Instituto de Ciencias Antonio Brack

ISCB REGIONAL Student GROUP

Peru Ayacucho

>SPB

Domingo 24 de octubre 2021

Turno mañana (8:30 am - 12:30 pm)

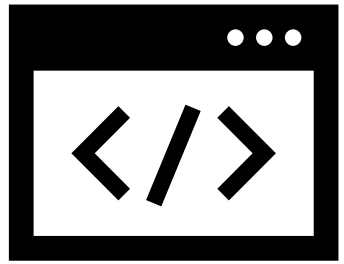
Turno tarde (3:30 - 7:30 pm)

Introducción a Data Science y Programación en RStudio

MEd(c) Blgo. Irwing S. Saldaña Ugaz

Dpto. Ecoinformática y Biogeografía,
Instituto de Ciencias Antonio Brack

Lo que aprenderás



Módulo 1. Introducción al Lenguaje de Programación R

- Variables: clases y estructuras de datos
- Paquetes y Funciones: creación de funciones y funciones básicas
- Importación de bases de datos desde Hojas de Cálculo

Módulo 2. Estadística descriptiva

- Probabilidades
- Función de densidad y pruebas de normalidad
- Análisis exploratorio básico: gráficos importantes, histogramas, gráficos de barras, gráficas de densidad.
- Cálculo de medidas de tendencia central: media, moda, mediana
- Medidas de dispersión: varianza, desviación estándar.

Módulo 3 . Estadística Inferencial Básica

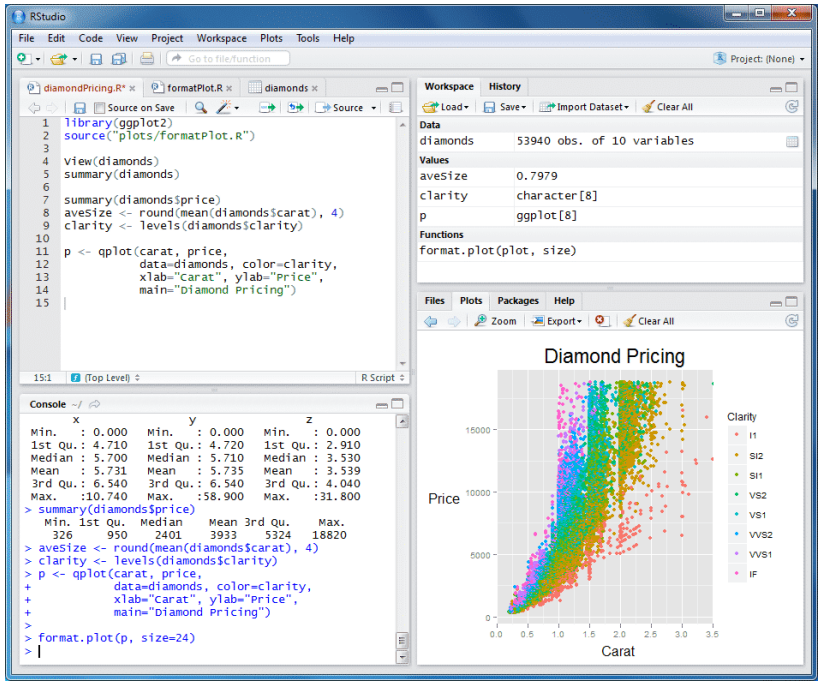
- Regresión Lineal OLS
- Comparaciones pareadas:
- Pruebas de t, ANOVA de una y dos vías.

Módulo I.

Introducción al lenguaje de programación R

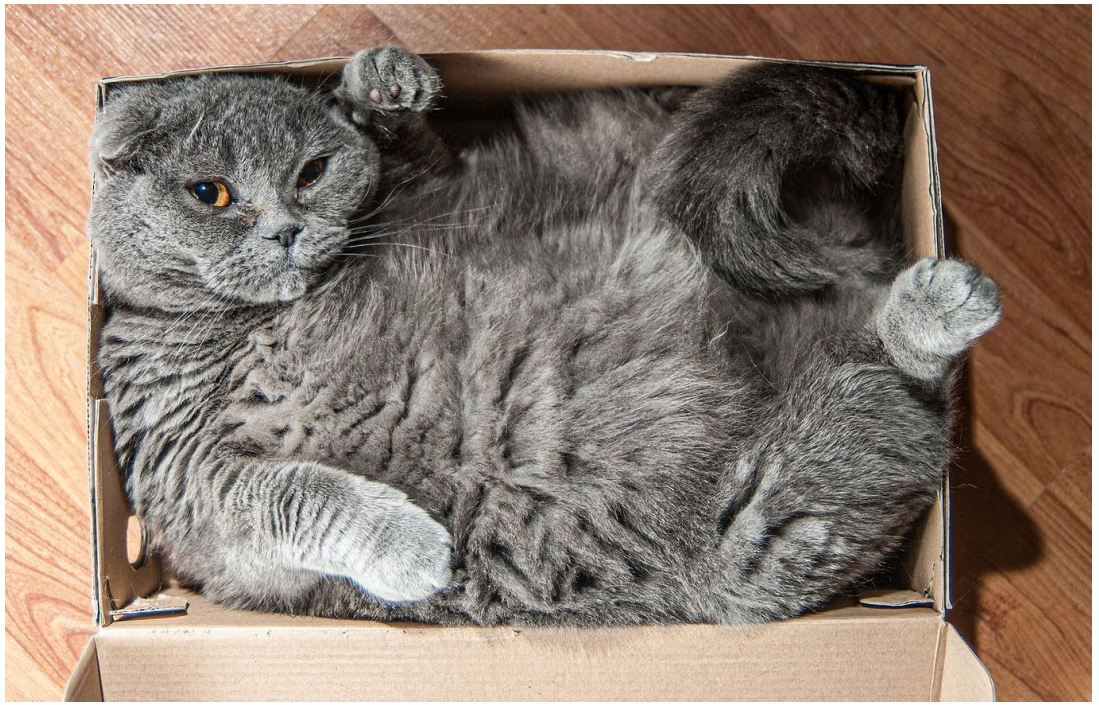
Lenguaje de programación R

- **Programar:** dar ordenes a una computadora esperando un resultado.
- **Lenguaje de programación:** sintaxis con la cual se escriben las órdenes.



