Clase 10 Inferencia estadística

Curso Introducción al Análisis de datos con R para la acuicultura.

Dr. José A. Gallardo | jose.gallardo@pucv.cl | Pontificia Universidad Católica de Valparaíso

11 July 2021

PLAN DE LA CLASE

1.- Introducción

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

2.- Práctica con R y Rstudio cloud

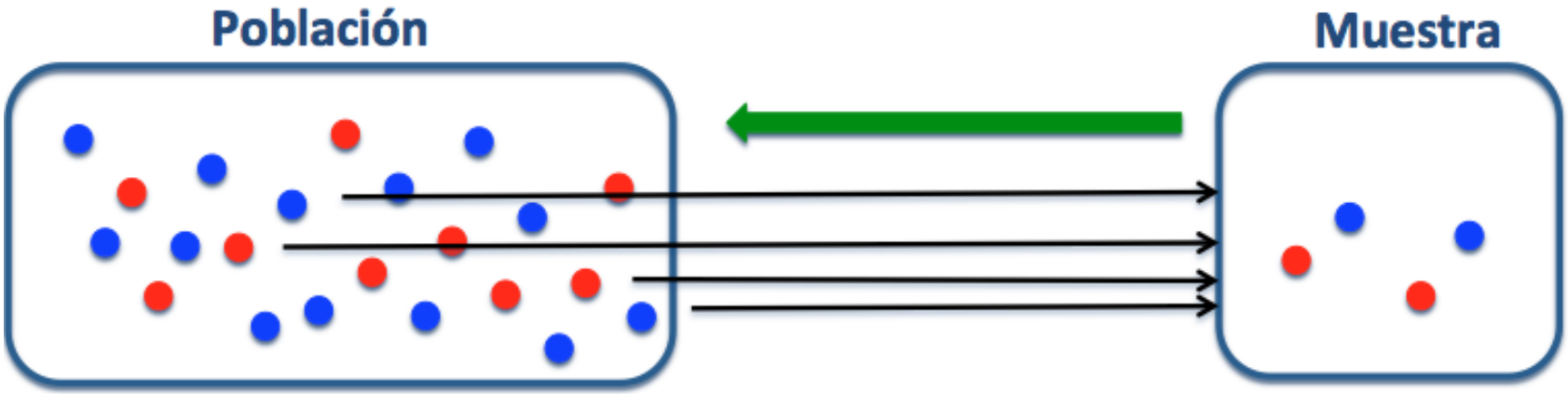
- Realizar pruebas de hipótesis: Correlación, comparación de medias (2 muestras independientes y pareadas).
- Realizar gráficas avanzadas con ggplot2.
- Elaborar un reporte dinámico en formato pdf.

INFERENCIA ESTADÍSTICA

¿Qué es la inferencia estadística?

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

¿Qué inferencia puede hacer de este experimento?



INFERENCIA ESTADÍSTICA 2

¿Para qué es útil?

- **Es más económico que hacer un Censo.**
  ¿Cuál es la biomasa en un estanque?
  ¿Cuántas larvas tengo para sembrar?
- **Bajo ciertos supuestos permite hacer afirmaciones.**
  Si alimento mis camarones con la dieta A crecerán más que con la dieta B.
  La eficacia de la vacuna es menor cuando los peces sufren una coinfección de patógenos.

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 UN PARÁMETRO

Objetivo Estimar parámetros de la población a partir de la muestra de una variable aleatoria.

Ejemplo Estimar el promedio del peso del cuerpo de una población a partir de una muestra de 30 animales.

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.

¿PUEDO ESTIMAR ERRONEAMENTE UN PARÁMETRO?

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

- **Las muestras deben tomarse de forma aleatoria.**
  No descartar animales pequeños en un muestreo (sobreestimo biomasa).
- **Ley de los grandes números.**
  Mis variables están correlacionadas (3 muestras v/s 300 muestras).
  La biomasa del estanque es 100 kg (10 peces v/s 100 peces).
- **Evitar sesgo del investigador**
  Deseo rechazar la hipótesis, repito hasta que rechazo.
- **Otros**
  Errores y fraude.

