Clase 05 - Inferencia estadística

Curso Análisis de datos con R para Biociencias

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

21 January 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

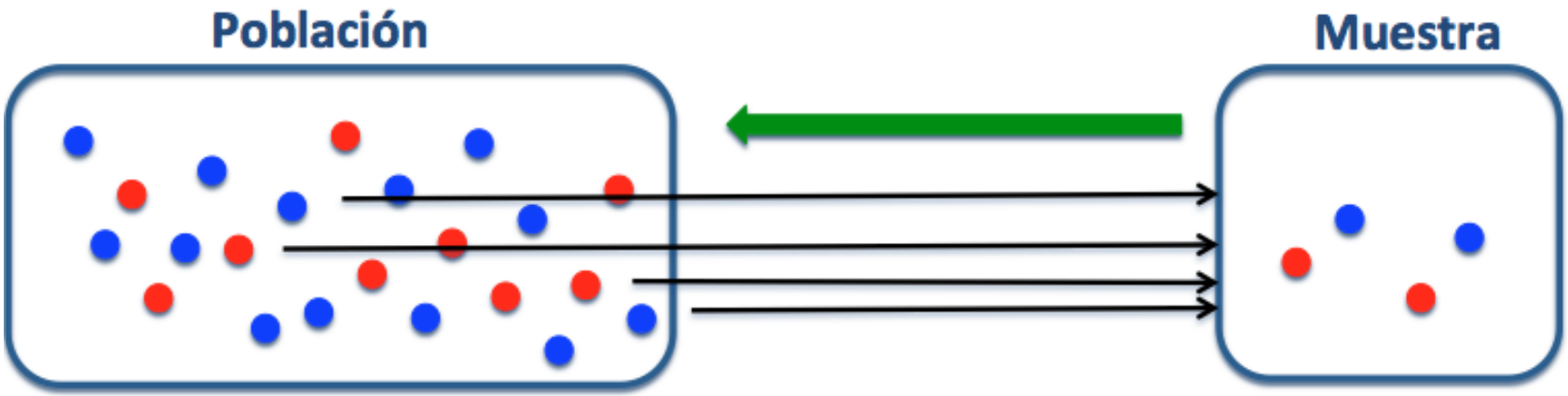
- Someter a prueba diferentes hipótesis estadísticas.
- Realizar gráficas avanzadas con ggplot2.

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 puede hacer de este experimento?



INFERENCIA ESTADÍSTICA 2

¿Para qué es útil?

- **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 el fertilizante A las plantas crecen más que con el fertilizante B.
  La eficacia de la vacuna en adultos es baja, pero en jóvenes es alta.

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

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

Ejemplo
A partir del muestreo de 30 individuos estimo el nivel medio de cortisol luego de un estrés agudo.

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.**
  Muestreo la diversidad de bacterias en una bahía justo en el efluente de una industria. Descarto animales pequeños porque quiero que sean grandes.
- **Ley de los grandes números.**
  ¿Mis variables están correlacionadas? 3 muestras v/s 300 muestras.
- **Evitar sesgo del investigador**
  Deseo aceptar la hipótesis "la vacuna funciona" repito el ensayo hasta que por azar funciona. No considerando las veces que no funcionó.
- **Otros**
  Errores, 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 mis peces es menor o 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 intención de ***rechazarla***.
- ¿Por qué no simplemente aceptar la hipótesis alternativa?
- Porque 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*.

¿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 $0{,}01$.

