Clase 05 - Inferencia estadística

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

Dr. José Gallardo Matus

Pontificia Universidad Católica de Valparaíso

17 July 2022

PLAN DE LA CLASE

1.- Introducción

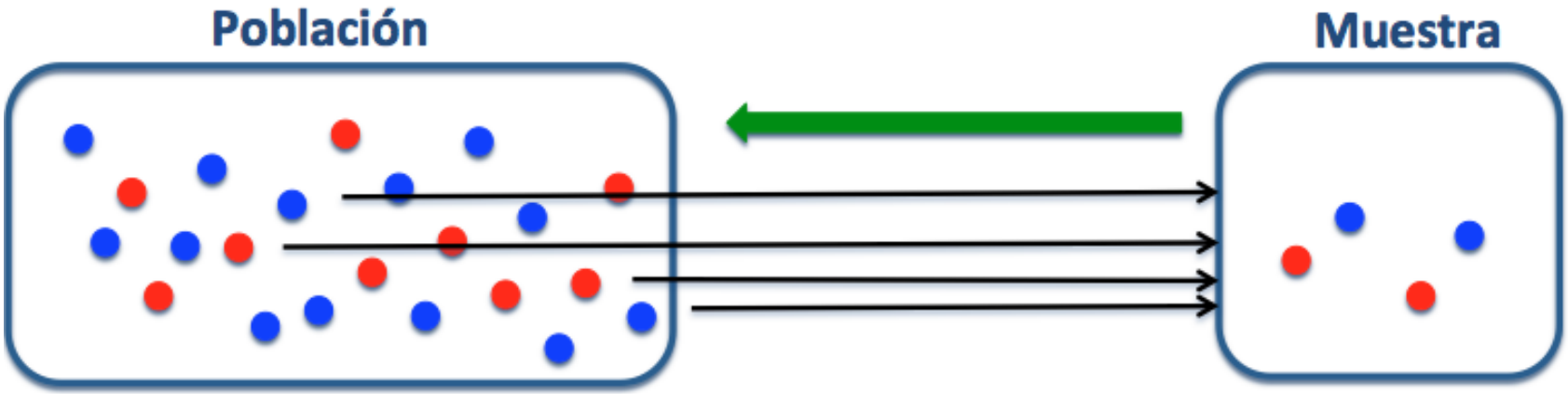
- ¿Qué es la inferencia estadística?.
- ¿Cómo someter a prueba una hipótesis?
- Pruebas paramétricas
- Interpretar resultados de análisis de datos con R.

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 aditivo A los peces crecen más que con aditivo B.
- **Bajo ciertos supuestos permite hacer predicciones.**
  Peces con genotipo GG maduran más que peces con genotipo AA.

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) = 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 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.
  Ej. media de cortisol en plasma es de 15 $\mu gramos/decilitro$ (N=30).
- **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. Ej. Intervalo de confianza: del 95% nuestro parámetro estará entre 10 y 20 $\mu gramos/decilitro$, 95 de 100 veces.

ERROR EN LA ESTIMACIÓN DE PARÁMETRO.

¿Puedo estimar erróneamente un parámetro?
Por supuesto, los errores se producen por violar algunas premisas.

- **Las muestras deben tomarse de forma aleatoria.**
  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.**
  Compare experimento de 3 muestras v/s 300 muestras.
- **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
$\mathbf{H_0}$: El nivel medio de cortisol es $= 15$ microgramos por decilitro.
$\mathbf{H_A}$: El nivel medio de cortisol es > 15 microgramos por decilitro.

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*.

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

¿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 $0{,}01$.