Variables: creación

- También conocidos como **objetos**
- Una variable es como un espacio virtual (de una clase y estructura específicos) que contiene algo.



```
# Crear es asignar un contenido a un nombre
# el operador de asignación es
<-

# Se usa de la siguiente manera
NOMBRE <- contenido
Contenido -> NOMBRE
```

Variables: clases

- Describe de qué tipo de elementos está constituida una variable.
  - Números (numeric)
  - Texto (character)
  - Elementos lógicos (logical)
- Podemos coercionar (cambiar) una variable de una clase a otra (con ciertas limitaciones).

```
# Funciones de preguntas lógicas
is.numeric()
is.character()
is.logical()
```

```
# Funciones de coerción
as.numeric()
as.character()
as.logical()
```

Variables: estructuras de datos

- Esta característica de una variable explica cómo es que se organizan los datos dentro de la variable.

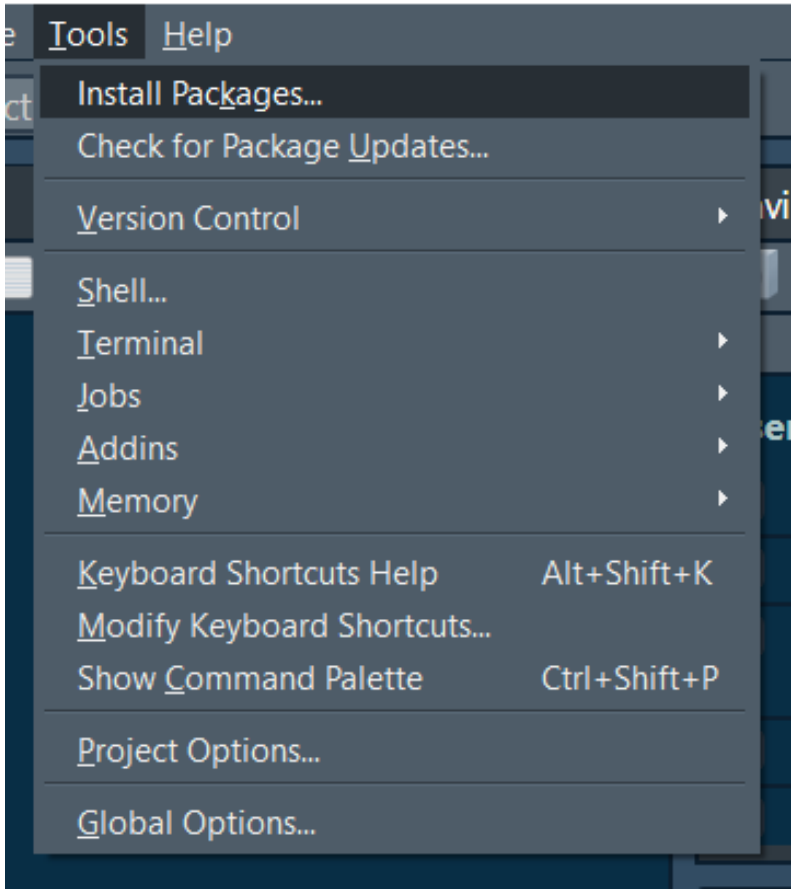
```
# Vectores 1D    # Tibbles 2D
c()              tibble()

# Factores 1D    # Listas 3D
factor()         list()

# Matrices 2D    # Arrays 3D
matrix()         array()

# Data.Frames 2D
data.frame()
```

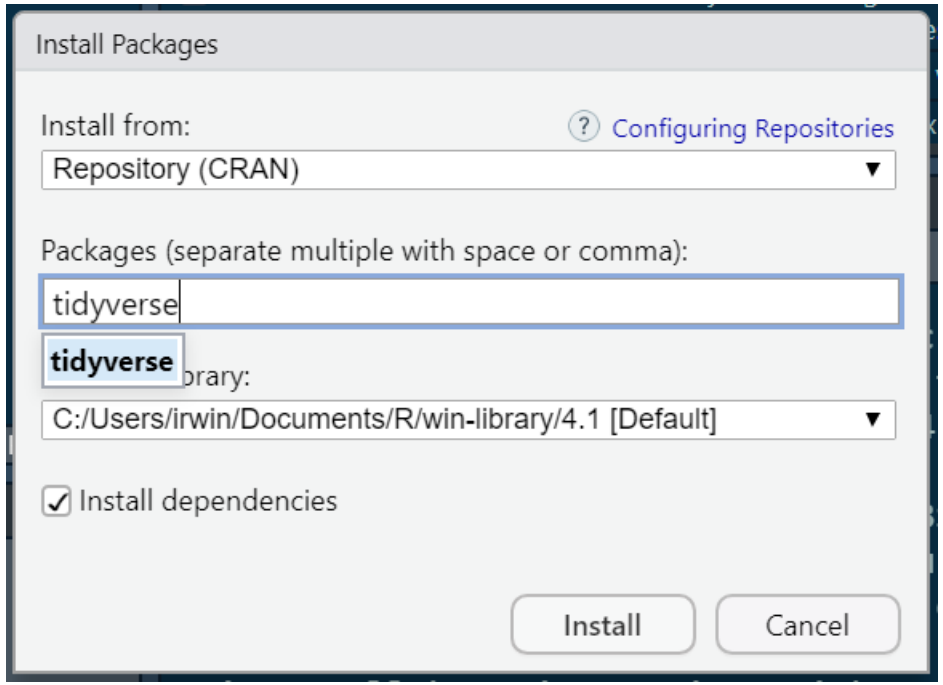
Variables: estructuras de datos

- Esta característica de una variable explica cómo es que se organizan los datos dentro de la variable.

```
# Funciones de preguntas lógicas
is.vector()
is.factor()
is.list()
is.matrix()
is.data.frame()

# Funciones de coerción
as.vector()
as.factor()
as.list()
as.matrix()
as.data.frame()
```

Paquetes en R






```
# Instalando con código desde CRAN
install.packages("BiocManager")

# Instalando con código desde Bioconductor
BiocManager::install()

# Instalando con código desde GitHub
devtools::install_github("hadley/stringr")
```

Creación de funciones

Asignarle un nombre a la nueva función

Usar la función `function()`

Argumento(s) de la función dentro de los paréntesis

```
FarCel <- function(x) {
  Centigrados <- (x - 32) * 5 / 9
  return(Centigrados)
}
```

Se suele colocar `return()` para imprimir el resultado de la función en la consola

