Clase 08 Inferencia estadística

OCE 386 - Introducción al análisis de datos con R.

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

26 October 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

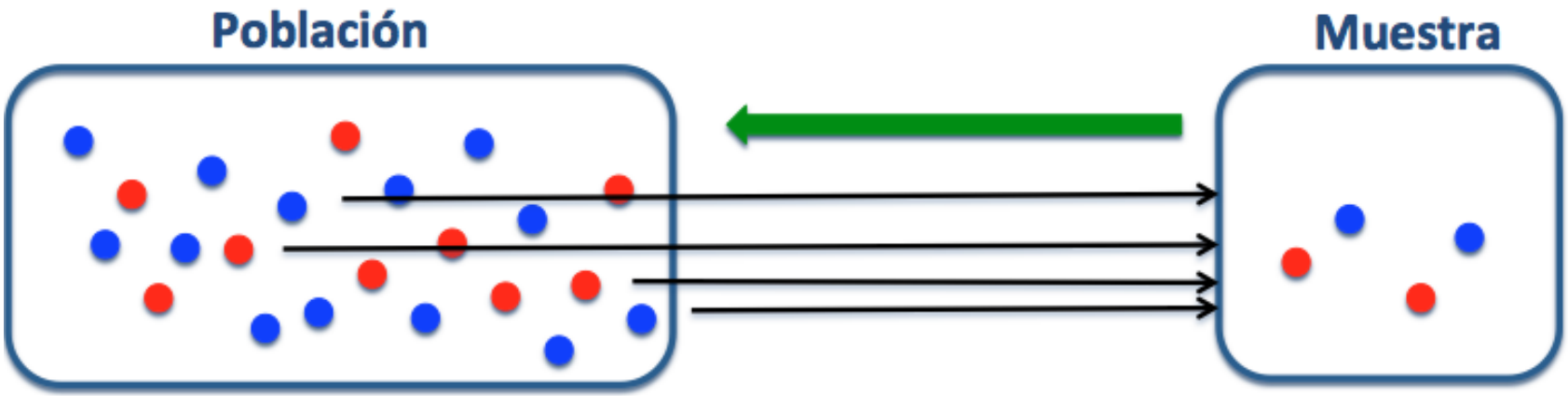
- Realizar pruebas de hipótesis: Correlación, comparación de medias (2 muestras independientes).
- Realizar gráficas avanzadas con ggplot2.
- Elaborar un reporte dinámico en formato pdf.

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 de ella.

¿Qué inferencia puedo hacer de este experimento?



INFERENCIA ESTADÍSTICA 2

¿Para qué es útil?

- **Es más económico que hacer un Censo.**
  ¿Cuántas larvas hay en la Bahía?
- **Bajo ciertos supuestos permite hacer afirmaciones.**
  La Bahía A está más contaminada que la Bahía B.

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) = μ.

- **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

¿Para qué? Estimar parámetros de la población a partir de la muestra de una variable aleatoria.

Ejemplo Medir la temperatura diaria de la superficie del mar en la Bahía de Valparaíso durante un año permite estimar la temperatura media por mes.

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 ERRÓNEAMENTE UN PARÁMETRO?

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

- **Las muestras deben tomarse de forma aleatoria.**
  Medir la temperatura de la bahía solo en la orilla de la Playa San Mateo no permite estimar adecuadamente la temperatura de la bahía.
- **Ley de los grandes números.**
  Mis variables están correlacionadas: Comparar 3 muestras v/s comparar 300 muestras.
- **Evitar sesgo del investigador**
  Deseo rechazar la hipótesis "la bahía no está contaminada", hago un muestreo cerca del efluente de una industria hasta que encuentro contaminación.
- **Otros**
  Equipos descalibrados, 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 asociada.

- **¿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 los peces es menor o igual a 1 Kg.
$\mathbf{H_A}$: El peso medio de los peces es mayor a 1 Kg.

¿POR QUÉ DOS HIPÓTESIS?

- Las pruebas estadísticas tienen como propósito someter a prueba una hipótesis nula con la intención 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