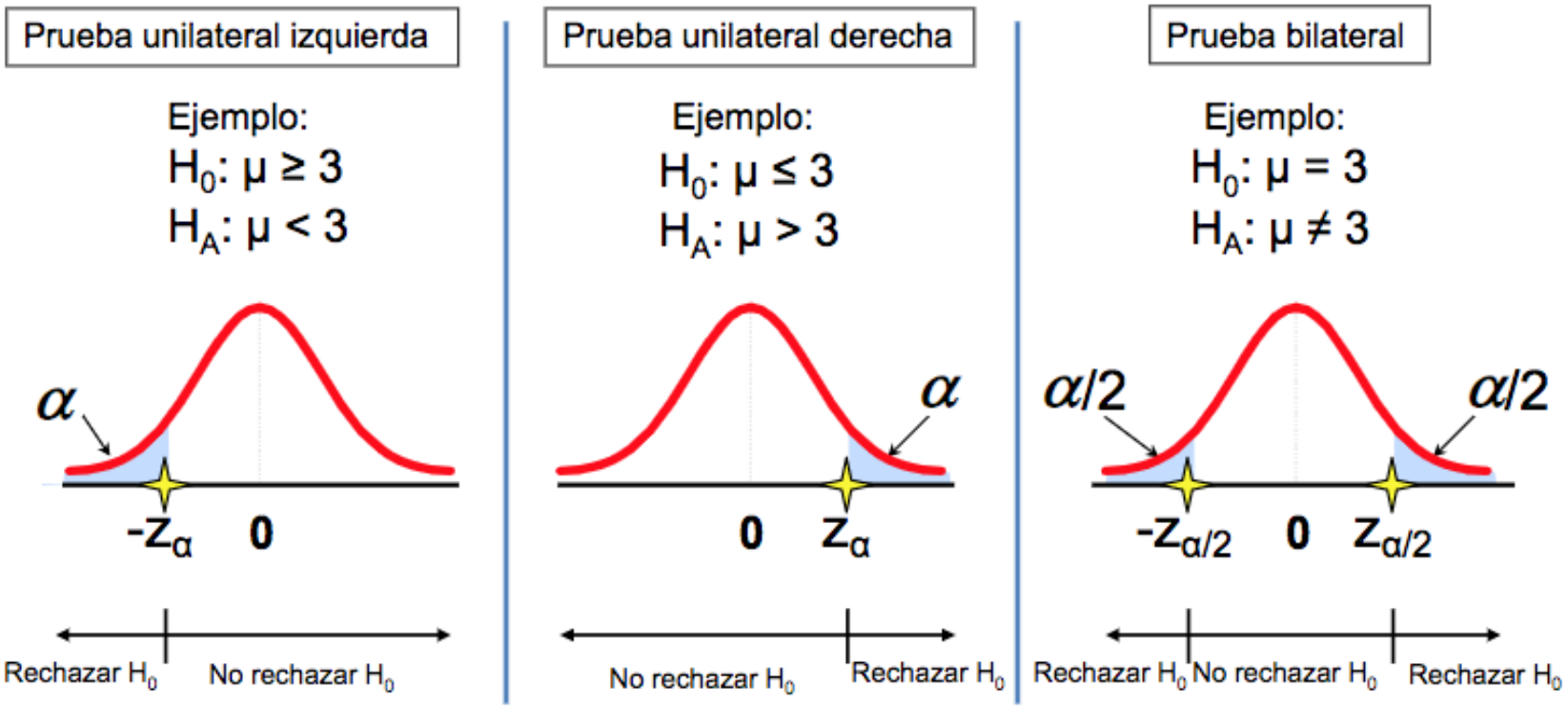
Es decir, rechazamos cuando el valor del estadístico está en el 5% inferior de la función de distribución muestral.

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



TIPOS DE PRUEBAS ESTADÍSTICAS

Según la forma de la distribución de la variable aleatoria.