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.
- **¿Cómo distribuye?**
  Si X $\sim$ *Normal*(μ_x,σ_x)

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

- **¿Por qué es importante?**
  Conocer la distribución de $\bar{X}$ nos permitirá hacer pruebas de hipótesis.

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
$\mathbf{H_0}$: El peso medio de mis peces es igual a 1 Kg.
$\mathbf{H_A}$: El peso medio de mis peces es mayor a 1 Kg.

¿POR QUÉ DOS HIPÓTESIS?

- Las pruebas estadísticas tienen como propósito someter a prueba una hipótesis nula con la intensión de *rechazarla*.
- ¿Por qué no simplemente aceptar la alternativa?

  We cannot conclusively affirm a hypothesis, but we can conclusively negate it Karl Popper

- Pueden existir otros fenómenos no conocidos o no considerados que posteriormente permitan a otro investigador rechazar nuestra hipótesis alternativa.
- Por lo tanto, los datos nos dirán si **existen o no** evidencias para rechazar la hipótesis nula.

ETAPAS DE UNA PRUEBA DE HIPÓTESIS

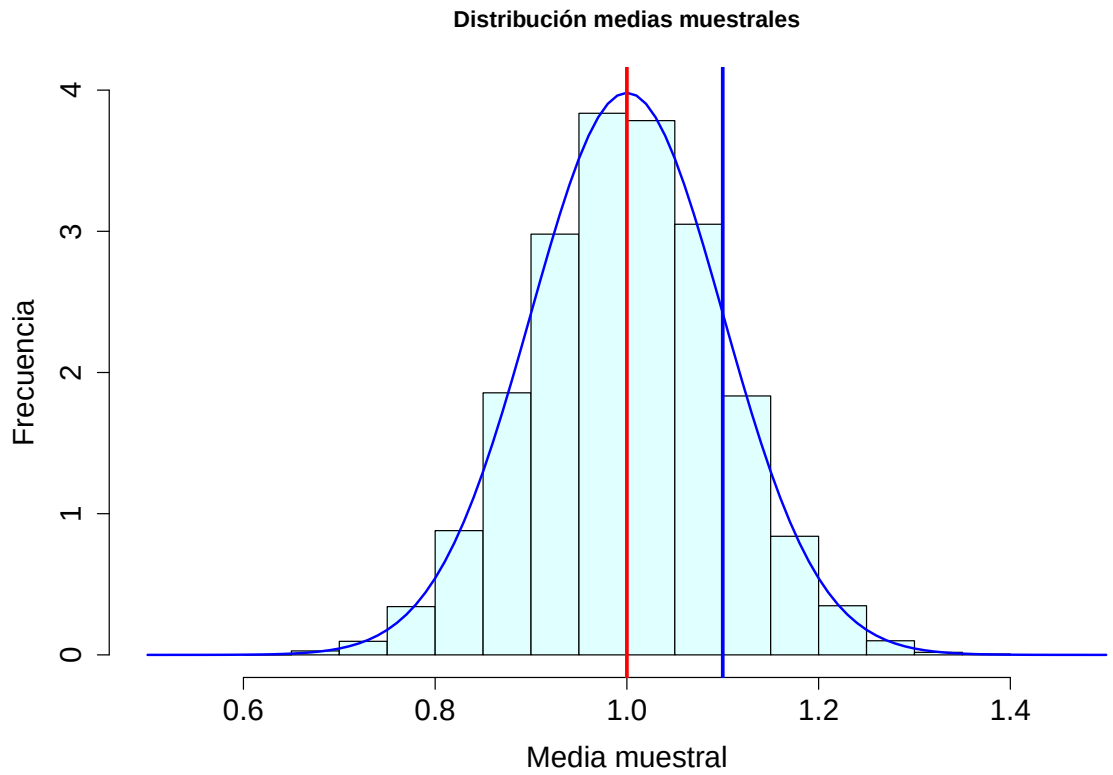
Para cualquier prueba de hipotesis necesitas lo siguiente:

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

PRUEBA DE HIPÓTESIS: NO RECHAZO.

H_0: El peso medio de sus peces es igual a 1 Kg.