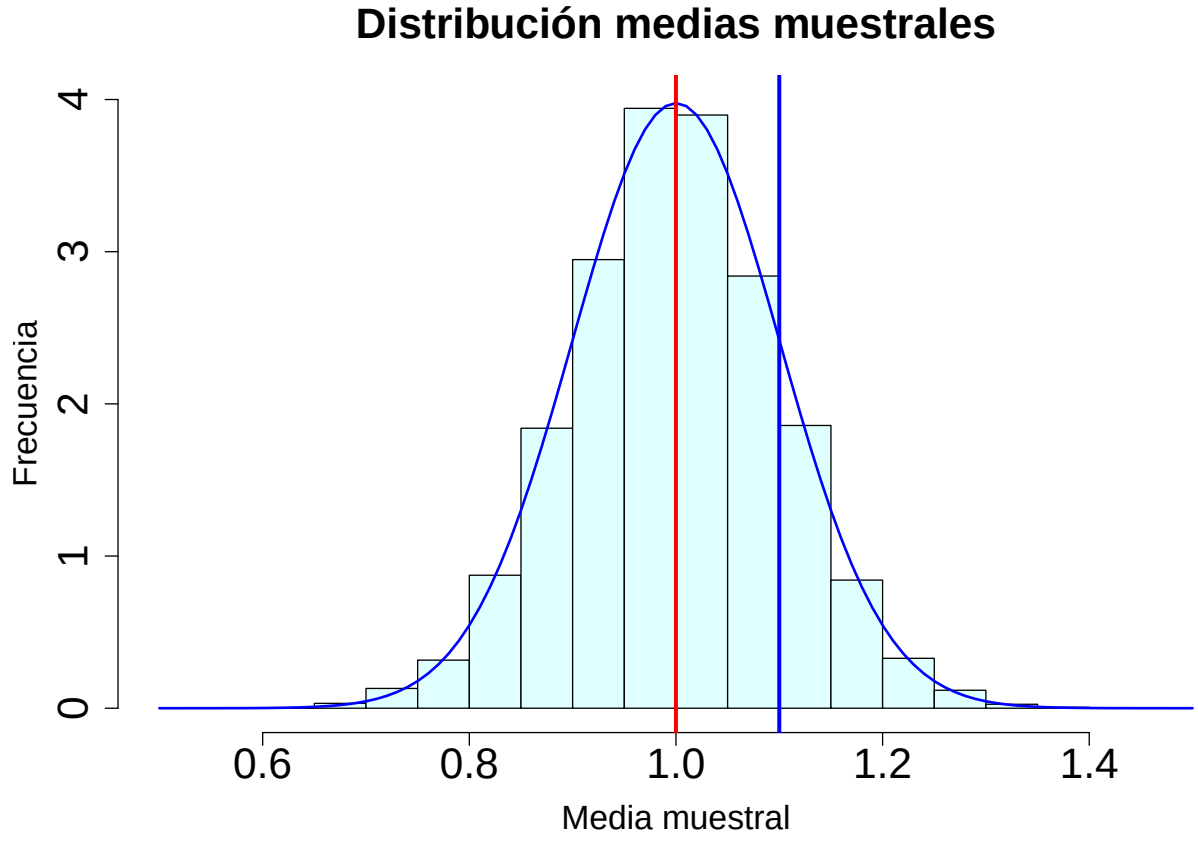
Para cualquier prueba de hipotesis necesitas lo siguiente:

- Tus ***datos*** (1).
- Una ***hipótesis nula*** (2).
- La ***prueba estadística*** (3) que se aplicará.
- El ***nivel de 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*.

PRUEBA DE HIPÓTESIS: NO RECHAZO.

H_0: El peso medio de los peces es menor o igual a 1 Kg.

