Manual de Prácticas

Estadística Aplicada

Dr. Miguel Ghebré Ramírez Elías

Profesor-Investigador

Facultad de Ciencias

Universidad Autónoma de San Luis Potosí

UASLP

FACULTAD DE CIENCIAS

Manual de Prácticas de Estadística Aplicada

Dr. Miguel G.Ramírez Elías

Facultad de Ciencias UASLP

Contents

1 Estadística Descriptiva: Variables Cualitativas **3**
Introducción . . . 3
Objetivos de aprendizaje . . . 3
Material . . . 3
Ejemplo 1 Estadísticos y Gráficos para para variables Cualitativas . . . 3
Ejercicios . . . 9
Bibliografía . . . 9

2 Estadística Descriptiva: Variables Cuantitativas **10**
Introducción . . . 10
Objetivos de aprendizaje . . . 12
Material . . . 12
Ejemplo 2 Estadistica Descriptiva: Datos cuantitativos. . . . 12
Ejercicios . . . 20
Bibliografía . . . 20

3. Estimación: Intervalos de Confianza (una muestra) **21**
Introducción . . . 21
Objetivos de aprendizaje . . . 21
Material . . . 21
Ejemplo 3: Consumo de Combustible . . . 22
Ejemplo 4: Concentración en Medicamentos . . . 23
Ejemplo 5: Intervalo de Confianza para Proporciones . . . 25
Ejercicios . . . 26
Bibliografía . . . 26

4 Inferencia Estadística: Pruebas de Hipótesis **27**
Introducción . . . 27
Objetivos de aprendizaje . . . 27
Material . . . 27
Ejemplo 6: Concentración de medicamentos (Contraste de hipótesis) . . . 27
Ejercicios . . . 30
Bilbiografía . . . 31

5 Estimación: Intervalos de Confianza (dos muestras) **32**
Introducción . . . 32
Objetivos de aprendizaje . . . 32
Material . . . 32
Ejemplo 9: Intervalos de confianza para la comparación de medias y proporciones de dos poblaciones 32
Ejercicios: . . . 34
Bibliografía . . . 34

6 Análisis de Varianza (ANOVA) 35
Introducción . . . 35
Objetivos de Aprendizaje . . . 35
Material . . . 36
Ejemplo 11: Pérdida de peso por dieta . . . 36
Ejemplo 12: ANOVA dos factores . . . 41
Ejemplo 13: ANOVA de 2 factores . . . 47
Ejercicios . . . 48
Bibliografía . . . 49

7 Regresión lineal Simple 50
Introducción . . . 50
Objetivos de aprendizaje . . . 50
Material . . . 50
Ejemplo 14: Contenido de alcohol en sangre . . . 51
Ejercicios . . . 60
Bibliografía . . . 60

8 Estadística No Paramétrica 61
Introducción . . . 61
Objetivos de Aprendizaje . . . 61
Material . . . 62
Ejemplo 15: Pruebas de normalidad . . . 62
Ejemplo 16: Pruebas de Normalidad . . . 65
Ejemplo 17: Prueba Chi-cuadrada . . . 68
Ejemplo 18: Prueba U de Mann Whitney . . . 69
Ejemplo 19: Prueba U de Mann Whitney . . . 71
Ejemplo 20: Wilcoxon Signed Rank Test . . . 71
Ejemplo 21: Prueba H de Kruskal-Wallis . . . 72
Ejercicios . . . 73
Bibliografía . . . 73

1 Estadística Descriptiva: Variables Cualitativas

Introducción

La estadística descriptiva es una rama de la estadística que se encarga de recolectar, organizar, presentar y analizar un conjunto de datos para describir las características principales. A diferencia de la estadística inferencial, que busca hacer predicciones o generalizaciones sobre una población a partir de una muestra, la estadística descriptiva se centra en resumir y visualizar los datos de manera clara y concisa. La estadística descriptiva es fundamental en diversas áreas como la investigación científica, la economía, la medicina y la ingeniería, entre otras. Permite a los investigadores y profesionales comprender mejor los datos y tomar decisiones informadas basadas en la evidencia.

Variables cualitativas

Las variables cualitativas, son aquellas que describen cualidades o características y no pueden ser medidas numéricamente. Las variables cualitativas producen datos que se pueden clasificar de acuerdo a similitudes diferencias en clase; por lo tanto, con frecuencia se denominan datos categóricos.Estas variables se dividen en categorías o grupos. Ejemplos comunes incluyen:Color de ojos (azul, verde, marrón), Estado civil (soltero, casado, divorciado), Tipo de sangre (A, B, AB, O).