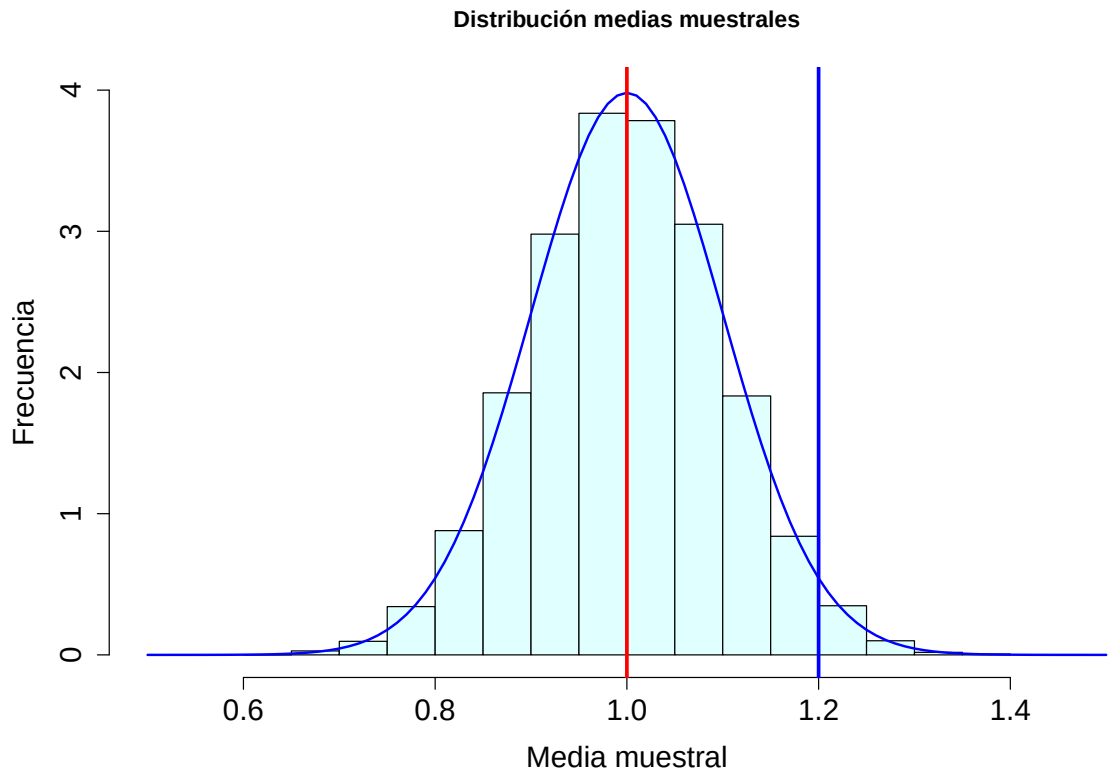
Si $\bar{X} = 1{,}1$ Kg, rechaza la hipótesis?



PRUEBA DE HIPÓTESIS: RECHAZO.

H_0: El peso medio de sus peces es igual a 1 Kg.

Si $\bar{X} = 1{,}2$ Kg, rechaza la hipótesis?



¿CUÁNDO RECHAZAR $\mathbf{H_0}$?

Regla de deisión
Rechazo $\mathbf{H_0}$ cuando la evidencia observada es poco probable que ocurra bajo el supuesto de que la hipótesis sea verdadera.