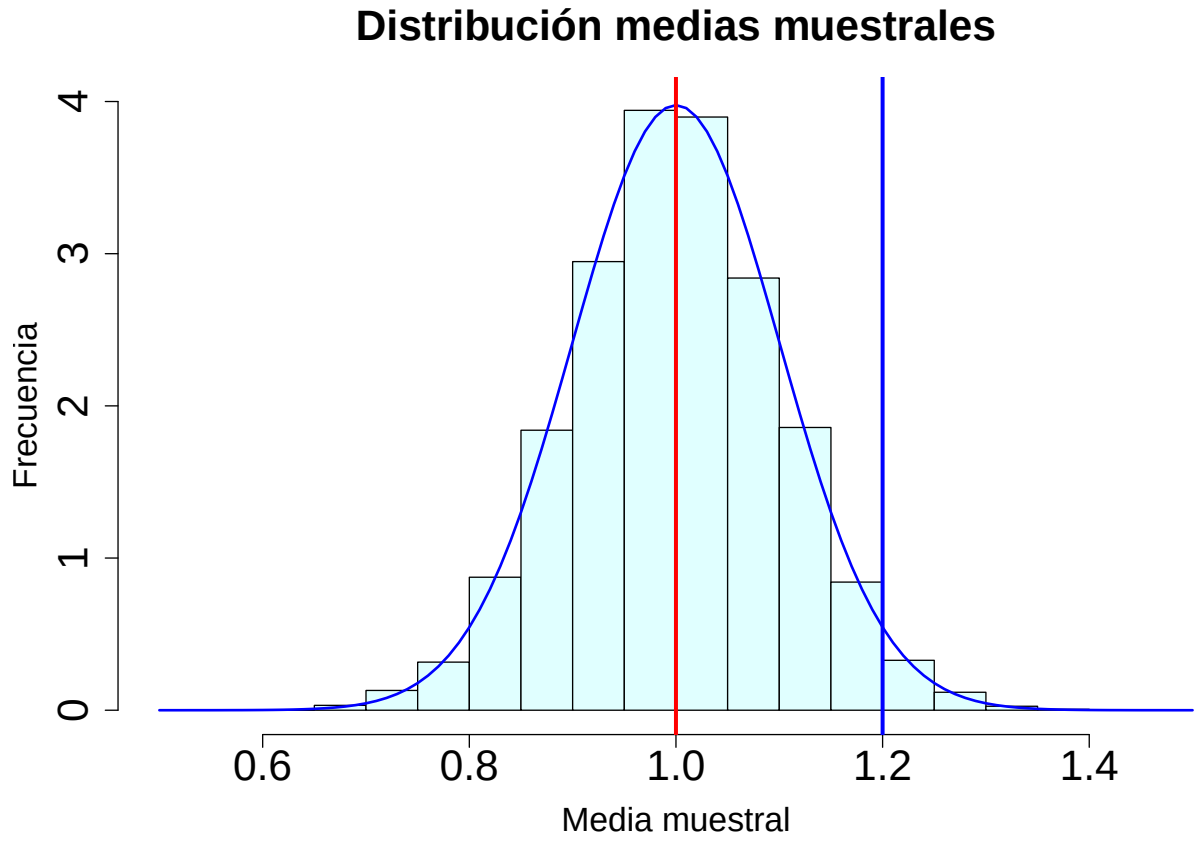
Si $\bar{X} = 1{,}1$ Kg, rechaza la hipótesis?



PRUEBA DE HIPÓTESIS: RECHAZO.

H_0: El peso medio de los peces es menor o igual a 1 Kg.

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



¿CUÁNDO RECHAZAR $\mathbf{H_0}$?

Regla de decisión
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 $0{,}01$.

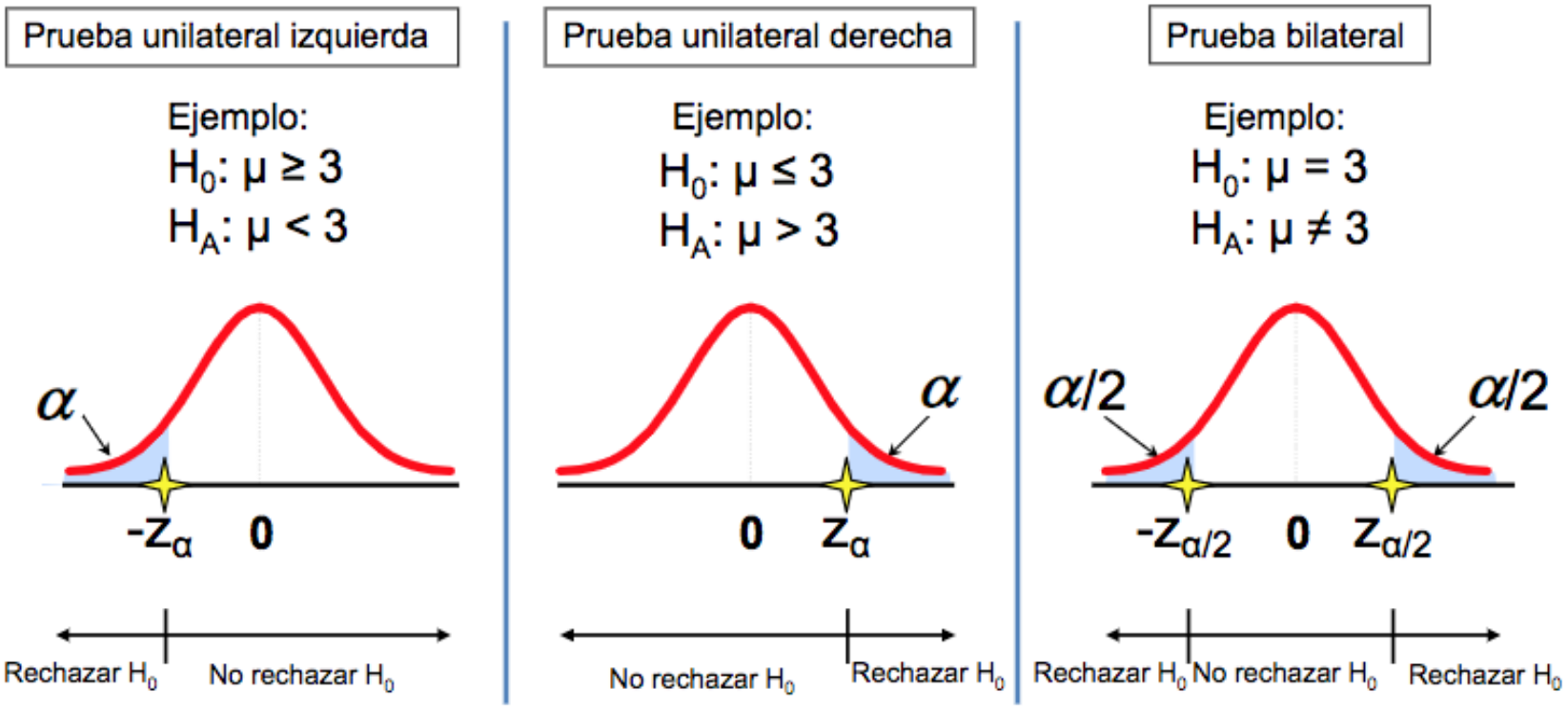
Es decir, rechazamos cuando el valor del estadístico está en el 5% inferior de la función de distribución muestral.