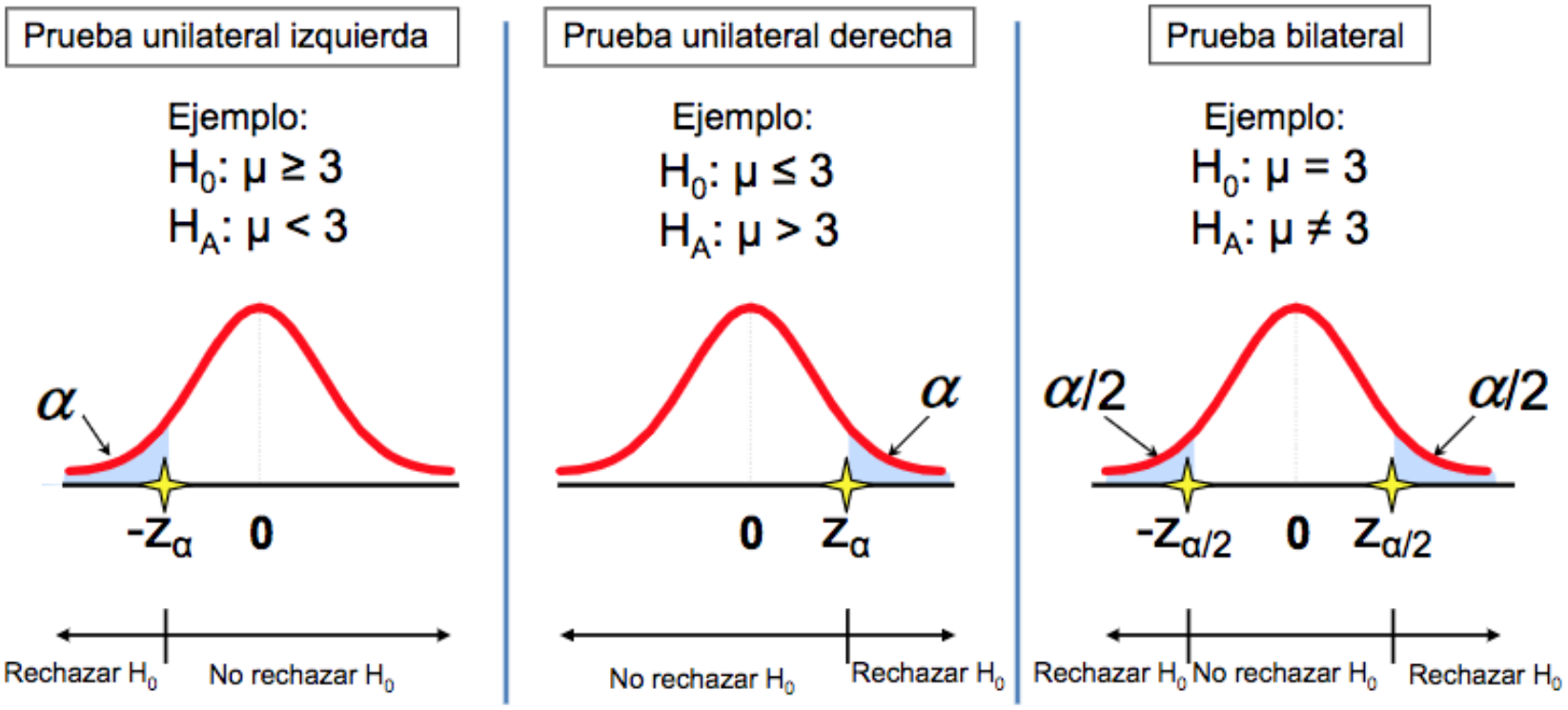
Es decir, rechazamos cuando el valor del estadístico está en el 5% inferior de la función de distribución muestral.

Corrección de Bonferroni 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



¿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

TIPOS DE PRUEBAS ESTADÍSTICAS

Según la forma de la distribución de la variable aleatoria.

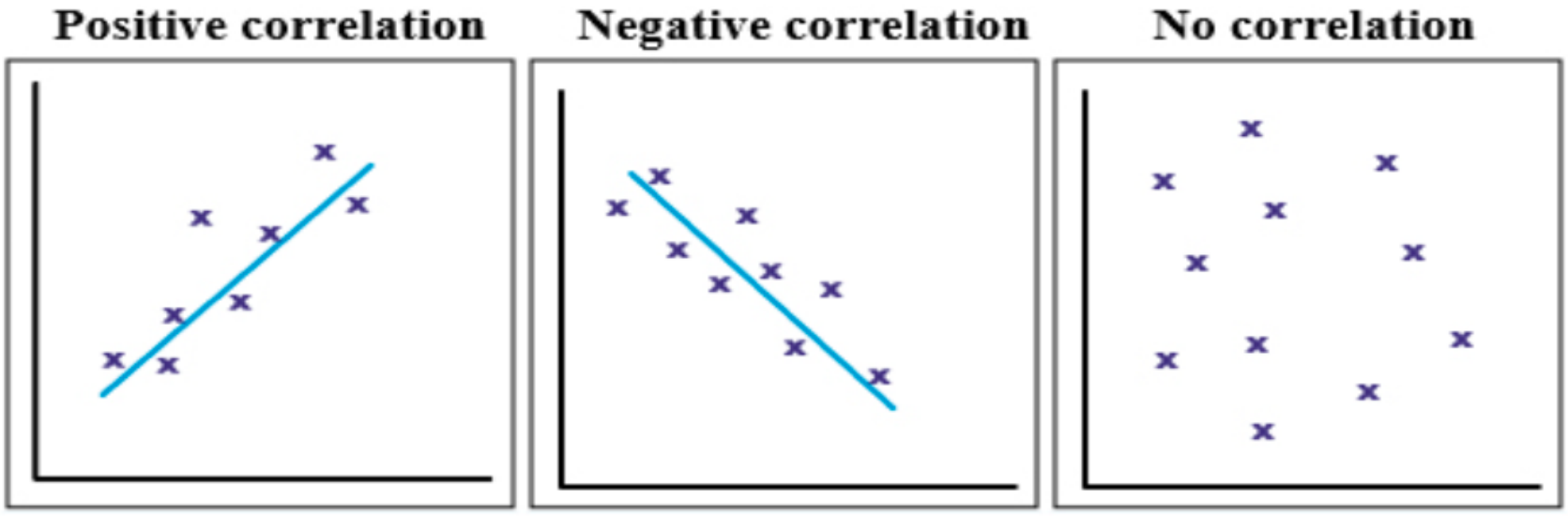
- **Métodos paramétricos** Las pruebas de hipótesis usualmente asumen una distribución normal de la variable aleatoria.