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

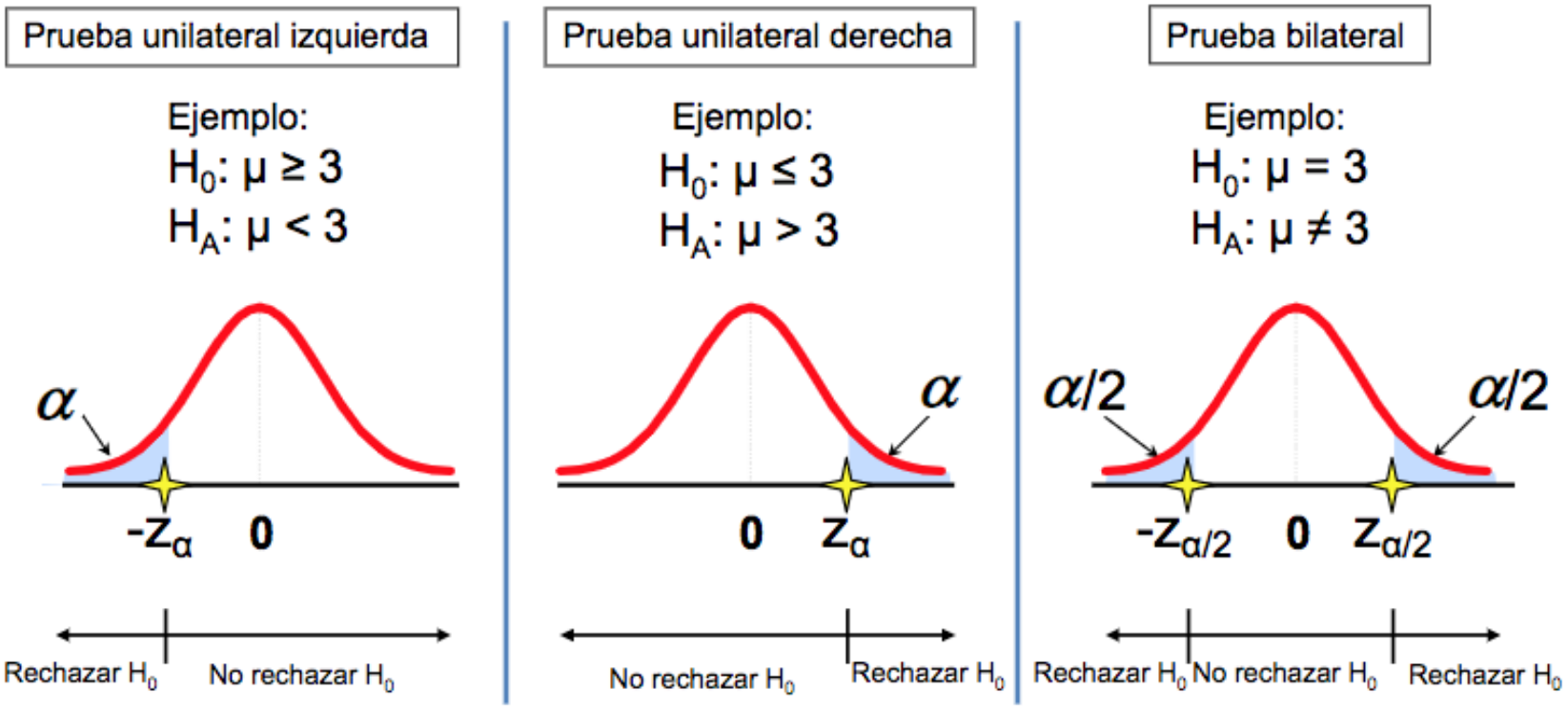
Es decir, rechazamos cuando el valor del estadístico está en el 5% inferior de la función de distribución muestral.

Corrección de Bonferroni comparaciones múltiples

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

Ejemplo: 50.000 genotipos asociados a un fenotipo. Solo por azar 2.500 estarán asociados con $P < 0{,}05$

PRUEBA DE HIPÓTESIS: UNA COLA O DOS COLAS



¿PUEDO COMETER UN ERROR EN LAS PRUEBAS DE HIPÓTESIS?

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

Tipos de errores

Tipo I (α) y tipo II (β), ambos están inversamente relacionados.

Decisión	**H_0 es cierta**	**H_0 es falsa**
Aceptamos H_0	Decisión correcta	Error tipo II
Rechazamos H_0	Error tipo I	Decisión correcta

SIGNIFICANCIA ESTADÍSTICA v/s PRÁCTICA

- **Problema 1**
  La vacuna aumenta significativamente el número de anticuerpos.

Sin vacuna = 10 anticuerpos Con vacuna = 11 anticuerpos (10 % de mejora).