Corrección de Bonferroni comparaciones múltiples

Pero a veces α debe ser menor que 0,01 (ej. 10^{-8})

Ejemplo: Correlación entre 20 variables del sedimento (180 correlaciones posibles). Solo por azar 5% (9) 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
Nuevo filtro disminuye significativamente el número de Coliformes fecales vertidos al río.

Sin filtro = 100 Coliformes fecales.
Con filtro = 90 Coliformes fecales (10 % de mejora).

¿Cuál es la importancia práctica de este hallazgo?

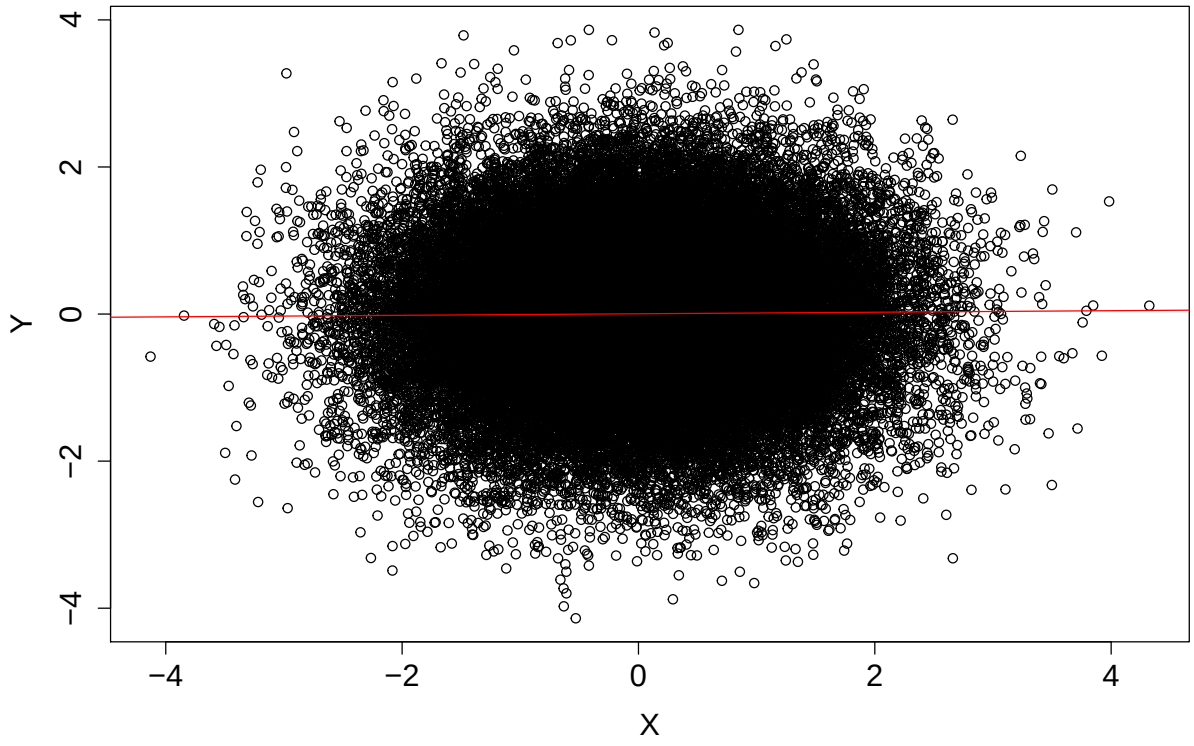
¿Mejorará la salud del Río?