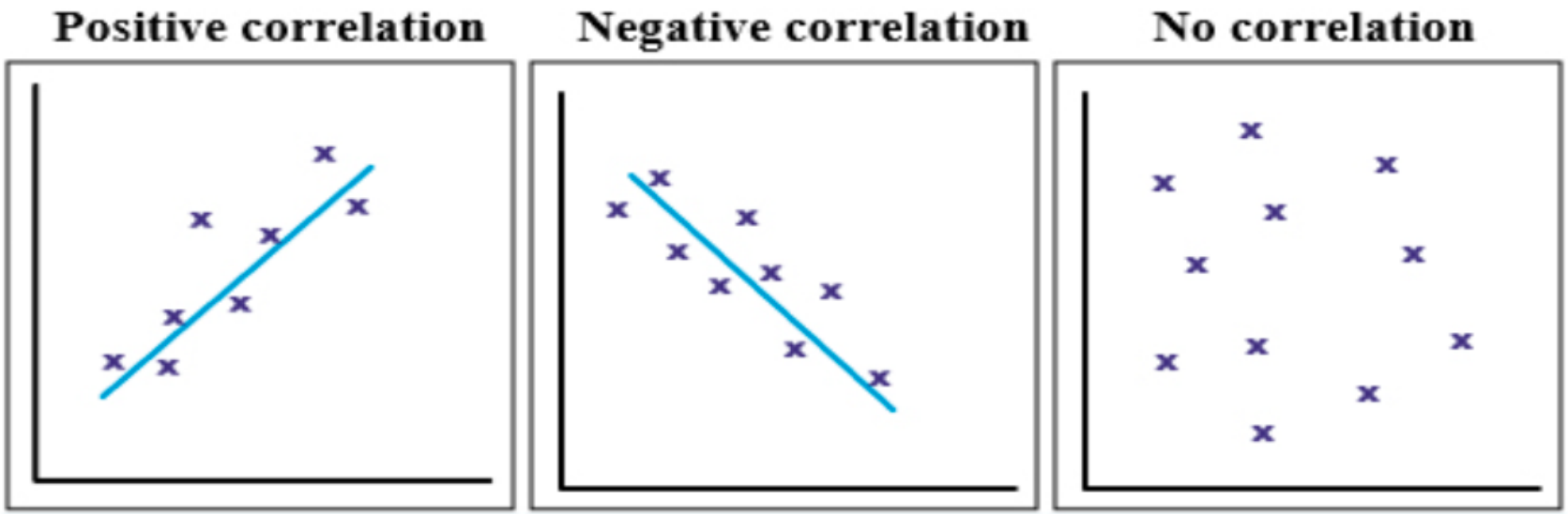
1. **Métodos paramétricos**

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

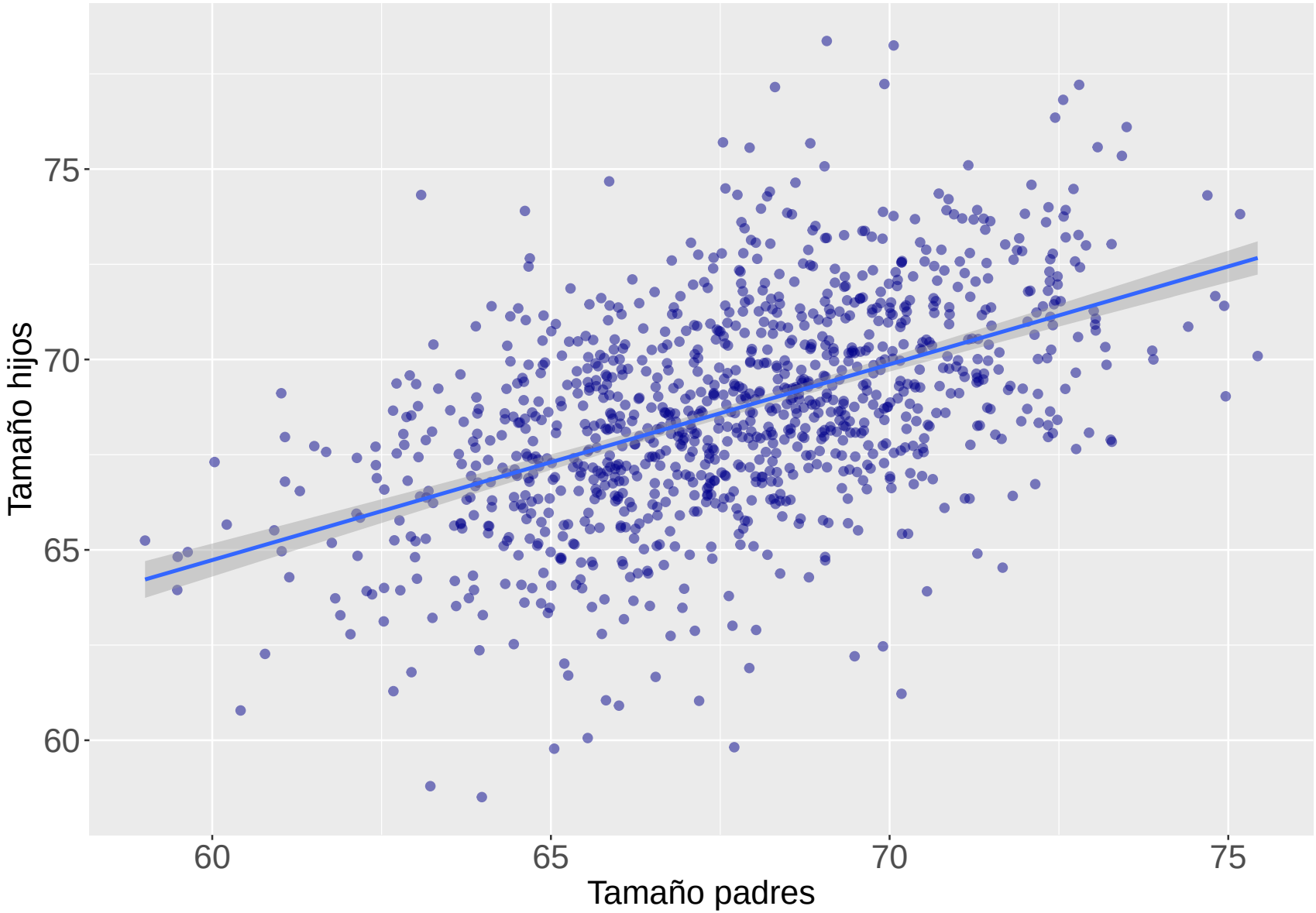
2. **Métodos NO paramétricos**

- 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



ESTUDIO DE CASO: TAMAÑO PADRES - HIJOS



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 en lineal.