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.

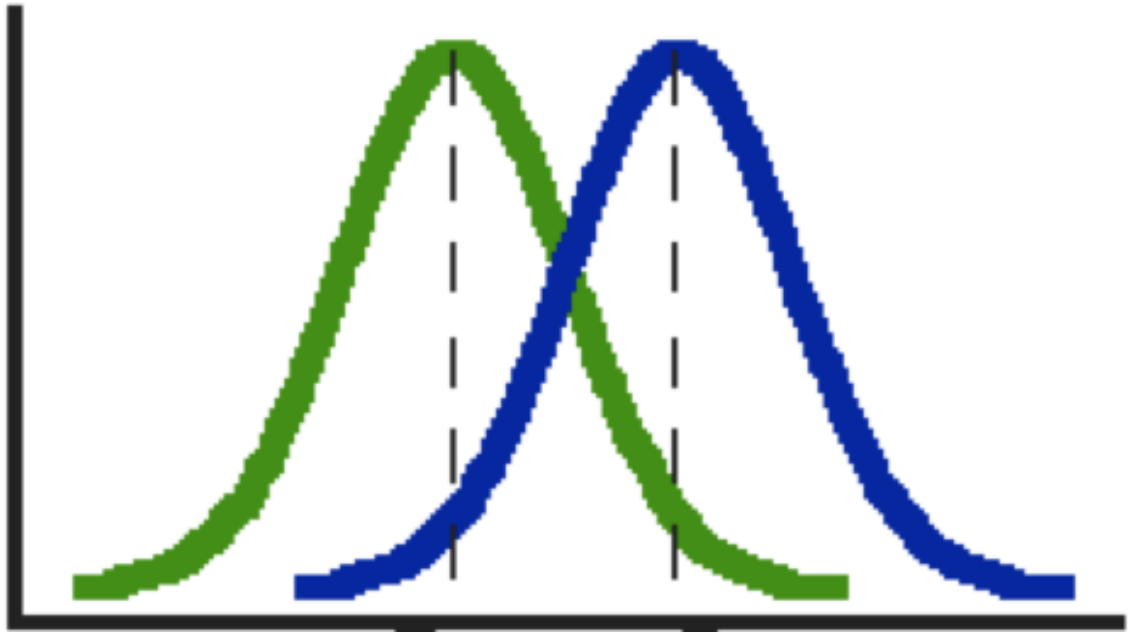
PRUEBA DE CORRELACIÓN



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

PRUEBA DE COMPARACIÓN DE MEDIAS