Objetivos de aprendizaje

Después de completar esta práctica, el estudiante será capaz de:

- Construir una tabla de distribución de frecuencias para variables cualitativas.
- Generar e interpretar Gráficas de Barras y Gráficas de Pastel
- Generar e interpretar gráficas para datos bivariados.

Material

- Computadora con el software R instalado o acceso a R studio cloud (https://posit.cloud/)
- Para la realización de esta práctica se requieren los siguientes paquetes:

```
library(tidyverse)
library(dplyr)
library(ggplot2)
library(knitr)
library(kableExtra)
```

Datos: https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset

Ejemplo 1 Estadísticos y Gráficos para para variables Cualitativas

Análisis estadístico del conjunto de datos: *Framingham heart study dataset*

El conjunto de datos está disponible públicamente en el sitio web de Kaggle y proviene de un estudio cardiovascular en curso sobre residentes de la ciudad de Framingham, Massachusetts. El conjunto de datos incluye más de 4240 registros, 16 columnas y 15 atributos.

1. Leer el archivo framingham.csv

Leemos el archivo mediante la función `read.csv`:

```
df.1 <- read.csv('framingham.csv')
```

El conjunto contiene una columna llamada “male” que tine valor de 0 y 1. Donde 1 significa “male”, es decir, el paciente es Masculino. Para poder realizar el proyecto vamos a convertir la variable “male” a categórica:

```r
# Convertir variable categórica
df.1$sex <- factor(df.1$male, levels = c(0, 1), labels = c("Female", "Male"))
```

2. Generar la Tabla de Frecuencias.

Para generar la tabla de Frecuencias, primero debemos calcular la frecuencia absoluta (ni), la frecuencia relativa (fi), la frecuencia absoluta acumulada (Ni) y la frecuencia relativa acumulada (Fi). Para las absolutas usamos la función `table`y para las relativas usamos `prop.table`:

```r
# Frecuencias absolutas.
ni <- table(df.1$sex)
# Frecuencias relativas
fi <- prop.table(ni)
# Frecuencias acumuladas.
Ni <- cumsum(ni)
# Frecuencias relativas acumuladas.
Fi <- cumsum(fi)
# Creación de un data frame con las frecuencias.
tabla_frec <- cbind(ni, fi, Ni, Fi)
tabla_frec
```

```
##          ni        fi   Ni        Fi
## Female 2420 0.5707547 2420 0.5707547
## Male   1820 0.4292453 4240 1.0000000
```

Otra alternativa para generar la tabla de frecuencias, es usar la función `count` del paquete `dplyr`:

```r
count(df.1, sex) |>
    mutate(fi = n/sum(n), Ni = cumsum(n), Fi = cumsum(n)/sum(n)) |>
    kable() |>
    kable_styling(bootstrap_options = "bordered", full_width = T)
```

sex	n	fi	Ni	Fi
Female	2420	0.5707547	2420	0.5707547
Male	1820	0.4292453	4240	1.0000000

Usando el argumento `bootstrap_options` de la función `kable_styling` podemos cambiar el formato de la tabla (https://www.w3schools.com/bootstrap/bootstrap_tables.asp).

3. Crear una nueva variable llamada "Obesidad" basada en la clasificación del índice de masa corporal (IMC):

Tabla de Clasificación del IMC

IMC	Estado
Por debajo de 18.5	Bajo peso
18.5–24.9	Peso normal
25.0–29.9	Sobrepeso
Por encima de 30.0	Obesidad

Algunas variables cualitativas son ordinales. Además de las categorías ordenadas (p. ej., excelente, muy buena, buena, regular, mala), los investigadores a veces recopilarán información sobre medidas distribuidas de forma continua, pero luego categorizarán estas medidas porque facilita la toma de decisiones.

Agregamos a nuestro conjunto de datos una nueva variable llamada “Obesidad” que contenga las categorías del IMC definidas en la tabla anterior.

```
df.1.1 <- df<-na.omit (df.1) |>
    mutate(Obesidad = cut(BMI, breaks = c(0, 18.5, 24.5, 30, Inf), labels = c("Bajo peso", "Peso Normal
```

Generamos la tabla de distribución de frecuencias:

```
# Frecuencias absolutas.
ni <- table(df.1.1$Obesidad)
# Frecuencias relativas
fi <- prop.table(ni)
# Frecuencias acumuladas.
Ni <- cumsum(ni)
# Frecuencias relativas acumuladas.
Fi <- cumsum(fi)
count(df.1.1, Obesidad) |>
    mutate(fi = n/sum(n), Ni = cumsum(n), Fi = cumsum(n)/sum(n)) |>
    kable() |>
    kable_styling(bootstrap_options = "bordered", full_width = T)
```

Obesidad	n	fi	Ni	Fi
Bajo peso	49	0.0133953	49	0.0133953
Peso Normal	1435	0.3922909	1484	0.4056862
Sobrepeso	1723	0.4710224	3207	0.8767086
Obeso	451	0.1232914	3658	1.0000000

La tabla de distribución de frecuencias generada corresponde a la categorización de valores de IMC como una variable ordinal llamada Obesidad. Las frecuencias, o la cantidad de participantes en cada categoría de respuesta, se muestran en la segunda columna y las frecuencias relativas, como porcentajes, se muestran en la tercer columna. La cuarta columna muestra la frecuencia acumulada y la última columna la frecuencia relativa acumulada.