Las llaves encierran el contenido de la función

58 Funciones básicas en R

```
set.seed()      sample()        na.omit()       coef()
names()         setwd()         par()           sort()
View()          getwd()         log()           subset()
str()           file.choose()   log10()         library()
unique()        plot()          sqrt()          data()
relevel()       density()       scale()         lenght()
colSums()       boxplot()       lm()            nrow()
colMeans()      hist()          aov()           ncol()
lapply()        mean()          tukeyHSD()      cbind()
apply()         median()        t.test()        rbind()
summary()       sd()            leveneTest()
rnorm()         var()           wilcox.test()
paste()         min()           exp()
rep()           max()           shapiro.test()
seq()           range()         confint()
```

- [R Functions List (+ Examples) | All Basic Commands of R Programming (statisticsglobe.com)]

Importación de bases de datos en R

- Normalmente se trabaja con bases de datos manejadas en hojas de cálculo
  - .xlsx
  - .csv
  - .txt

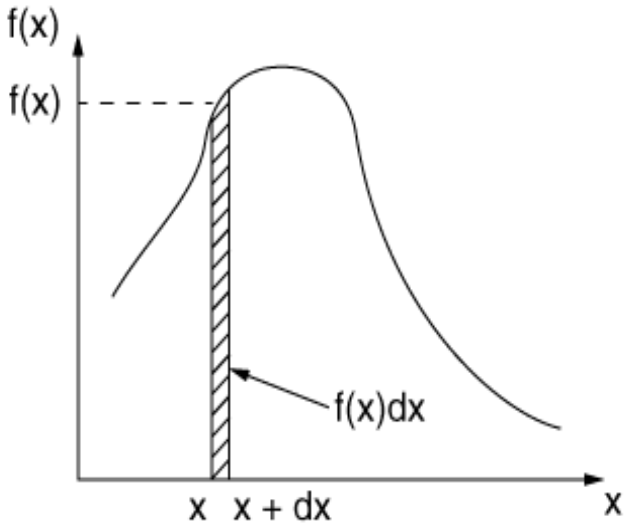
```
# Cargar un Excel en R
openxlsx::read.xlsx()

# Cargar un csv en R
read.csv()

# Cargar un Excel en R
read.delim()
```

Módulo II.

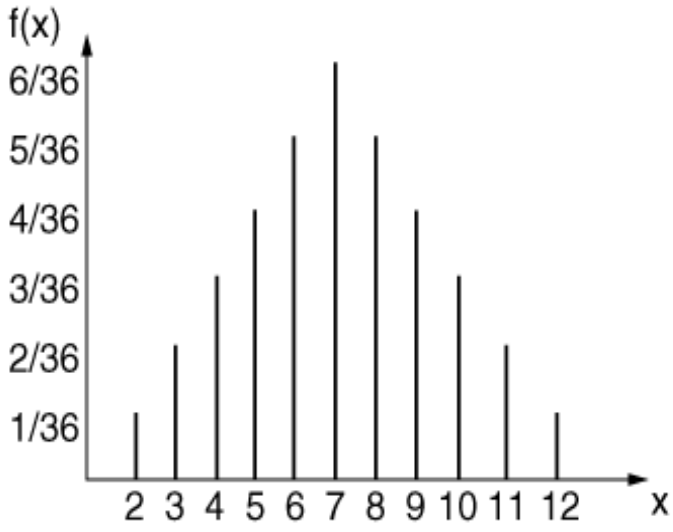
Estadística descriptiva

Todo conjunto de datos tiene su propia distribución de **probabilidades**

- **Probabilidad:** es el grado de certidumbre de que dicho suceso pueda ocurrir (de 0 a 1).
- **Discretas**
  - La probabilidad de un evento es igual su proporción.
- **Continuas**
  - La probabilidad se toma como un área bajo la curva de función de densidad donde es probable que se encuentre un valor dado, mas ya no es la probabilidad de un valor específico.



Función de Densidad de Probabilidad



Función de Masa de Probabilidad

Distribución de Probabilidades Teóricas Discretas

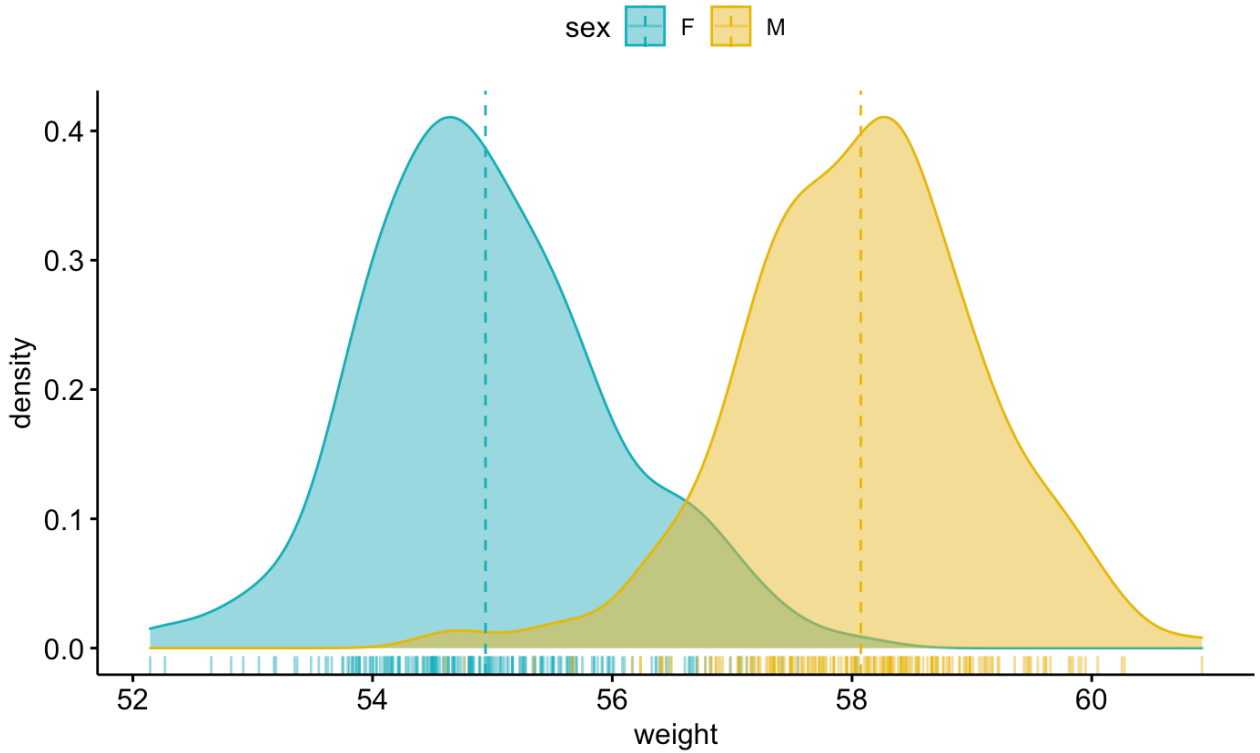
Distribución	Descripción
Bernoulli	Cada experimento genera un resultado. Estos son independiente y solo tiene dos opciones de resultado (0,1). Ambos tiene la misma prob. Se anota éxito o fracaso.
Binomial	Cada experimento genera una secuencia de resultados (“N”). De ellos, “n” son exitosos. Cada experiment tiene entonces una proporción asociada del tipo n/N, el cual se encuentra siempre entre 0 y 1. En otras palabras, es el número de éxitos de cada secuencia de experimentos.
Poisson	Modela cuantas veces es probable que ocurra un evento (conteo) en un periodo de tiempo definido.
Binomial Negativa	Alternativa especial a Poisson cuando hay sobredispersión (u/var) de los datos.