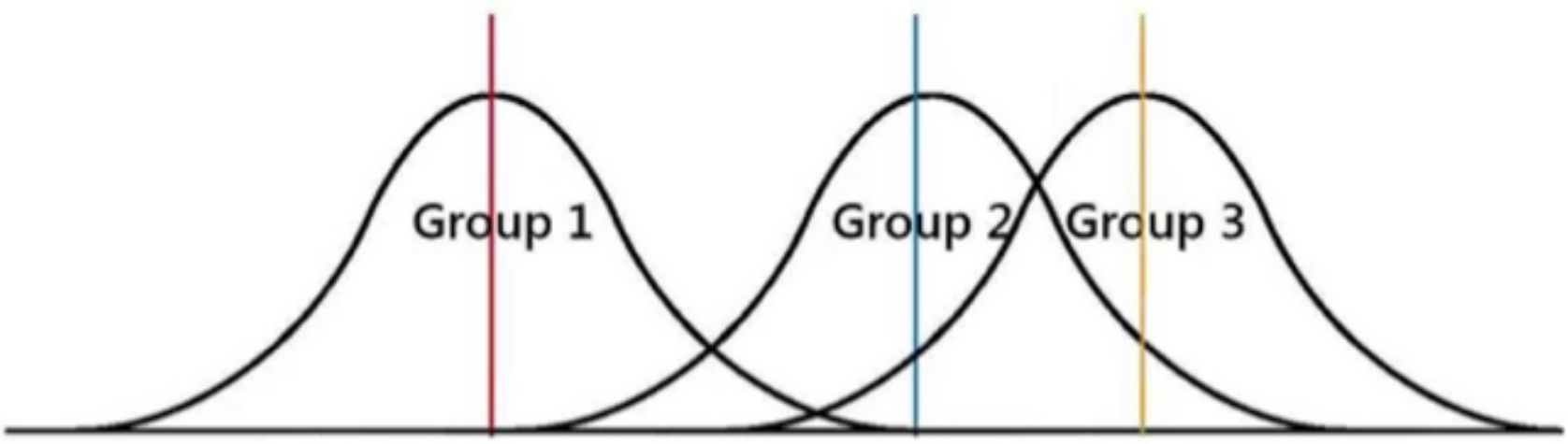
Hipótesis $H_0 : \mu_1 = \mu_2$ ***vs***. $H_1 : \mu_1 \neq \mu_2$

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

ANOVA

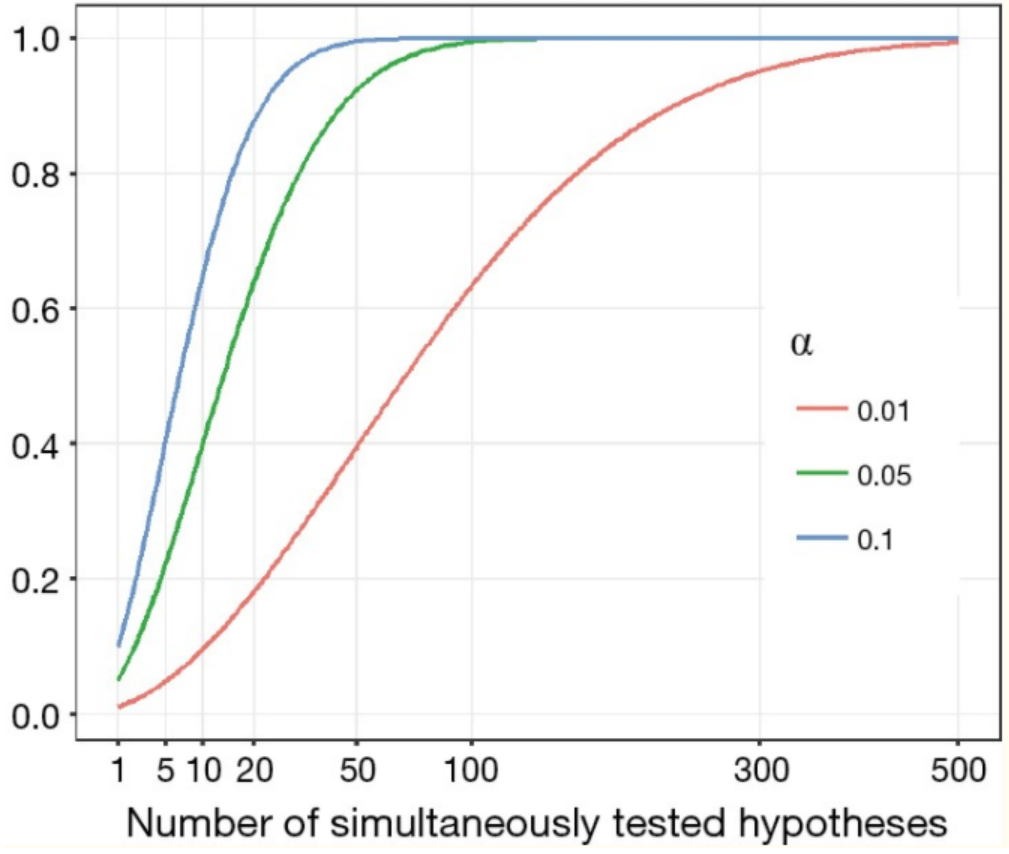
¿Qué es el análisis de varianza?