2) La distribución conjunta de (X,Y) es una distribución Bivariable normal.

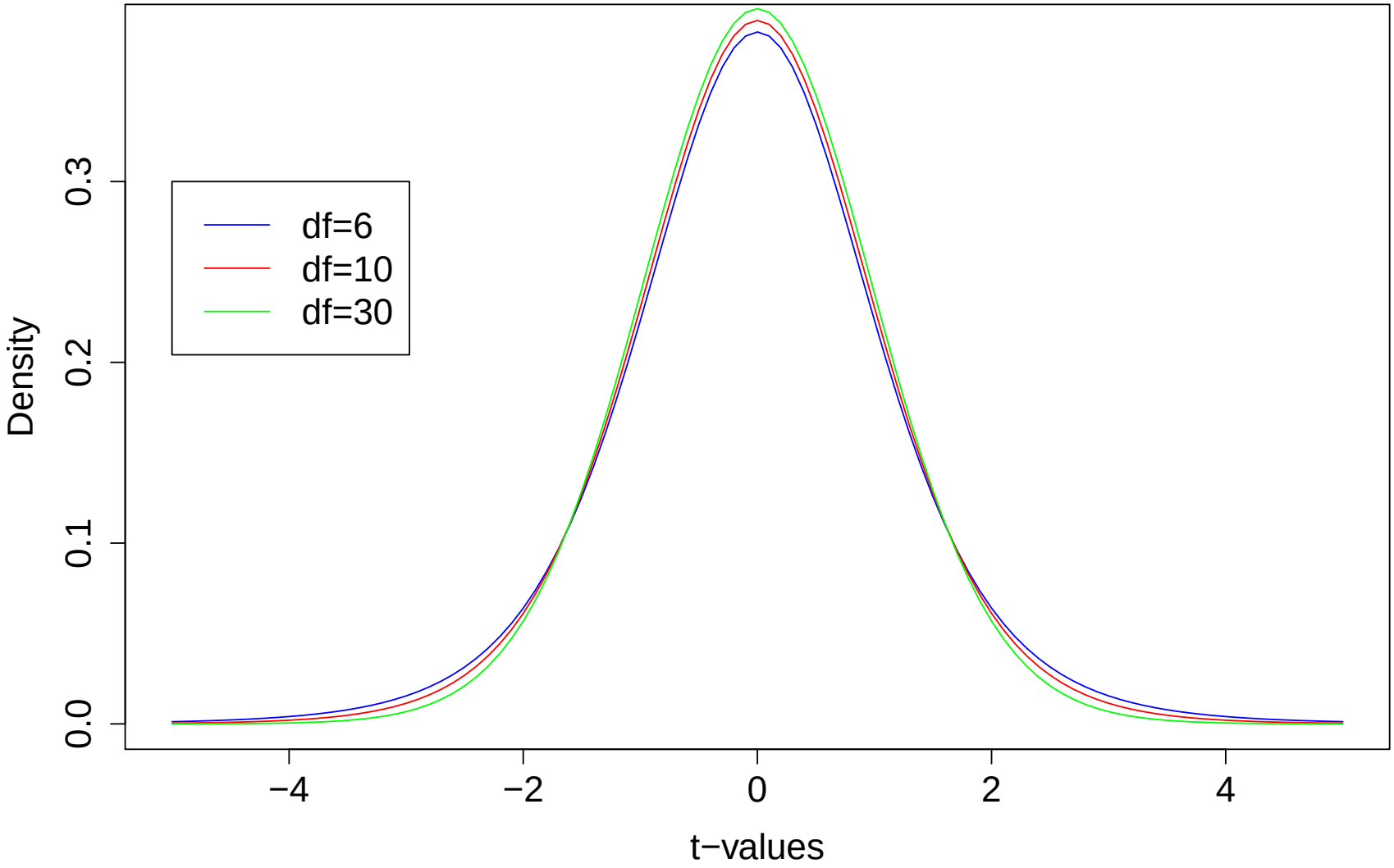
PRUEBA DE CORRELACIÓN DE PEARSON

```
cor.test(father.son$fheight, father.son$sheight,
         alternative = c("two.sided"))
```

```
##
##  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
## 95 percent confidence interval:
##  0.4552586 0.5447396
## sample estimates:
##       cor
## 0.5013383
```