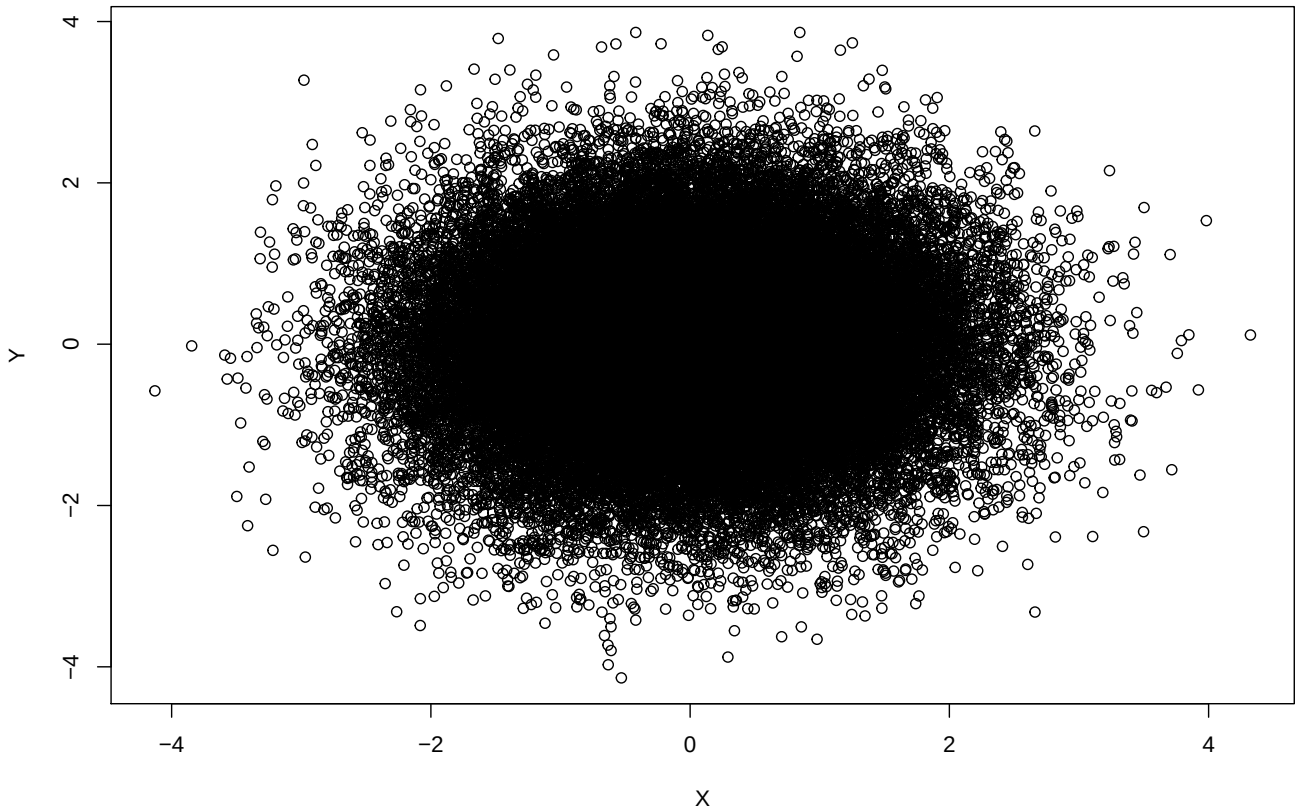
¿Cuál es la importancia práctica de este hallazgo?

¿Mejorará la salud de mis peces?

SIGNIFICANCIA ESTADÍSTICA v/s PRÁCTICA 2

Problema 2 Si aumento N siempre lograré rechazar la hipótesis nula, cada vez para diferencias más pequeñas. ¿Esto tiene significancia práctica?