SIGNIFICANCIA ESTADÍSTICA v/s PRÁCTICA 2

Problema 2

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

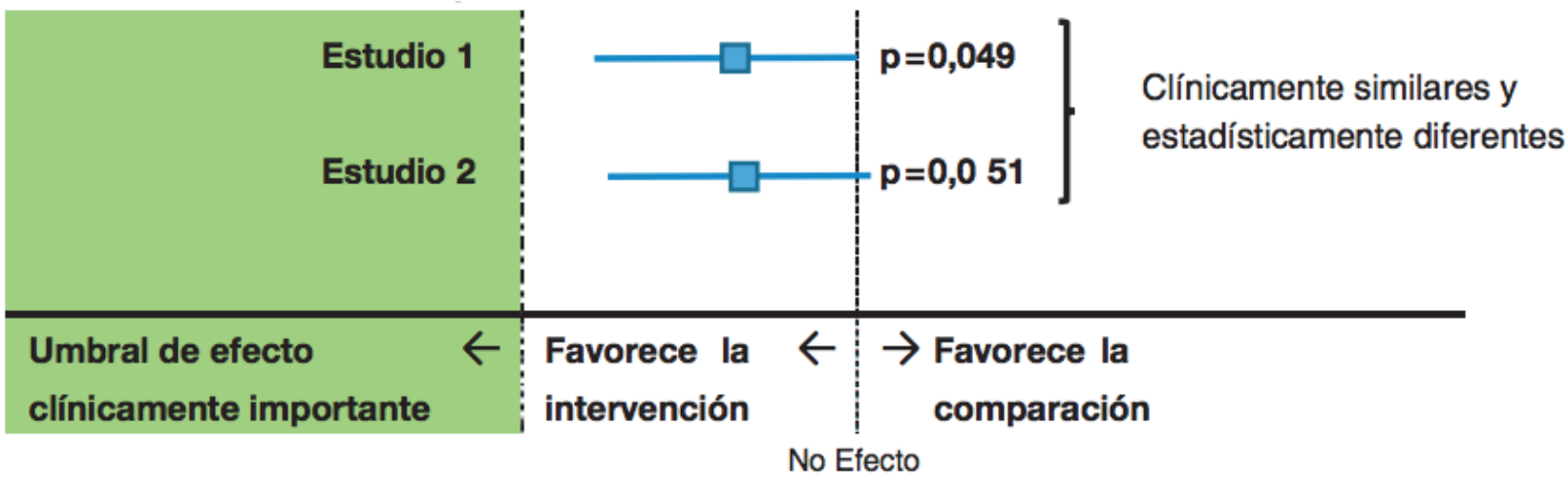
X e Y están significativamente correlacionados $\rho = 0{,}01$ (p-value = 0.01901)



SIGNIFICANCIA ESTADÍSTICA v/s PRÁCTICA 3

Problema 3

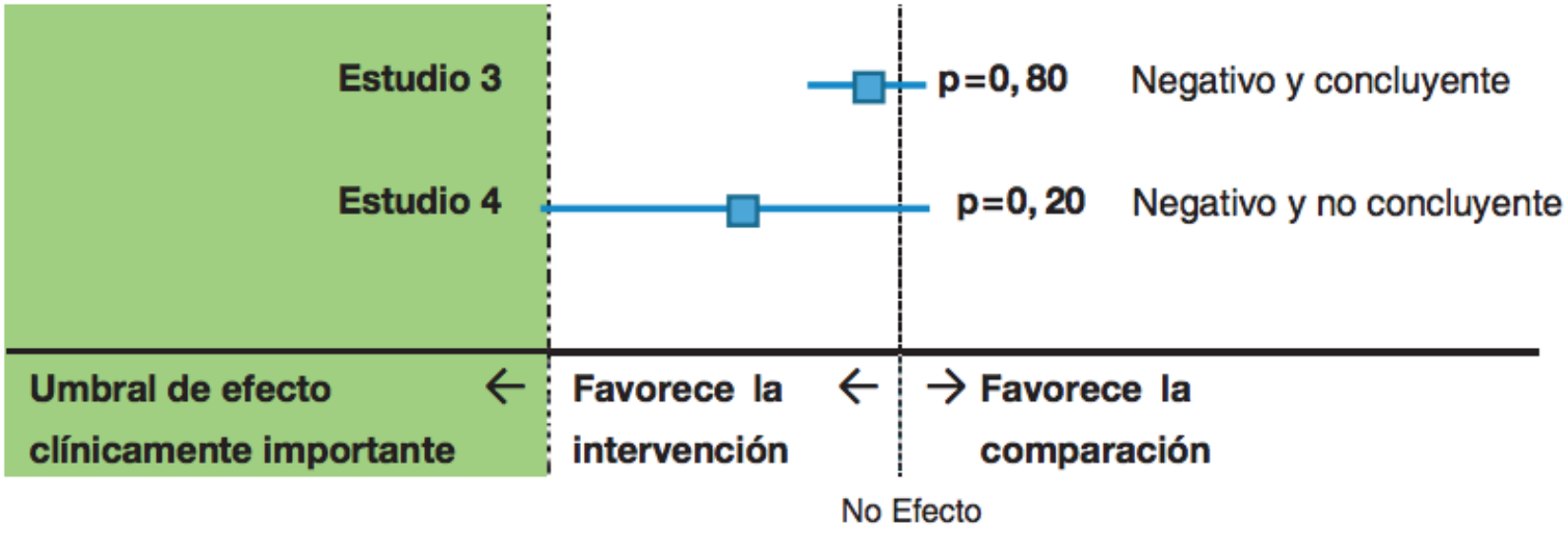
Clasificación basada en un punto de corte arbitrario.