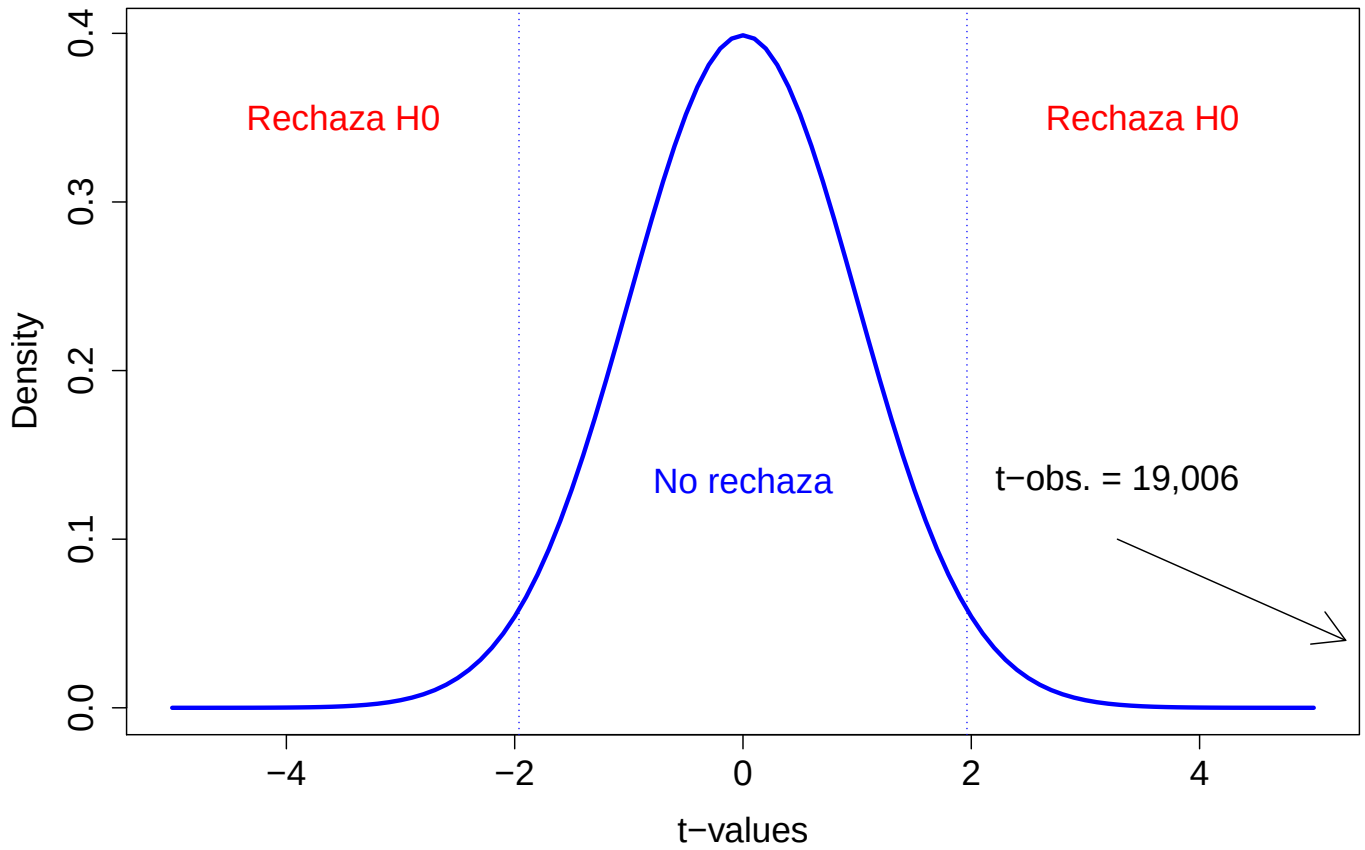
DISTRIBUCIÓN T STUDENT

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

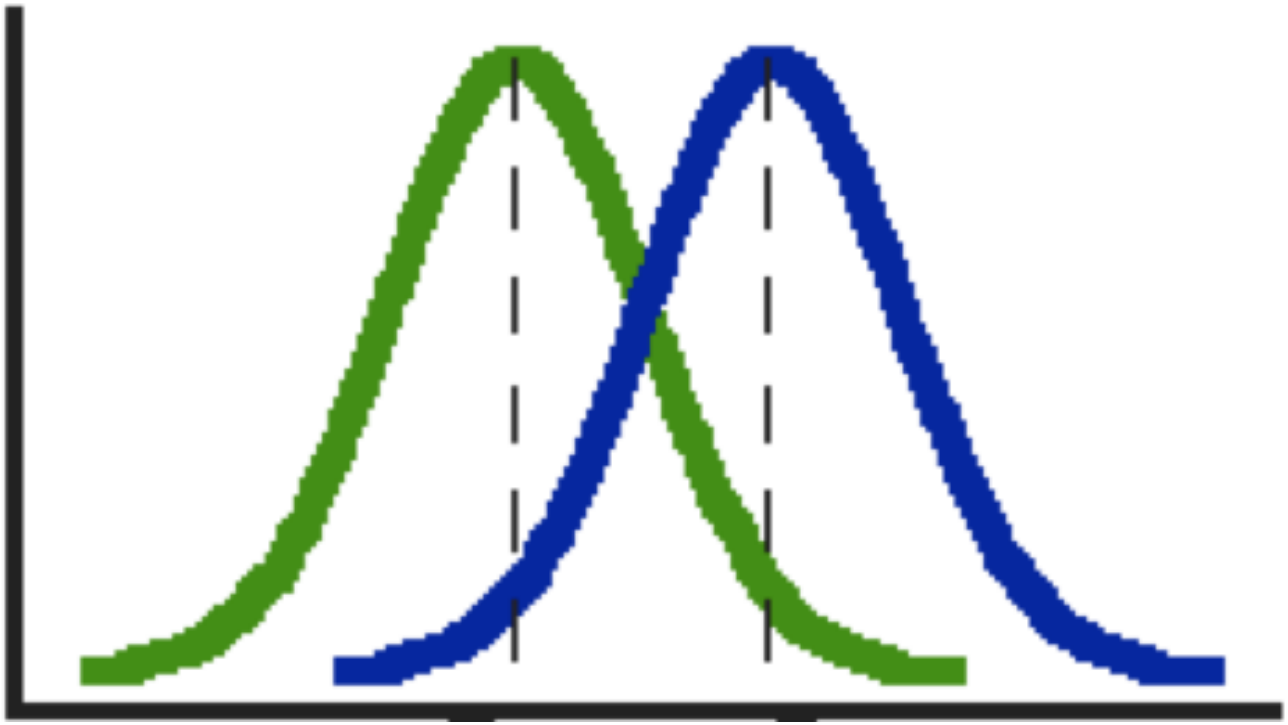