Las frecuencias acumuladas reflejan el número de pacientes con un nivel particular de IMC o por debajo de él. Por ejemplo, 1706 pacientes tienen bajo peso o peso normal. Hay 3685 pacientes con bajo peso, peso normal, o sobrepeso. Las frecuencias relativas acumuladas son útiles para resumir las variables ordinales e indican la proporción (entre 0 y 1) o el porcentaje (entre 0 % y 100 %) de pacientes con un nivel particular o por debajo de él. En este ejemplo, el 87 % de los pacientes NO están clasificados como Obesos (es decir, tienen bajo peso, peso normal o sobrepeso).

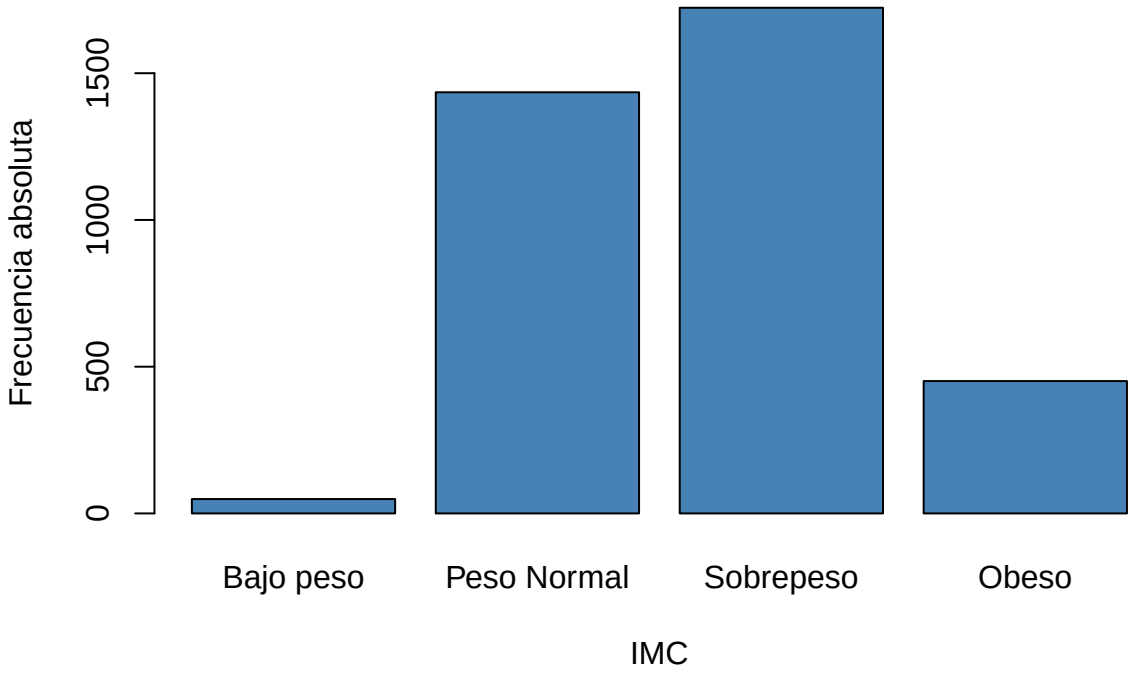
4. Crear una gráfica de barras de frecuencias absolutas y relativas del IMC para cada categoria (bajo peso, peso normal, sobrepeso y obesidad)

Una vez que a las mediciones se les hayan dado categorías y se resumieron en una tabla estadística, se puede usar ya sea una gráfica de pastel o una gráfica de barras para mostrar la distribución de los datos.

Las representaciones gráficas son útiles para resumir datos, y es recomendable representar las variables cualitativas con gráficos de barras. Las opciones de respuesta o categorías (ej: sí/no,masculino/femenino) se muestran en el eje horizontal, y las frecuencias o frecuencias relativas se representan en el eje vertical.