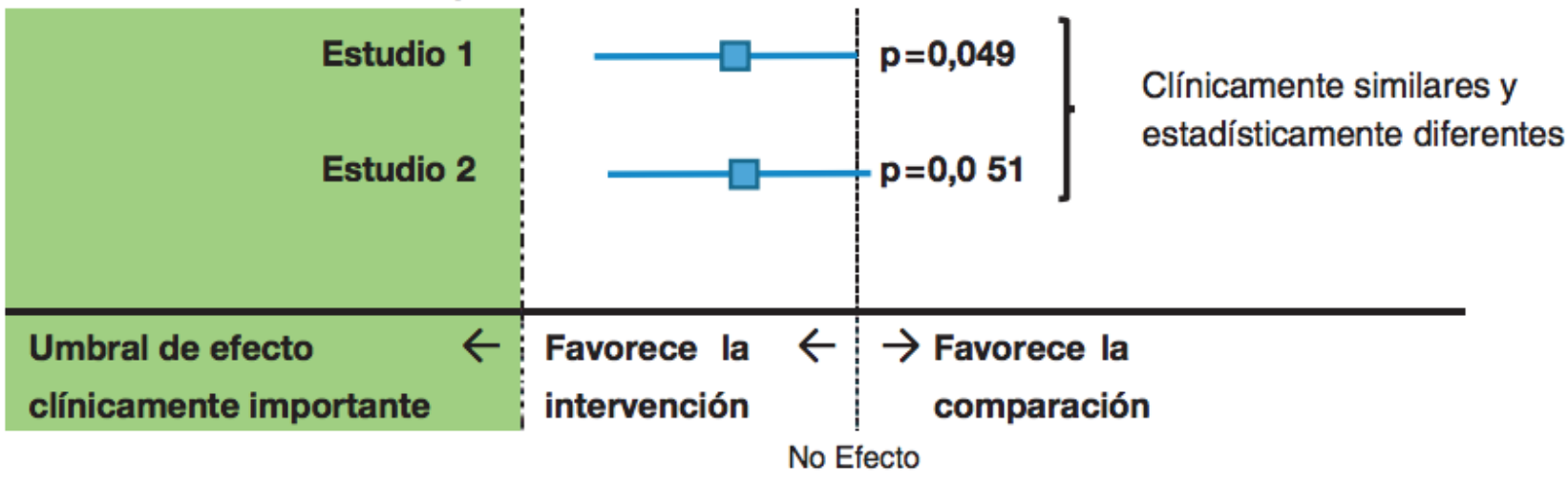
X e Y están significativamente correlacionados r = 0,01 (p-value = 0.01901)



SIGNIFICANCIA ESTADÍSTICA v/s PRÁCTICA 3

Problema 3

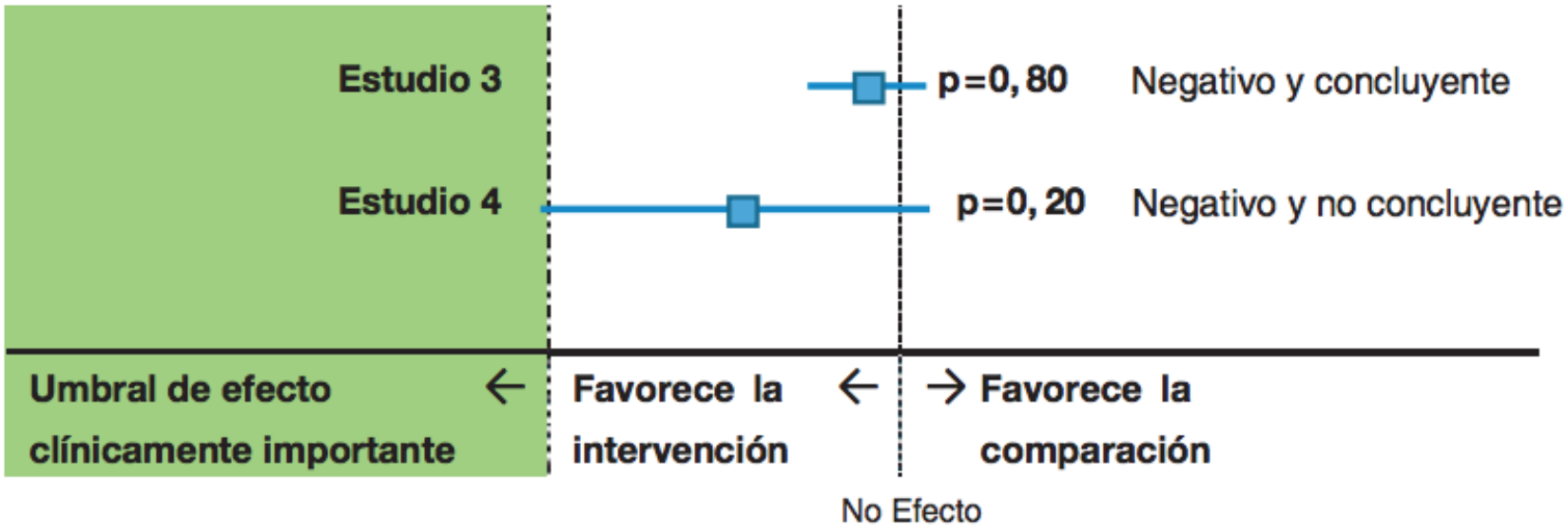
Clasificación basada en un punto de corte arbitrario.