PRUEBA DE HIPÓTESIS

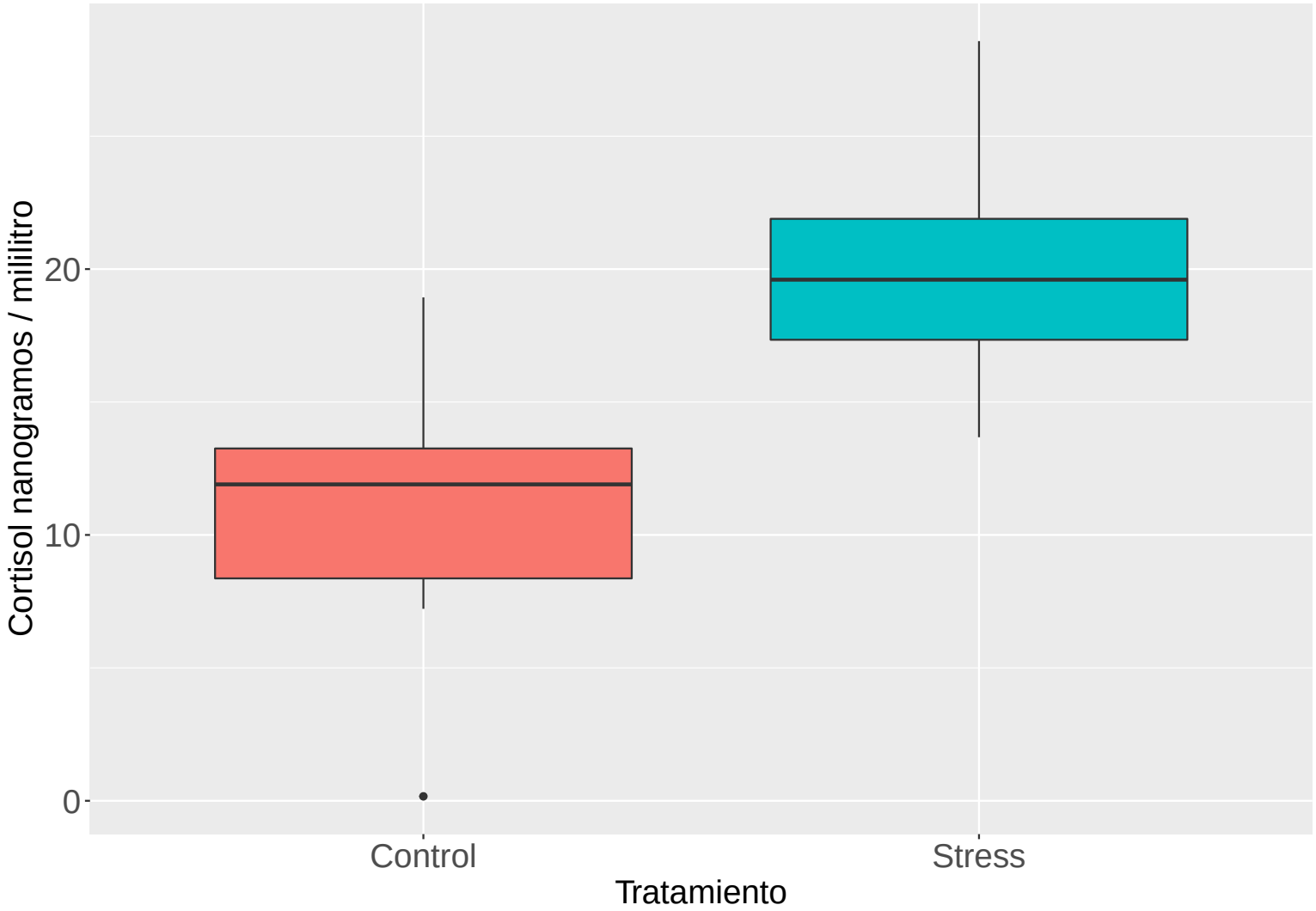
- gl correlación = Nº datos - 2 = 1078 - 2
- Región de no rechazo= t-student (gl=1076) = -1.96 - 1.96