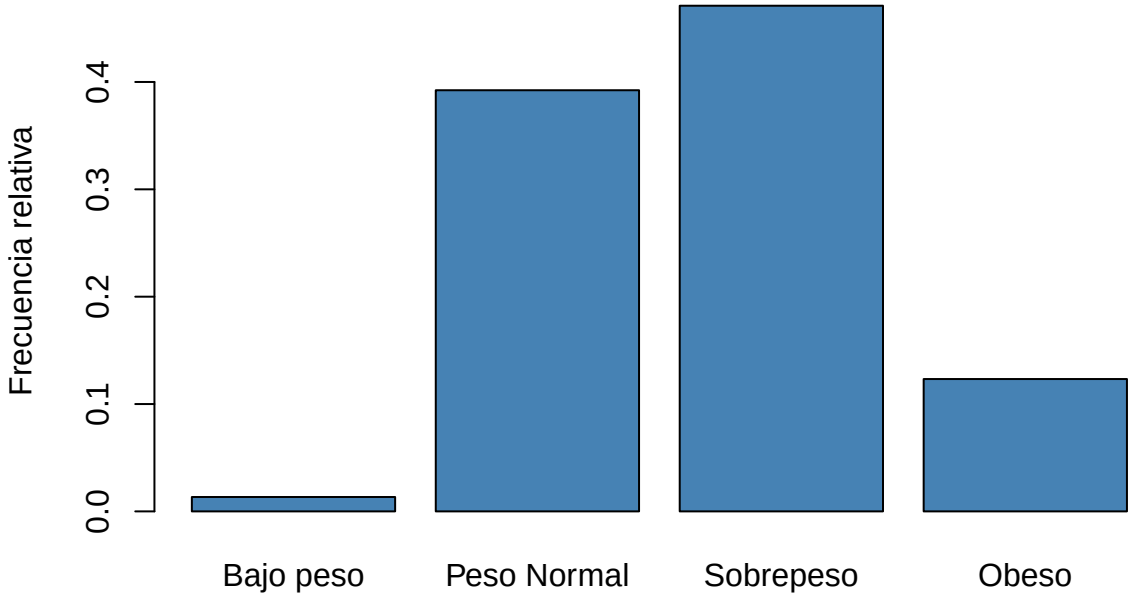
Para dibujar un diagrama de barras se puede usar la función `barplot` del paquete `graphics`:

```
# Diagrama de barras de frecuencias absolutas.
barplot(ni, col = "steelblue", main="", xlab="IMC", ylab="Frecuencia absoluta")
```



Y para la gráfica de barras de frecuencias relativas:

```
# Diagrama de barras de frecuencias relativas.
barplot(fi, col = "steelblue", main="", xlab="", ylab="Frecuencia relativa")
```



5. Generar un gráfico de pastel para la variable categórica Obesidad

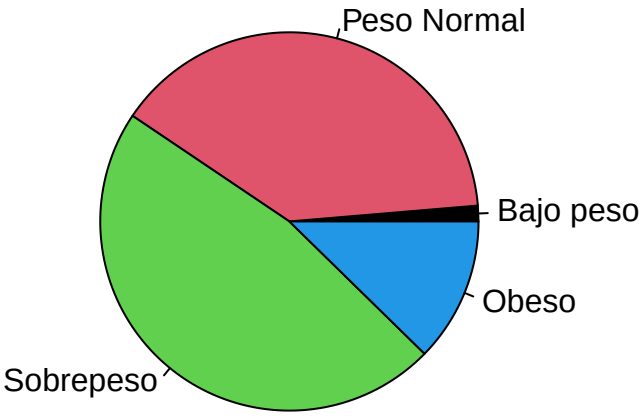
Una gráfica de pastel es la conocida gráfica circular que muestra la forma en que están distribuidas las medidas entre las categorías. La gráfica de pastel se usa para mostrar las relaciones de las partes con respecto al todo (100%).

Para dibujar una gráfica de pastel, se puede usar la función `pie` del paquete `graphics`. También podemos especificar los colores de las porciones del gráfico circular usando `col =`. Además de crear un gráfico visualmente agradable, nos permitirá hacer coincidir el texto de la leyenda con la parte correcta del gráfico circular.

Los colores se especifican después de `col =` usando un vector que contiene la misma cantidad de colores que secciones del gráfico circular.