Distribución de Probabilidades Teóricas Continuas

Distribución	Descripción
Normal	Conjuntos de datos numéricos continuos simétricos, que tienen curtosis y simetría con valores entre -2 y +2.
Beta	Permite modelar conjuntos de datos limitados al rango de 0 a 1 (proporciones), exceptuando 0 y 1.
Gamma	Conjuntos de datos numéricos continuos que son siempre positivos y tienden a la normalidad, pero estos muestras sesgo hacia un lado (son asimétricos: valores de simetría superiores o inferiores al rango -2 a +2).
Chi square	Usada principalmente en pruebas de hipótesis, para testear la relación entre variables categóricas, la bondad de ajuste de datos vs distribuciones teóricas, test de independencia en tablas de contingencia, o Likelihood Ratio Test en ANOVAs.
T-Student	Es usada para estimar la distribución de los parámetros y suma de cuadrados residuales en regresiones lineales. Es muy robusto cuando tamaño de la muestra es pequeño y/o cuando se desconoce la varianza de la población.

Funciones de densidad

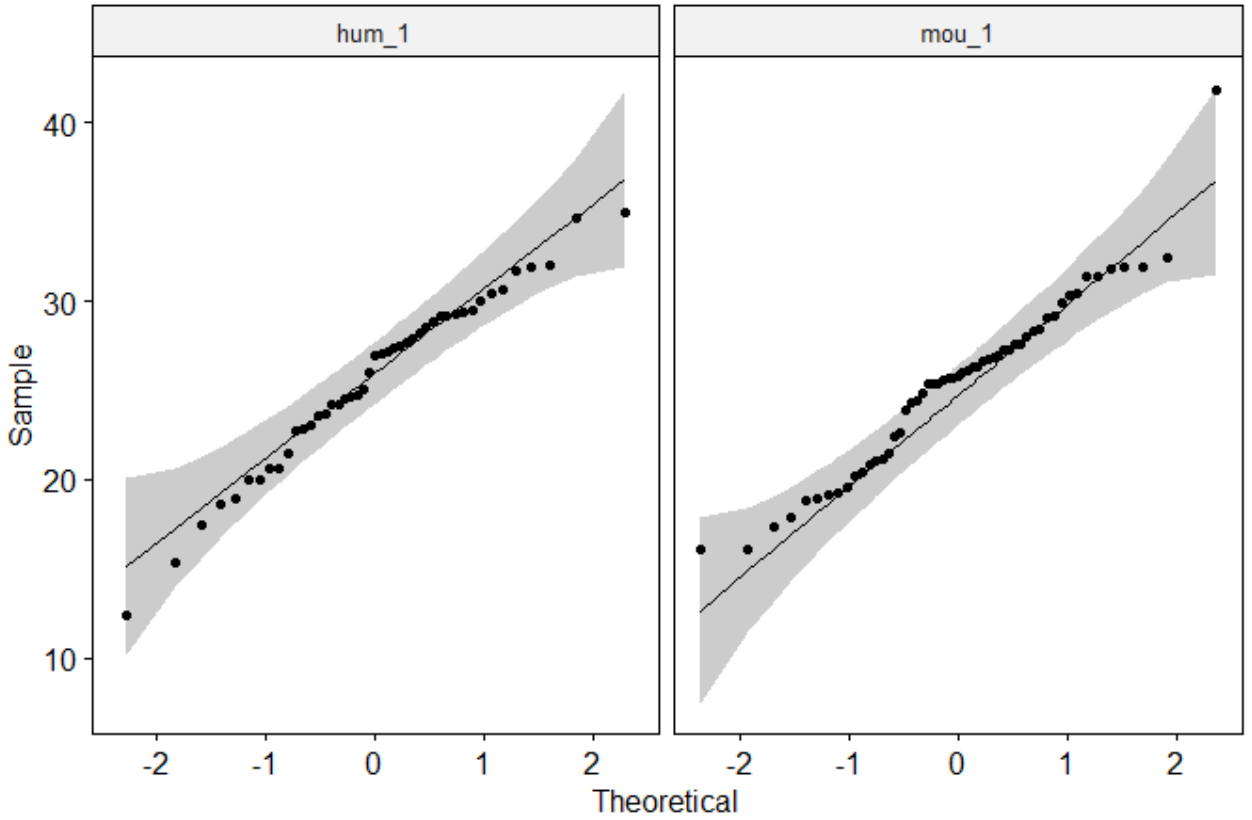
```
# Con código simple
plot(density(VECTOR))

# Con código asociado a ggplot2
library(ggpubr)
ggdensity(x="COLUMNA", data=TABLA)
```



Pruebas de Normalidad

- La tendencia actual es a usar graficación para identificar normalidad.

```
# Test de Shapiro-Wilk
shapiro.test(x)

# Test de Kolmogorov-Smirnov
ks.test(x, "pnorm", mean(x), sd(x))

# Test de Anderson-Darling
nortest::ad.test(x)

# Gráfico Q-Q Plot (Cuantil-Cuantil)
library(ggpubr)
ggqqplot(x)
```