SIGNIFICANCIA ESTADÍSTICA v/s PRÁCTICA 4

Problema 4

Resultados "estadísticamente no significativos" pueden ser o no ser concluyentes.



TÍPOS DE PRUEBAS ESTADÍSTICAS

Según la forma de la distribución de la variable aleatoria.

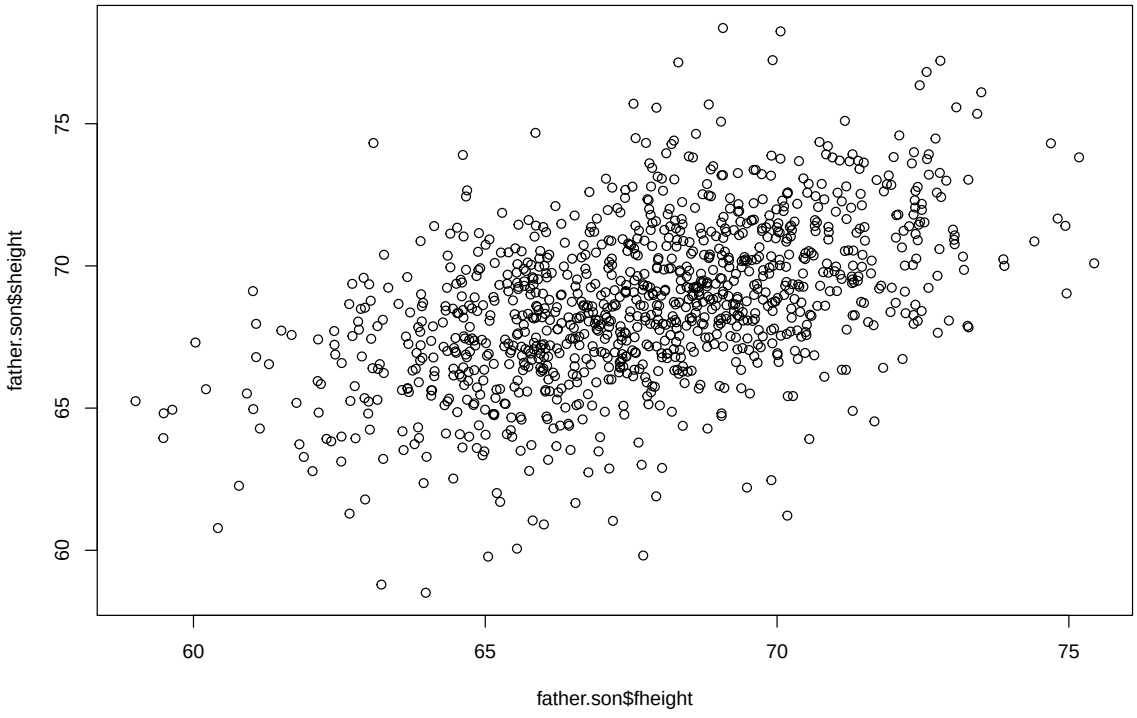
- **Métodos paramétricos** Las pruebas de hipótesis usualmente asumen una distribución normal de la variable aleatoria.

Util para la mayoría de las variables cuantitativas continuas.

- **Métodos NO paramétricos** Las pruebas de hipótesis no asumen una distribución normal de la variable aleatoria.