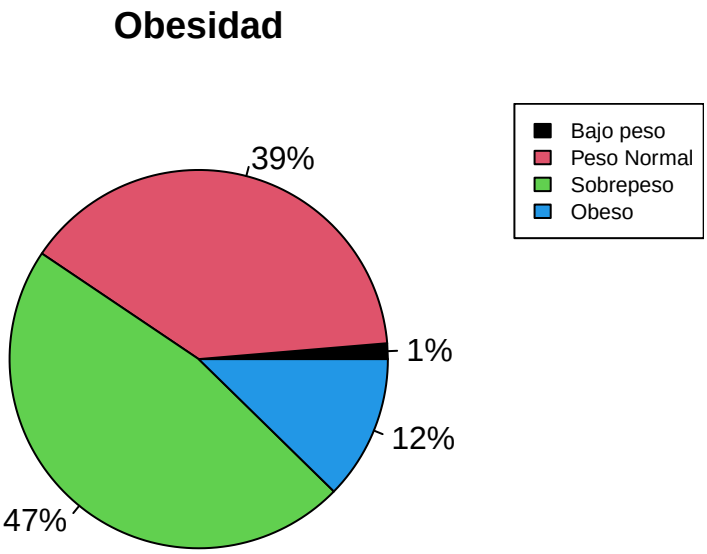
Gráfica de pastel por variable Obesidad:

```
ni.ob <- table(df.1.1$Obesidad)
pie(ni.ob, col = 1:length(ni), main = "")
```



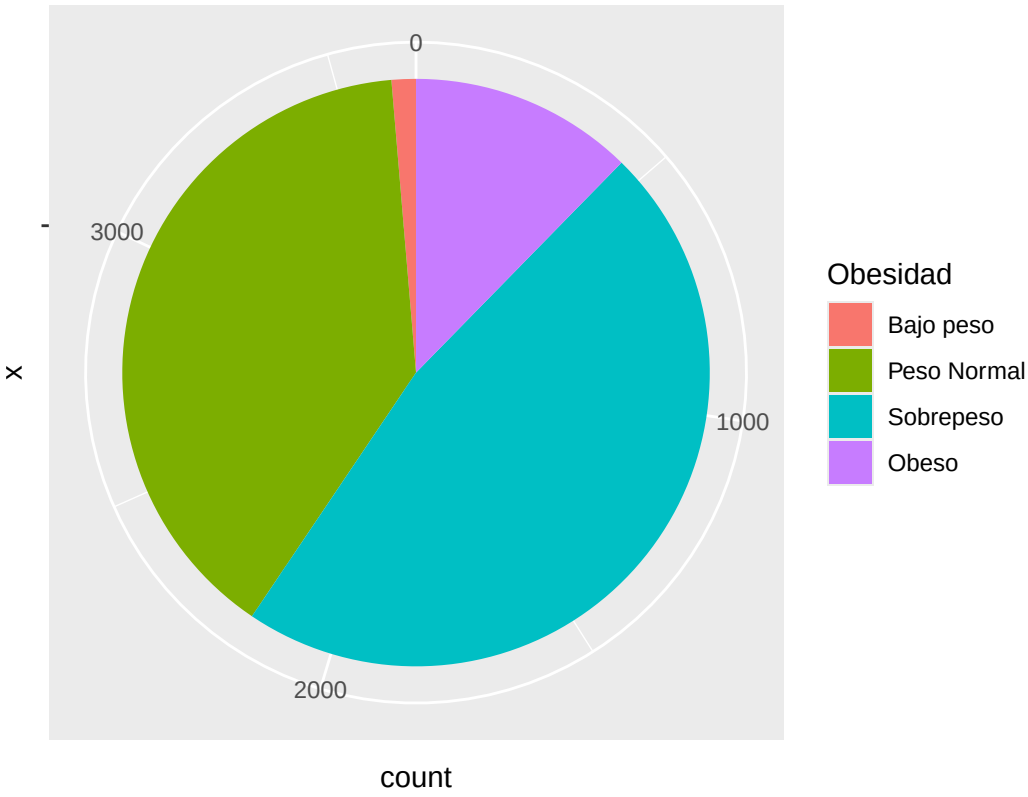
Si bien el gráfico circular nos da una idea general de la frecuencia relativa, resulta útil incluir el porcentaje real de cada categoría en el gráfico. Podemos hacerlo agregando una leyenda:

```
prop.ob<-prop.table(ni.ob)
pie(ni.ob, col = 1:length(ni),
    labels = paste(round(prop.ob*100), "%", sep = ""),main = "Obesidad")
legend("topright", levels(df.1.1$Obesidad), cex=0.7, fill=1:length(ni))
```



Otra alternativa para hace una gráfica de pastel, es usar las funciones `geom_bar` y `coor_polar` del paquete `ggplot2`:

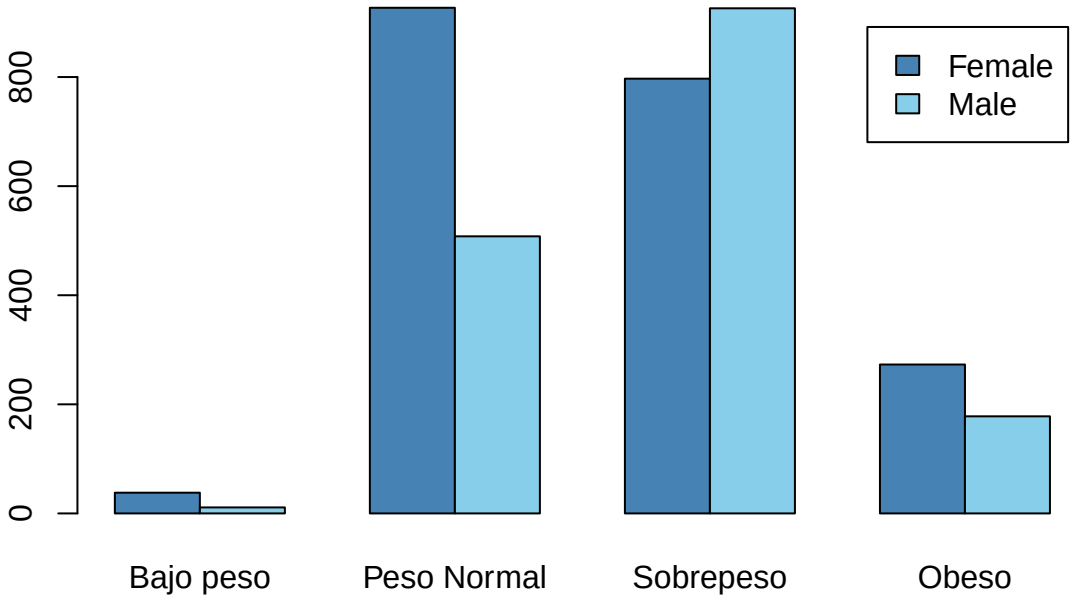
```
ggplot(df.1.1, aes(x = "", fill = Obesidad)) +
    # Añadir la capa de las barras.
    geom_bar() +
    # Añadir el sistema de coordenadas polares
    coord_polar(theta = "y") +
    labs(title = "")
```



6. Generar una gráfica de barras "lado a lado" que compare dos variables (Género y Obesidad)

La función barplot también puede crear gráficos de barras "lado a lado" al comparar la frecuencia de una variable agrupada por otra. La primera variable en la función `table` será la que se graficará en el eje y, y la segunda se utilizará como categorías para crear las barras "apiladas o en paralela"lado a lado". El argumento `beside = TRUE` indica que queremos un gráfico de barras lado a lado, y el argumento `legend.text = TRUE` agregará automáticamente una leyenda al gráfico de barras.

```
barplot(
table(df.1.1$sex, df.1.1$Obesidad),
main="",
beside = TRUE,
legend.text = TRUE,
xlab = "",
ylab = "", col = c("steelblue","skyblue")
)
```



Ejercicios

1.Una empresa desea comprender las preferencias de café de sus empleados para optimizar las compras y mejorar la satisfacción. Se realizó una encuesta a 50 empleados donde se les preguntó su tipo de café preferido (Espresso, Latte, Americano, Cappuccino, Filtrado) y si consumen azúcar en su café (Sí/No).

Datos:Los datos se encuentran en el archivo 01_cafe.xlsx.*(https://github.com/ghebrer82/EstadisticaAplicada/blob/main/01_ex_cafe.xlsx)*

Realice las siguientes actividades:

a) Generar la Tabla de Frecuencias para la variable "Tipo de Café Preferido", incluyendo frecuencias absolutas, frecuencias relativas y porcentajes.
b) Crear una gráfica de barras de frecuencias absolutas y de frecuencias relativas para la variable "Tipo de Café Preferido".
c) Generar un gráfico de pastel (o circular) para la variable "Tipo de Café Preferido".
d) Generar una gráfica de barras "lado a lado" (o agrupada) que compare la variable "Tipo de Café Preferido" con la variable "Consumo de Azúcar". Es decir, mostrar la distribución de tipos de café para quienes consumen azúcar y para quienes no.

Bibliografía

- Estadística, Mario Triola, 12va Edición, Pearson, 2018.
- Probabilidad y Estadística para Ingeniería y Ciencias, Ronald Walpole, Pearson Educación, 2012.

2 Estadística Descriptiva: Variables Cuantitativas

Introducción

La estadística descriptiva es una rama de la estadística que se encarga de recolectar, organizar, presentar y analizar un conjunto de datos para describir las características principales. A diferencia de la estadística inferencial, que busca hacer predicciones o generalizaciones sobre una población a partir de una muestra, la estadística descriptiva se centra en resumir y visualizar los datos de manera clara y concisa. La estadística descriptiva es fundamental en diversas áreas como la investigación científica, la economía, la medicina y la ingeniería, entre otras. Permite a los investigadores y profesionales comprender mejor los datos y tomar decisiones informadas basadas en la evidencia.