Medidas de tendencia central

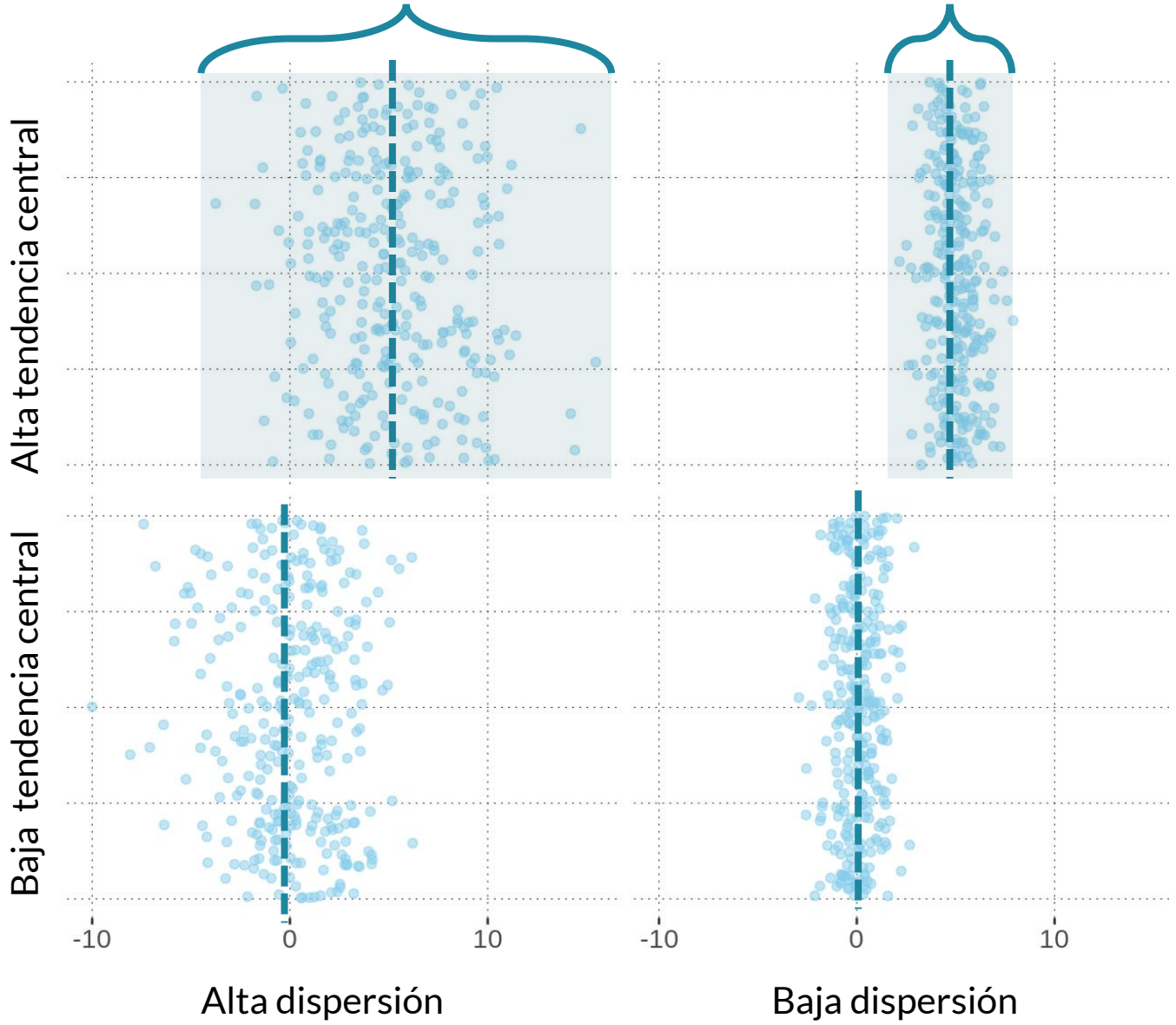
- Es un valor que busca describir un conjunto de datos identificando una posición central en el mismo.
- Es el primero de los dos descriptores principales de cualquier conjunto de datos.

```
# Promedio aritmético
mean()
# Promedio aritmético
psych::geometric.mean()
# Promedio aritmético
psych::harmonic.mean()
# Mediana
median()
# Moda
moda() # No existe, debemos crearla
```

Medidas de dispersión (o variabilidad)

- Es un valor que busca describir la dispersión de un conjunto de datos.
- Es el primero de los dos descriptores principales de cualquier conjunto de datos.

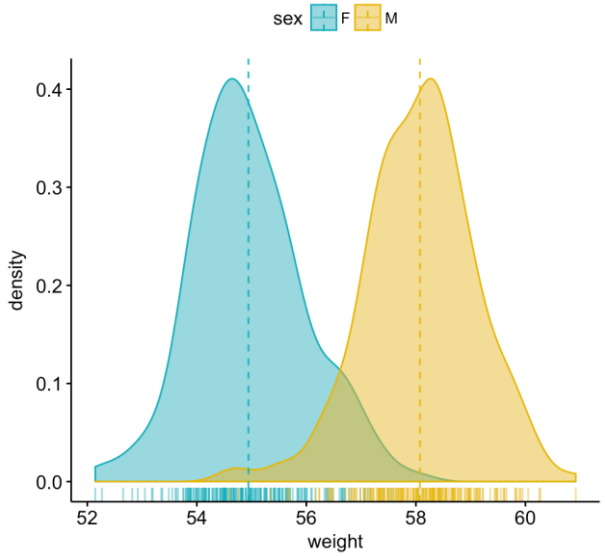
```
# Desviación estándar
sd()
# Varianza
var()
# Mínimo
min()
# Máximo
max()
# Rango
range()
```


Alta tendencia central
Baja tendencia central
-10
0
10
-10
0
10
Alta dispersión
Baja dispersión

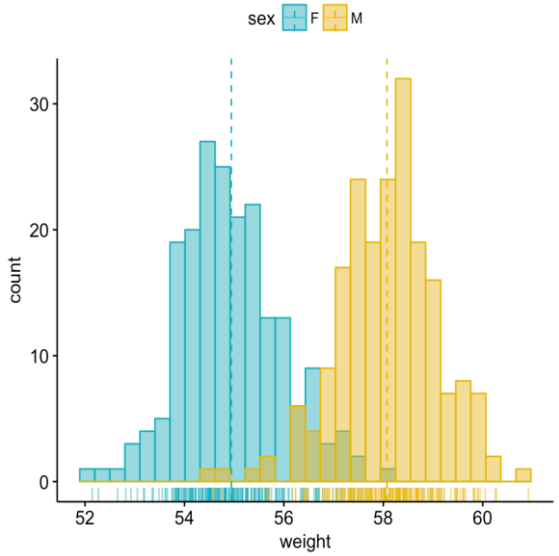
Análisis exploratorio básico

- Básicamente implica:
  - Importar los datos
  - Limpiarlos (eliminar filas sin data, limpiar NA)
  - Procesamiento de datos (especificar la clase/estructura de datos de cada variable de análisis)
  - Visualización (gráficos)