SIGNIFICANCIA ESTADÍSTICA v/s PRÁCTICA 4

Problema 4

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



TIPOS DE PRUEBAS ESTADÍSTICAS

Según la forma de la distribución de la variable aleatoria.

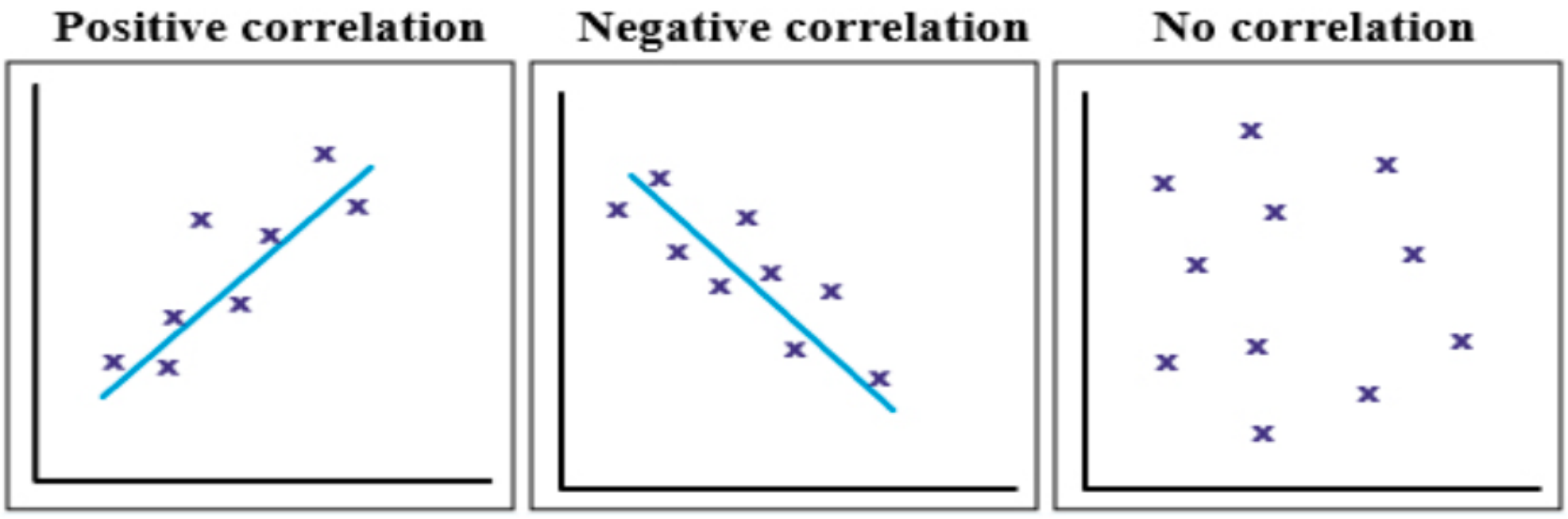
- **Métodos paramétricos** Las pruebas de hipótesis usualmente asumen una distribución normal de la variable aleatoria.

Útil para la mayoría de las variables cuantitativas continuas.

- **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.

