA DOS MUESTRAS INDEPENDIENTES

```
t.test(Cortisol ~ Zona, dat, alternative = c("two.sided"),  
       var.equal=TRUE)
```

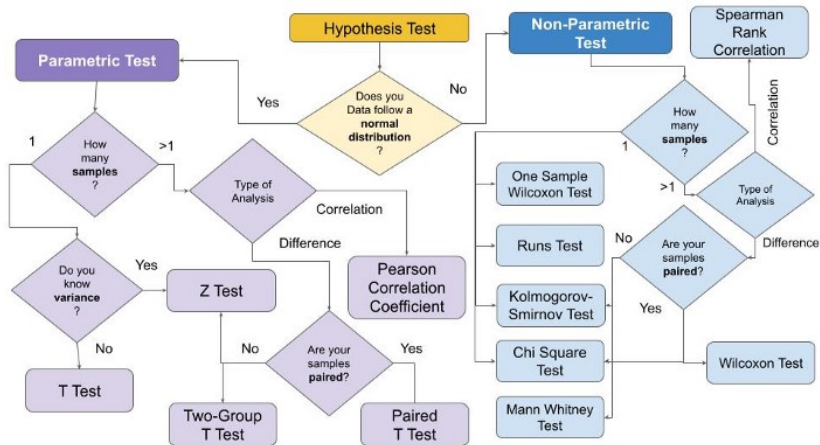
```
##  
## Two Sample t-test  
##  
## data: Cortisol by Zona  
## t = -3.2134, df = 18, p-value = 0.004818  
## alternative hypothesis: true difference in means is not  
## 95 percent confidence interval:  
## -4.5099156 -0.9440824  
## sample estimates:  
## mean in group Bosque mean in group Urbano  
## 7.805148 10.532147
```

# PRUEBA DE HIPÓTESIS

- ▶  $gl \text{ correlación} = N^{\circ} \text{ datos} - 2 = 20 - 2$
- ▶ Región de no rechazo distribución t-student ( $gl=18$ ) =  $-2.1 - 2.1$



# PRÁCTICA ANÁLISIS DE DATOS





# RESUMEN DE LA CLASE

## 1. Conceptos básicos de inferencia estadística

- ▶ Estadístico y parámetro.

## 2. Conceptos básicos de pruebas de hipótesis

- ▶ Hipótesis nula, alternativa.

## 3. Distribución de probabilidad

- ▶ t-student.

## 4. Realizar pruebas de hipótesis

- ▶ Test de correlación.
- ▶ Test de comparación de medias para 2 muestras independientes.

## 5. Realizar gráficas avanzadas con ggplot2.