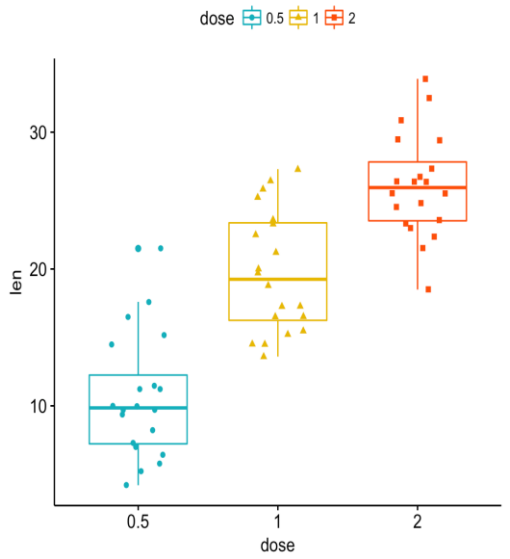
```
# Boxplot (categórica vs continua)
ggboxplot()
# de Dispersión de puntos (continua vs continua ó continua vs discreta)
ggscatter()
# Histograma (continua)
gghistogram()
# de Densidad (continua)
ggdensity()
# de Barras (tablas de frecuencia de una variable categórica)
ggbarplot()
```



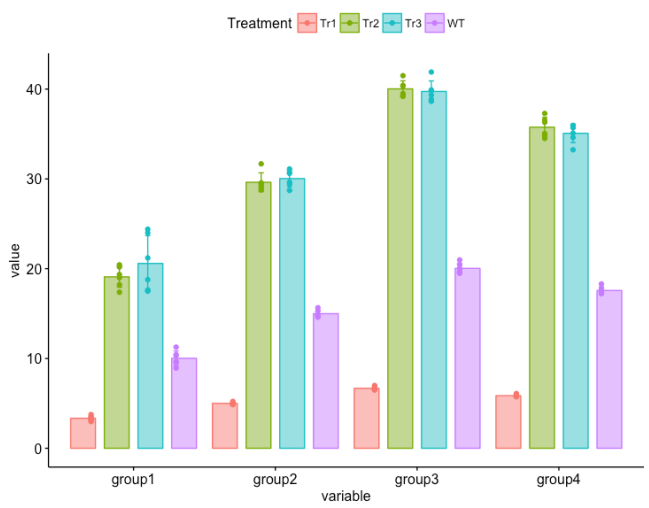
ggdensity()



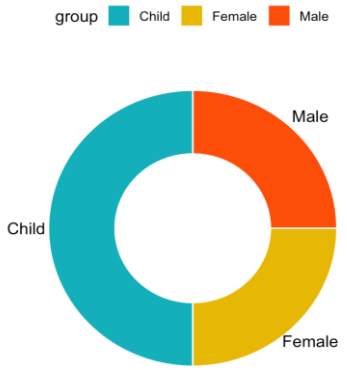
gghistogram()



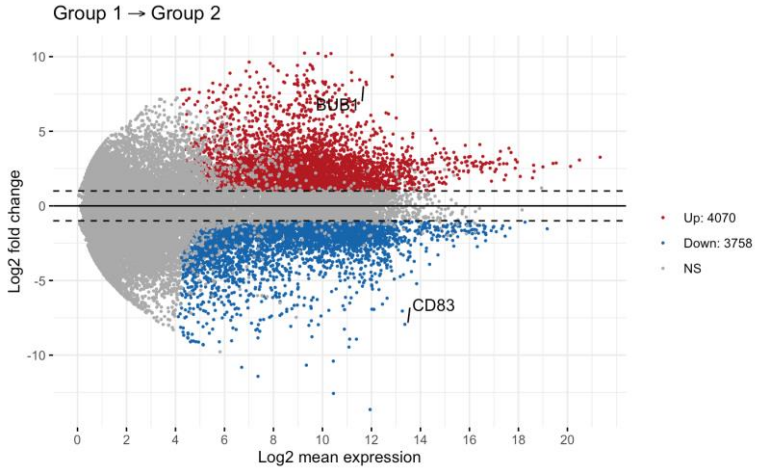
ggboxplot()



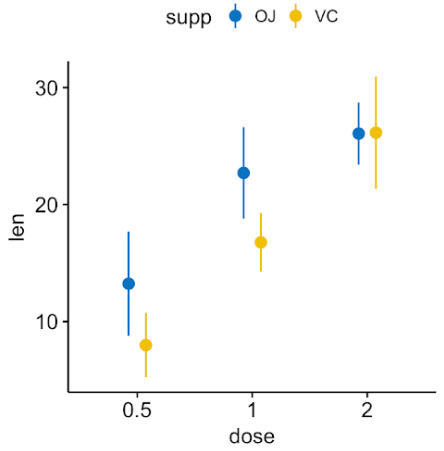
ggbarplot()



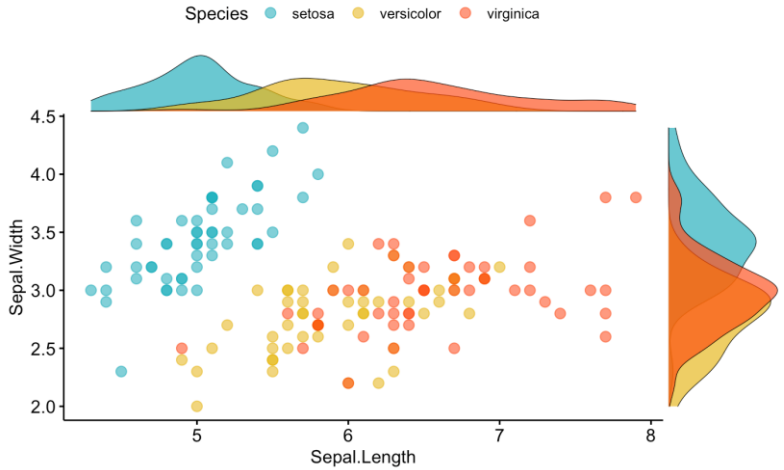
ggdonutchart()



ggmaplot()



ggerrorplot()



ggscatterhist()

Módulo III.