COEFICIENTE CORRELACIÓN DE PEARSON

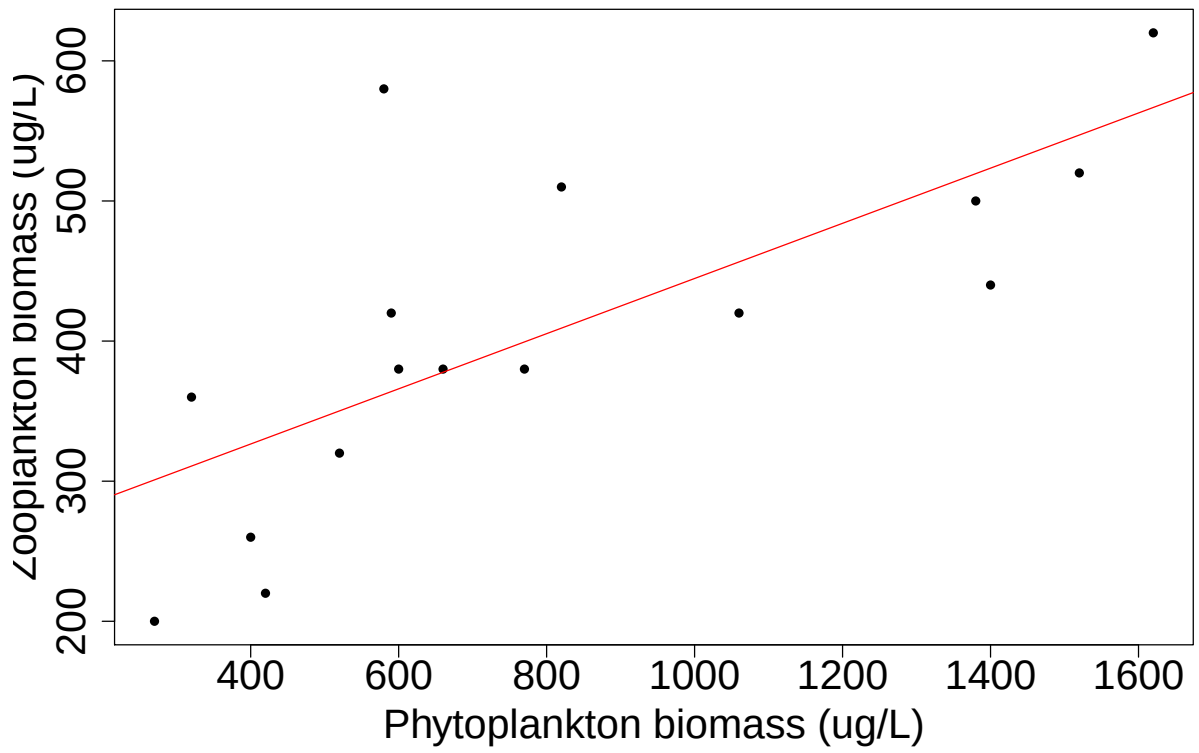


Hipótesis $H_0 : \rho = 0$ ausencia de correlación ***vs***. $H_1 : \rho \neq 0$ existencia de correlación.

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.

Estudio de caso: Relación entre biomasa de fito y zooplancton

Datos simulados desde estudio Lago Balaton de Europa (Hidrobiología, 2020)

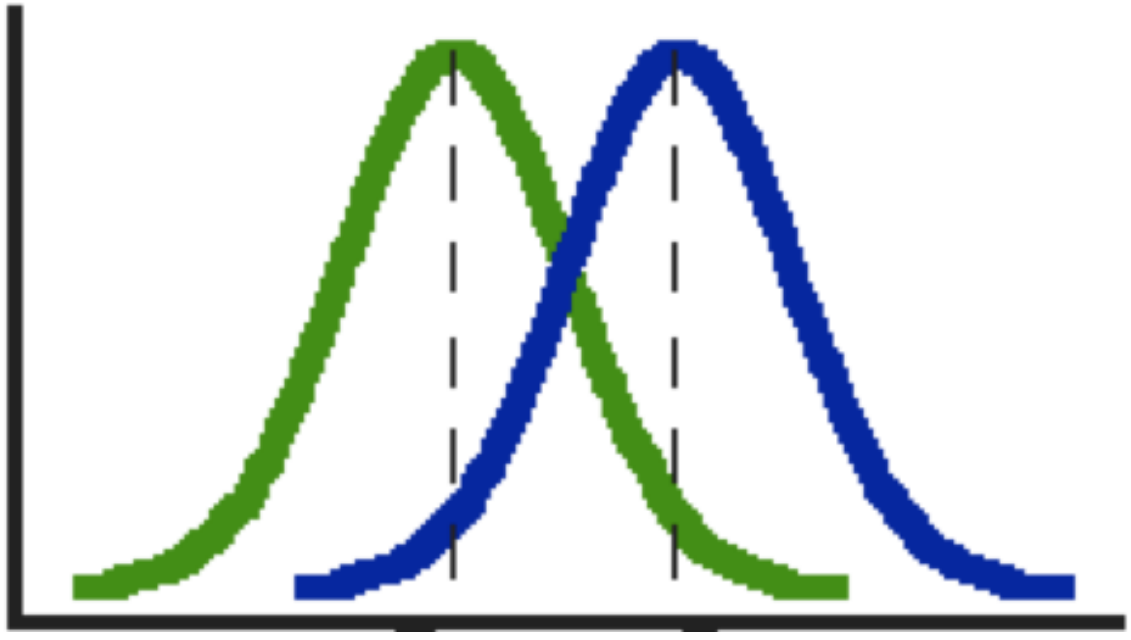


Pearson's product-moment correlation

```
cor.test(lake$Fito,lake$Zoo, method="pearson",
         alternative = "two.sided")
```

```
##
##  Pearson's product-moment correlation
##
## data:  lake$Fito and lake$Zoo
## t = 3.9612, df = 14, p-value = 0.00142
## alternative hypothesis: true correlation is not equal to
## 95 percent confidence interval:
##  0.3615508 0.8987850
## sample estimates:
##       cor
## 0.7269671
```