Util para todas las variables, incluyendo cuantitativas discretas y cualitativas.

COEFICIENTE CORRELACIÓN DE PEARSON



Hipótesis H0 : r = 0 ausencia de correlación H1 : r $\neq$ 0 existencia de correlación

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

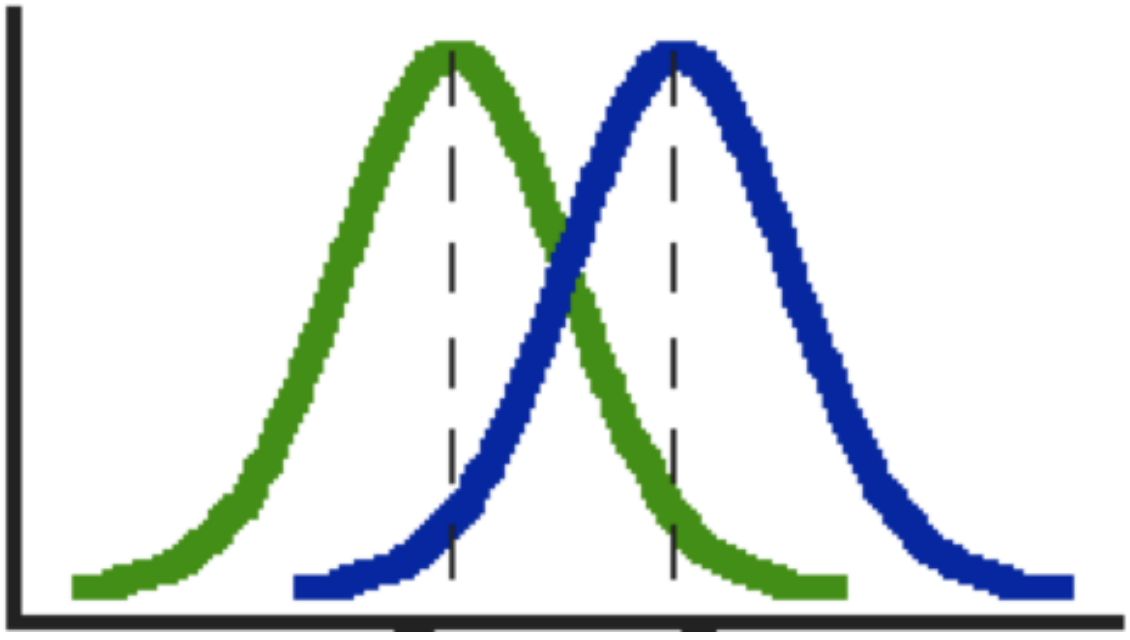
R Documentation cor.test {stats}

```
cor.test((x, y,
     alternative = c("two.sided", "less", "greater"),
     method = c("pearson", "kendall", "spearman"),
     conf.level* = 0.95, ...)
```

Pearson's product-moment correlation

data: X - Y
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

PRUEBA DE COMPARACIÓN DE MEDIAS



Hipótesis H0 : $\mu_1 = \mu_2$ H1 : μ_1 r $\neq \mu_2$

Supuestos: 1) Las variables X es continua. 2) Distribución normal.

R Documentation Student's t.test {stats}

t.test(x, y = NULL,
alternative = c("two.sided","less","greater"),
paired = FALSE,
var.equal = FALSE,
conf.level = 0.95, . . .)

Two Sample t-test data: Peso by Sexo
t = -11.315, df = 18, p-value = 1.292e-09
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-41.18421 -28.28530
sample estimates:
mean in group Female 141.9380 mean in group Male 176.6727

PRÁCTICA ANÁLISIS DE DATOS

- Guía de trabajo práctico disponible en drive y Rstudio.cloud.
  Clase_10
- El trabajo práctico se realiza en Rstudio.cloud.
  10 Inferencia estadística

RESUMEN DE LA CLASE

- **Elaborar hipótesis**
- **Realizar pruebas de hipótesis**
  - Test de correlación.
  - Test de comparación de medias para 2 muestras independientes.
  - Test para muestras pareadas.
- **Realizar gráficas avanzadas con ggplot2**