Herramienta básica para analizar el efecto de uno o más factores (cada uno con dos o más niveles) en un experimento.



PROBLEMA DE LAS COMPARACIONES MÚLTIPLES

¿Por qué preferir anova y no múltiples t-test?
Porque con una t-test normal se incrementa la tasa de error al aumentar el número de comparaciones múltiples.



Fuente[1]: [1]:doi:10.21037/jtd.2017.05.34

ANOVA: MODELOS LINENALES

Una forma muy conveniente de representar una ANOVA es mediante un modelo lineal.

Modelo lineal para ANOVA de una vía

$y \sim \mu + \alpha + \epsilon$

Modelo lineal para ANOVA de dos vías

$y \sim \mu + \alpha + \beta + \epsilon$

Modelo lineal para ANOVA de dos vías con interacción

$y \sim \mu + \alpha + \beta + \alpha^{*}\beta + \epsilon$

ANOVA: HIPÓTESIS Y SUPUESTOS

Hipótesis factor 1

H_0 : $\alpha_{1.1} = \alpha_{1.2} = \alpha_{1.3}$

Hipótesis factor 2

H_0 : $\beta_{2.1} = \beta_{2.2} = \beta_{2.3}$

Hipótesis interacción

H_0 : $\alpha^*\beta = 0$

Hipótesis Alternativa

H_A : No todas las medias son iguales

Supuestos:

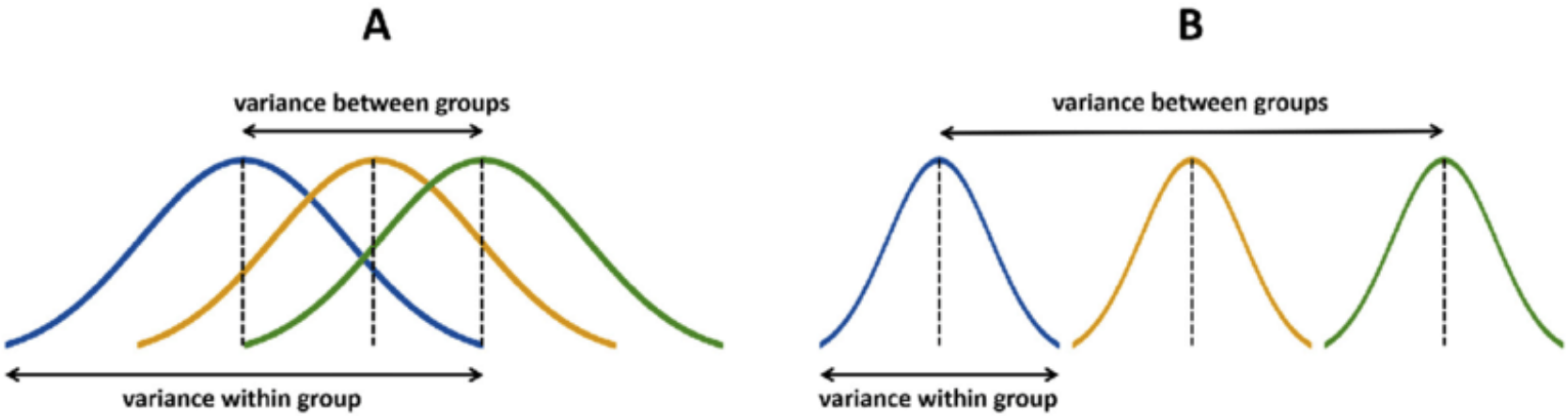
1) Independencia de las observaciones.
2) Normalidad.
3) Homocedasticidad: homogeneidad de las varianzas.

ANOVA PARA COMPARAR MEDIAS

¿Por qué se llama ANOVA si se comparan medias?
Por que el estadístico **F** es un cociente de varianzas.

$$\mathbf{F} = \frac{\sigma^2_{entregrupos}}{\sigma^2_{dentrogrupos}}$$

Mientras mayor es el estadístico **F**, más es la diferencia de medias entre grupos.



PRÁCTICA ANÁLISIS DE DATOS

- Guía de trabajo práctico disponible en drive y Rstudio.cloud.
- El trabajo práctico se realiza en Rstudio.cloud.

RESUMEN DE LA CLASE

- **Conceptos básicos de inferencia estadística**
- **Conceptos básicos de pruebas de hipótesis**
- **Realizar pruebas de hipótesis**
  - Test de correlación.
  - Test de comparación de medias para 2 muestras independientes.
  - Test de ANOVA.
- **Realizar gráficas avanzadas con ggplot2**.