PRUEBA DE COMPARACIÓN DE MEDIAS



Hipótesis $H_0 : \mu_1 = \mu_2$ ***vs***. $H_1 : \mu_1 \neq \mu_2$

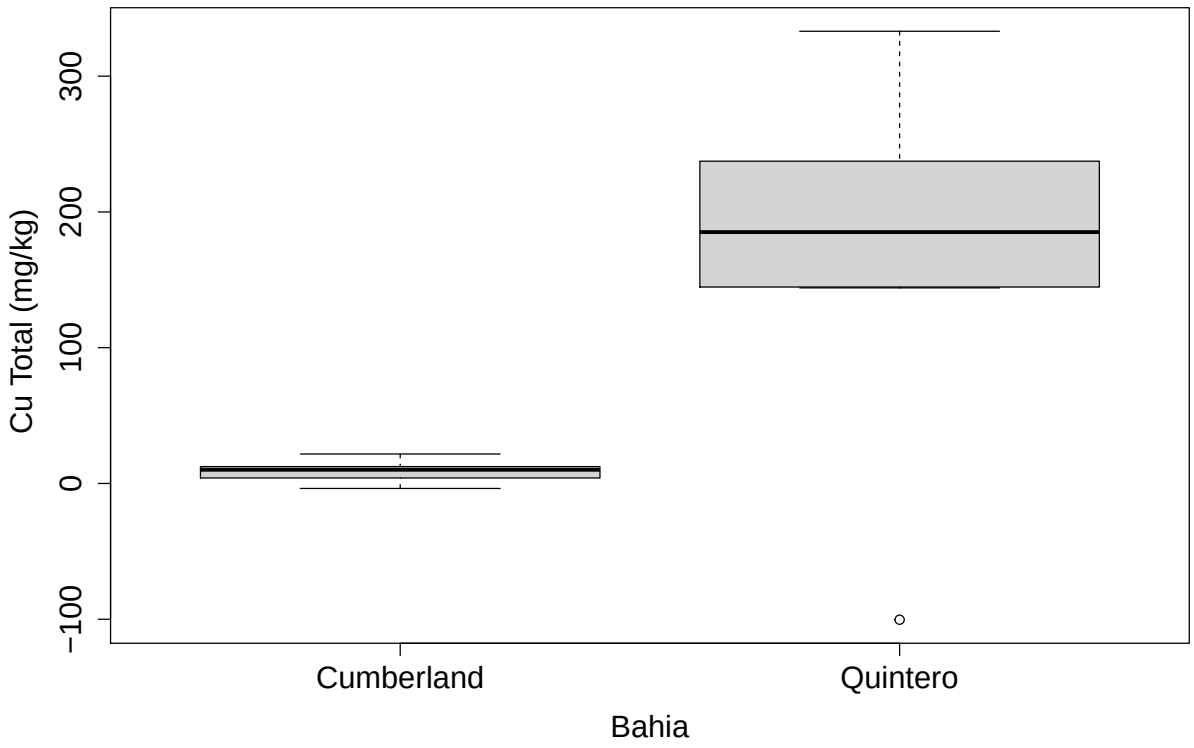
Supuestos:

1) La variable X es continua.

2) Distribución normal.

Estudio de caso: Contaminación en bahías Quintero y Cumberland.

Datos simulados desde estudio ambiental de Quintero (Directemar, 2019)



Two Sample t-test

```
t.test(Cobre ~ Bahia, bahia, alternative = c("two.sided"),
       var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  Cobre by Bahia
## t = -4.5919, df = 18, p-value = 0.0002262
## alternative hypothesis: true difference in means is not
## 95 percent confidence interval:
##  -244.25504  -90.90816
## sample estimates:
## mean in group Cumberland   mean in group Quintero
##                 9.025739                176.607336
```

PRÁCTICA ANÁLISIS DE DATOS

- Guía de trabajo práctico disponible en drive y Rstudio.cloud.
  Clase_8
- El trabajo práctico se realiza en Rstudio.cloud.
  8 Inferencia estadística

RESUMEN DE LA CLASE

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