PRUEBA DE COMPARACIÓN DE MEDIAS



ESTUDIO DE CASO: CORTISOL EN SALMON



Adaptado de Fast, et al. 2006

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.

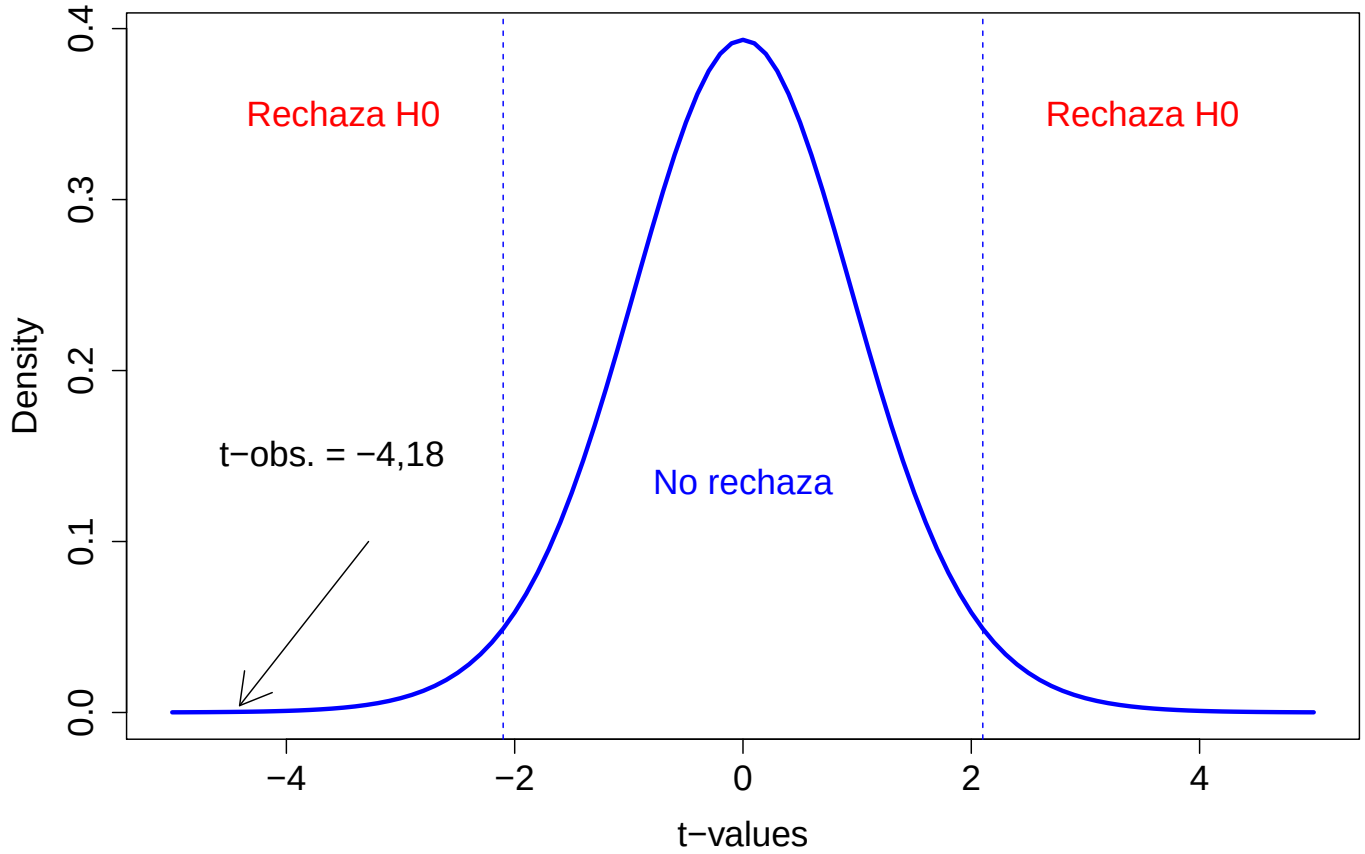
PRUEBA DE T PARA DOS MUESTRAS INDEPENDIENTES