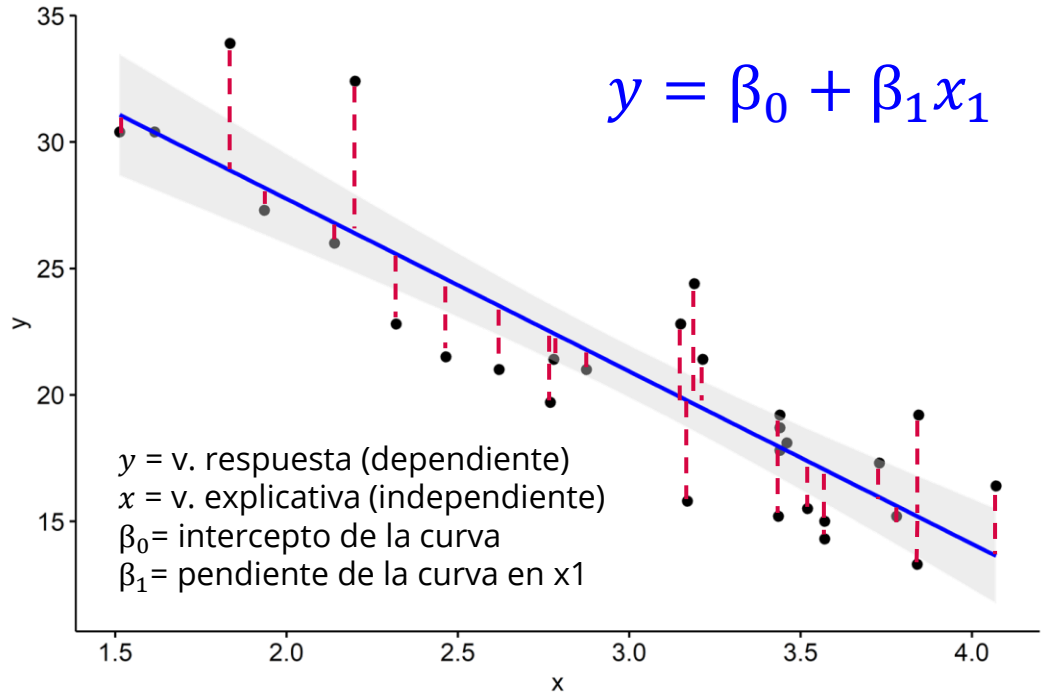
Estadística inferencial

Regresión Lineal OLS

$$\begin{cases} \text{var}[y_i] = {\sigma^2}/{w_i} & \text{.....Comp. aleatorio} \\ \text{E}[y_i] = \mu_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji} & \text{.....Comp. sistemático} \end{cases}$$

- Modelo: abstracción de la realidad.
- Ecuación matemática más sencilla (línea recta) para representar la relación entre variables.

```
# Modelo lineal simple
lm(y ~ x, data=DF)

# Modelo lineal múltiple (efectos
principales aditivos)
lm(y ~ x1 + x2, data=DF)

# Modelo lineal múltiple (efectos
principales aditivos e interacción)
lm(y ~ x1 + x2 + x1:x2, data=DF)
lm(y ~ x1*x2, data=DF)
```



Regresión Lineal OLS

$$\begin{cases} \text{var}[y_i] = {\sigma^2}/{w_i} & \text{.....Comp. aleatorio} \\ \text{E}[y_i] = \mu_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji} & \text{.....Comp. sistemático} \end{cases}$$

- **A1:** La variable respuesta debe ser continua.
  - **OPCIONES**: GLM u otras regresiones avanzadas.
- **A2:** La relación entre X e Y debe ser lineal (ver gráfica de dispersión de puntos).
  - **OPCIONES**: linealizar la relación transformando la variables, regresiones cuadráticas, cubicas, GAM.
- **A3:** Los **residuales** son independientes, es decir, provienen de observaciones independientes.
- **A4:** La variable respuesta no debe tener outliers en ningún nivel de análisis.
- **A5:** Los **residuales** tienen distribución normal.
  - **OPCIONES**: segurar un n muestral grande.
- **A6:** Los **residuales** son homocedásticos.
  - **OPCIONES**: Redefinir la variable para comprimir los valores (dividir: ratios, per capita). Regresión lineal pesada (Weighted Linear Regression). Tranformar la variable respuesta (Y).

Regresión Lineal OLS

$$\begin{cases} \text{var}[y_i] = {\sigma^2}/{w_i} & \text{.....Comp. aleatorio} \\ \text{E}[y_i] = \mu_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji} & \text{.....Comp. sistemático} \end{cases}$$

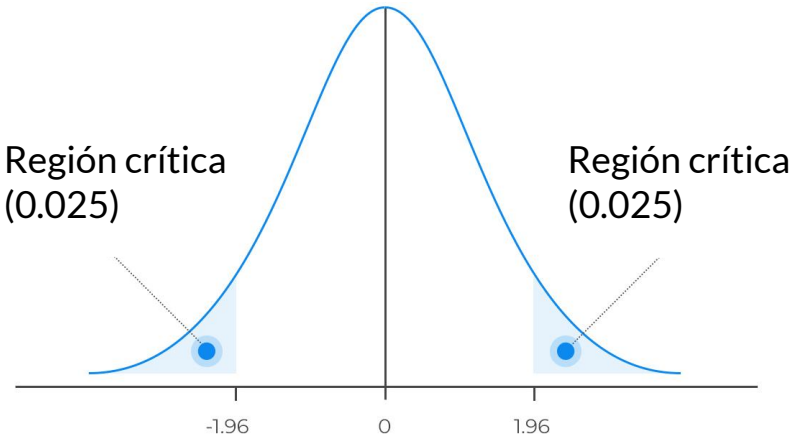
Método	Tipo de Var. Respuesta (Y)	Tipo de Var. Explicativa (X)	Número de Var. Explicativas	Número de Niveles
Regresión Lineal Simple	Continua	Continua	1	
t-Test	Continua	Categórica	1	2
ANOVA	Continua	Categórica	1 (ANOVA una vía), 2 (ANOVA de dos vías), o más	3 o más
ANCOVA	Continua	Continua y Categórica	2 o más	2 o más (si X es categórica)
Regresión Lineal Múltiple	Continua	Continua	2 o más	

Comparaciones pareadas: pruebas de t

- **A1:** Los datos a analizar son resultados de mediciones (valores continuos).
- **A2:** Los datos a analizar fueron obtenidos por un muestreo aleatorio.
- **A3:** Debe haber homogeneidad de varianzas entre los grupos evaluados. No obstante, R lidia con esto aplicando la corrección de Welch cuando var.equal = FALSE.
- **A4:** Los datos de cada grupo tienen distribución normal (Test de Normalidad, Q-Q Plot). A mayor N muestral, se reduce la necesidad de testear esta asunción, siempre y cuando se cumplan las anteriores.