Estadísticas descriptivas para variables cuantitativas

Para proporcionar una descripción detallada de los cálculos utilizados para los resúmenes numéricos y gráficos de las variables cuantitativas, usaremos el conjunto de datos de participantes en el Estudio del corazón de Framingham.

Los dos componentes clave de un resumen útil para una variable continua son:

Una descripción del centro o "promedio" de los datos y una indicación de la variabilidad (dispersión) de los datos.

Medidas de centro (tendencia central)

Un paso esencial para explorar los datos es obtener un "valor promedio" para cada característica (variable): una estimación de dónde se ubica la mayor parte de los datos (es decir, su tendencia central).

A primera vista, resumir los datos puede parecer bastante trivial: tome la media de los datos. Si bien la media es fácil de calcular y conveniente de usar, puede que no siempre sea la mejor medida para un valor central.

Media

La estimación más básica de las medidas de centro es la media. La media es la suma de todos los valores dividida por la cantidad de valores. Considere el siguiente conjunto de números: {3 5 1 2}. La media es (3 + 5 + 1 + 2) / 4 = 11 / 4 = 2,75. La fórmula para calcular la media para un conjunto de n valores x1, x2, ..., xn es:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

donde: - $\sum$ denota la suma de todos los números del conjunto. - x_i representa cada número individual del conjunto. - n es la cantidad total de números del conjunto.

N (o n) se refiere a la cantidad total de registros u observaciones. En estadística, se escribe con mayúscula si se refiere a una población y con minúscula si se refiere a una muestra de una población.

Otro tipo de media es la media ponderada, que se calcula multiplicando cada valor de datos xi por un peso especificado por el usuario wi y dividiendo su suma por la suma de los pesos. La fórmula para una media ponderada es:

$$\text{Media ponderada} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

donde: - w_i representa el peso de cada número. - x_i representa cada número individual en el conjunto. - $\sum$ denota la suma. - n es el número total de números en el conjunto.

Hay dos motivaciones principales para usar una media ponderada: - Algunos valores son intrínsecamente más variables que otros, y a las observaciones altamente variables se les da un peso menor. Por ejemplo, si

tomamos el promedio de varios sensores y uno es menos preciso, podríamos reducir el peso de los datos de ese sensor. - Los datos recopilados no representan de manera igualitaria a los diferentes grupos que nos interesa medir. Por ejemplo, debido a cómo se llevó a cabo un experimento en línea, es posible que no tengamos un conjunto de datos que refleje con precisión todos los grupos en la base de usuarios. Para corregir eso, podemos dar un peso mayor a los valores de los grupos subrepresentados.

Mediana

La mediana es el número del medio en una lista ordenada de datos. Supongamos que hay un número par de valores de datos. En ese caso, el valor del medio no está realmente en el conjunto de datos, sino que es el promedio de los dos valores que dividen los datos ordenados en mitades superior e inferior. En comparación con la media, que utiliza todas las observaciones, la mediana depende únicamente de los valores en el centro de los datos ordenados. Si bien esto puede parecer una desventaja, ya que la media es mucho más sensible a los datos, hay muchos casos en los que la mediana es una mejor métrica para la tendencia central.

La mediana es una estimación sólida de la tendencia central, ya que no está influenciada por valores atípicos (casos extremos) que podrían sesgar los resultados y como ocurre con la media.

Valores atípicos

Un valor atípico es cualquier valor que esté muy alejado de los demás valores de un conjunto de datos. El hecho de que sea un valor atípico no hace que un valor de datos sea inválido o erróneo. Sin embargo, los valores atípicos suelen ser el resultado de errores en los datos, como la mezcla de datos de unidades diferentes o lecturas erróneas del instrumento. Cuando los valores atípicos son el resultado de datos insuficientes, la media dará como resultado una estimación de tendencia central deficiente, mientras que la mediana seguirá siendo válida. En cualquier caso, los valores atípicos deben identificarse y, por lo general, vale la pena investigarlos más a fondo.

Medidas de Variabilidad

La variabilidad o dispersión, mide si los valores de los datos están muy agrupados o dispersos. Los conjuntos de datos pueden tener el mismo centro pero con aspecto diferente por la forma en que los números se dispersan desde el centro.

Desviación estándar

Las estimaciones de variabilidad más utilizadas se basan en las diferencias, o desviaciones, entre la estimación del centro y los datos observados. Para un conjunto de datos {1, 4, 4}, la media es 3 y la mediana es 4. Las desviaciones con respecto a la media son las diferencias: 1 – 3 = –2, 4 – 3 = 1, 4 – 3 = 1. Estas desviaciones nos indican cuán dispersos están los datos alrededor del valor central.