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

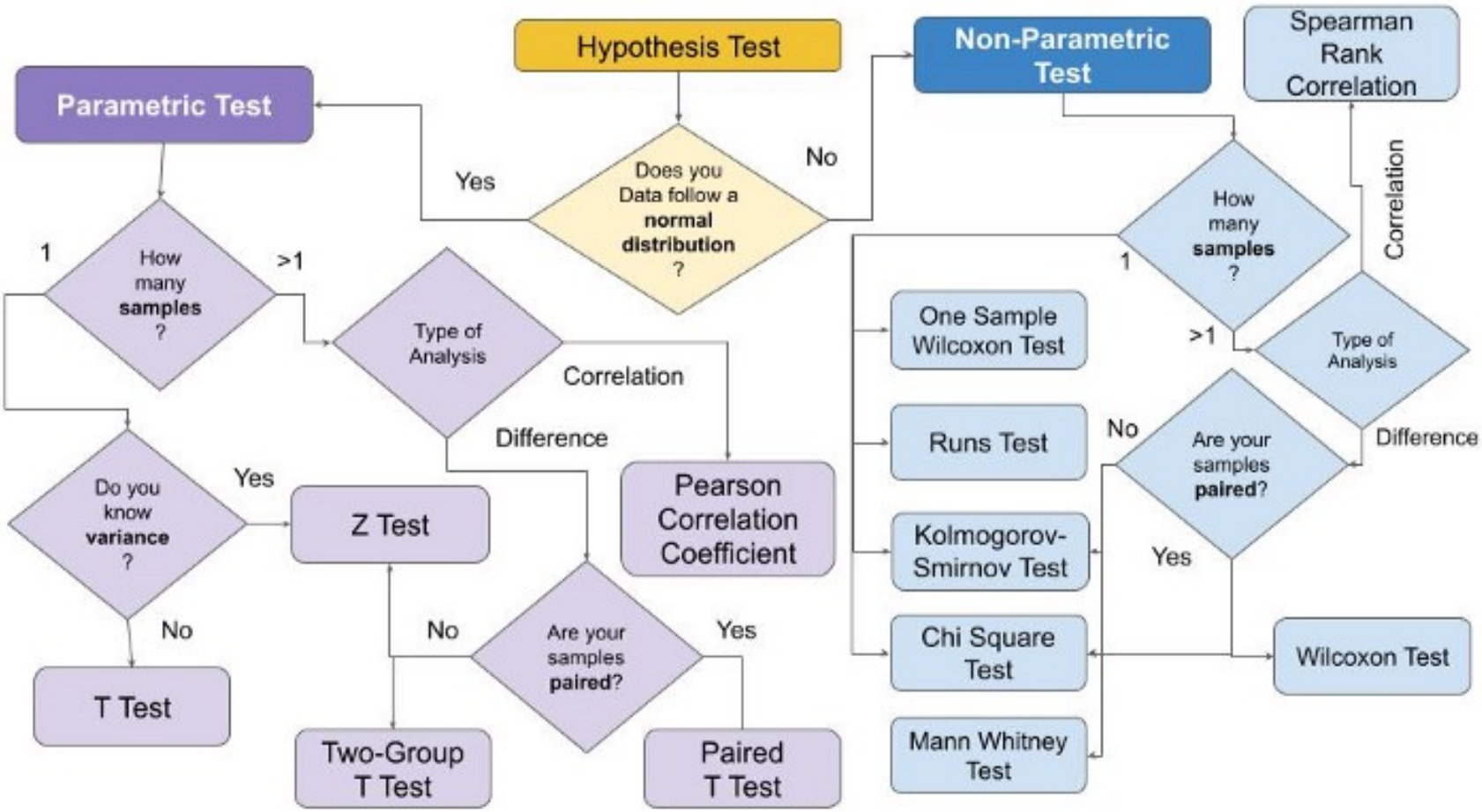
```
##
##  Two Sample t-test
##
## data:  Cortisol by Tratamiento
## t = -4.1858, df = 18, p-value = 0.0005554
## alternative hypothesis: true difference in means is not
## 95 percent confidence interval:
##  -14.012884  -4.647153
## sample estimates:
## mean in group Control  mean in group Stress
##              11.04311                20.37313
```

PRUEBA DE HIPÓTESIS

- gl correlación = Nº 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**.