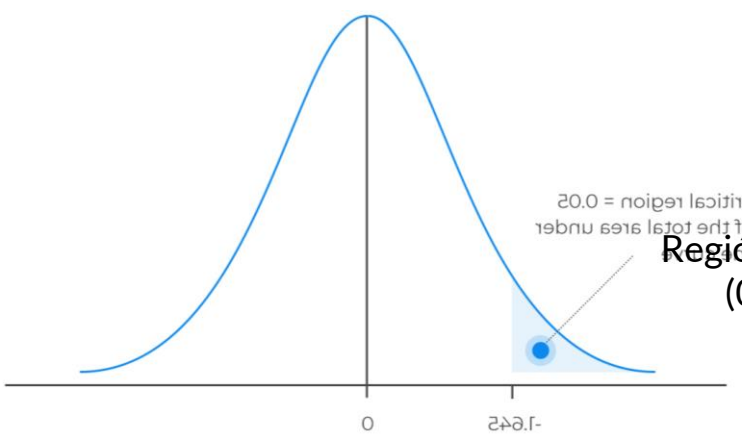
```
# Prueba de T para una muestra
t.test(x, mu=0)
# Prueba de T para dos muestras dependientes
t.test(x, y, paired=TRUE)
# Prueba de para dos muestras independientes con igual varianza
t.test(x, y, var.equal=TRUE)
# Prueba de para dos muestras independientes con varianza desigual
t.test(x, y)
```

Argumento `alternative`



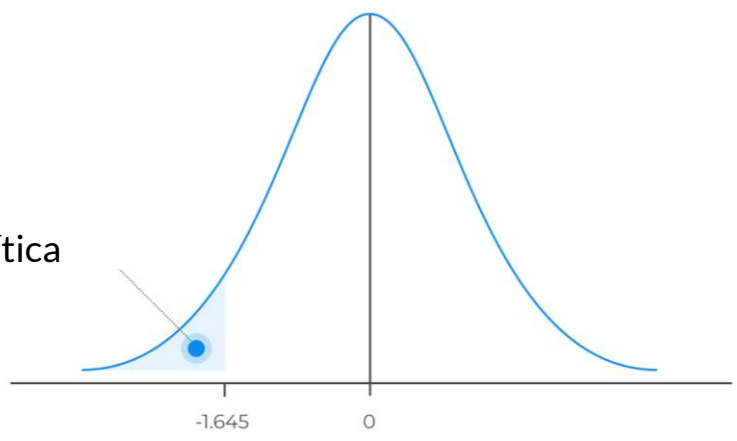
two.sided

¿Existen diferencias significativas entre el promedio del conjunto de dato x e y?



greater

¿El promedio del conjunto de datos x es significativamente mayor que el de y?



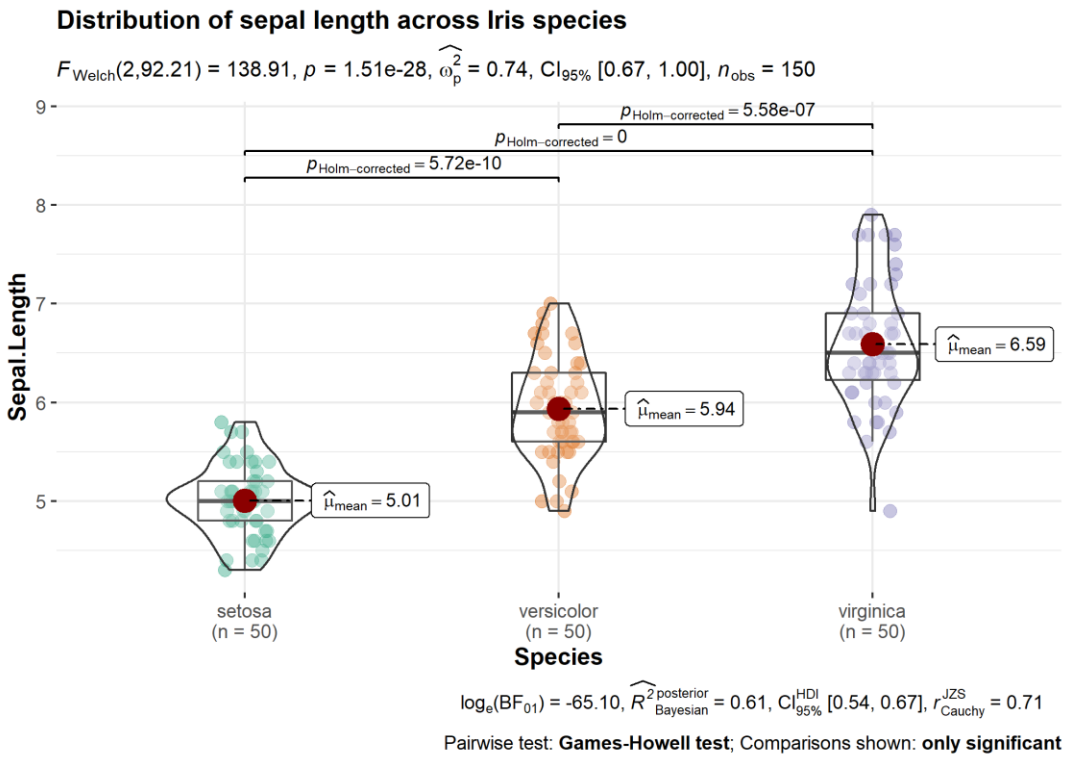
less

¿El promedio del conjunto de datos x es significativamente menor que el de y?

Comparaciones pareadas: anova (1 y 2 vías)

- Modelo lineal en el que la variable(s) explicativa(s) es(son) categóricas (factores).

A1: La variable respuesta debe ser continua
OPCIONES: GLM u otras regresiones avanzadas.
A2: Los **residuales** son independientes, es decir, provienen de observaciones independientes.
OPCIONES: Si las muestras son dependientes, realizar LMM.
A3: La variable respuesta no debe tener outliers en ningún nivel de análisis.
A4: Los **residuales** tienen distribución normal.
OPCIONES: Kruskal-Wallis.
A5: Los **residuales** son homocedásticos.
OPCIONES: Kruskal-Wallis (no paramétrico), o ANOVA con corrección de Welch.




```
# Anova de una vía
aov(lm(y ~ x, data=DF))

# Anova de dos vías
aov(lm(y ~ x1 + x2, data=DF))

# Prueba Post Hoc
tukeyHSD()
```