Las estimaciones de variabilidad más conocidas son la varianza y la desviación estándar, basadas en las desviaciones al cuadrado. La varianza es un promedio de las desviaciones al cuadrado y la desviación estándar es la raíz cuadrada de la varianza:

Varianza:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

Desviación estándar:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$$

La desviación estándar es mucho más fácil de interpretar que la varianza, ya que está en la misma escala que los datos originales.

Un enfoque diferente para estimar la dispersión se basa en la dispersión de los datos ordenados. Las estadísticas basadas en datos ordenados (clasificados) se denominan estadísticas de orden. La medida más básica

es el rango: la diferencia entre el números mayor y el menor. Los valores mínimo y máximo en sí mismos son útiles para identificar valores atípicos, pero el rango es susceptible a los valores atípicos y no es muy útil como una medida general de dispersión en los datos.

Cuando un conjunto de datos tiene valores atípicos o extremos, resumimos un valor típico utilizando la mediana en lugar de la media. Cuando un conjunto de datos tiene valores atípicos, la variabilidad suele resumirse mediante una estadística denominada rango intercuartil, que es la diferencia entre el primer y el tercer cuartil. El primer cuartil, denominado Q1, es el valor del conjunto de datos que contiene el 25 % de los valores por debajo de él. El tercer cuartil, denominado Q3, es el valor del conjunto de datos que contiene el 25 % de los valores por encima de él. Los cuartiles se pueden determinar siguiendo el mismo enfoque que utilizamos para determinar la mediana, pero ahora consideramos cada mitad del conjunto de datos por separado.

Una medida de variabilidad estándar es la diferencia entre el percentil 25 y el 75, denominada rango intercuartil (IQR o RIQ). A continuación, se muestra un ejemplo sencillo: {3,1,5,3,6,7,2,9}. Los ordenamos para obtener {1,2,3,3,5,6,7,9}. El percentil 25 está en 2.5 y el percentil 75 está en 6.5, por lo que el rango intercuartil es 6.5 – 2.5 = 4.

Objetivos de aprendizaje

Después de completar este proyecto, el estudiante será capaz de:

- Calcular la medidas de centro
- Calcular las medidas de variabilidad
- Generar e interpretar histograma y gráficas de caja.
- Identificar atípicos

Material

- Computadora con el software R instalado o acceso a R Studio Cloud (https://posit.cloud/)
- Para la realización de esta práctica se requieren los siguientes paquetes:

```
library(tidyverse)
library(vtable)
library(skimr)
library(summarytools)
library(knitr)
library(kableExtra)
library(readxl)
```

Datos: El conjunto de datos de enfermedades cardíacas *"Framingham"* incluye más de 4240 registros, 16 columnas y 15 atributos.

Fuente: https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset

Ejemplo 2 Estadistica Descriptiva: Datos cuantitativos.

Análisis estadístico del conjunto de datos: *Framingham heart study dataset*

El conjunto de datos está disponible públicamente en el sitio web de Kaggle y proviene de un estudio cardiovascular en curso sobre residentes de la ciudad de Framingham, Massachusetts. El conjunto de datos incluye más de 4240 registros, 16 columnas y 15 atributos.

Fuente: https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset

1. Calcular la media, mediana y moda de la presion diastólica (PAD)

Para calcular la media y la mediana usamos las funciones `meany` `median`. No existe una función de R que calcule directamente la moda, pero podemos definir la función:

```r
df.2 <- read.csv('framingham.csv')

media<-mean(df.2$diaBP, na.rm=T)
mediana<-median(df.2$diaBP, na.rm=T)


moda <- function(x) {
u <- unique(x) # Vector con los valores de la muestra sin repetir (sin ordenar).
tab <- tabulate(match(x, u)) # Frecuencias absolutas de los valores en u.
u[tab == max(tab)] # Valor con la mayor frecuencia.
}
moda<-moda(df.2$diaBP)

print(paste("Media:",round(media,2),sep=""))
```

```
## [1] "Media:82.9"
```

```r
print(paste("Mediana:",mediana,sep=""))
```

```
## [1] "Mediana:82"
```

```r
print(paste("Moda:",moda,sep=""))
```

```
## [1] "Moda:80"
```

2. Calcular Cuartiles y percentiles de la presion diastólica (PAD)

Para calcular cuartiles y percentiles usamos la función `quantile`

```r
quantile(df.2$diaBP, prob=c(0.25, 0.5, 0.75), na.rm=T)
```

```
## 25% 50% 75%
##  75  82  90
```

```r
quantile(df.2$diaBP, prob=c(0.05, 0.95),na.rm=T)
```

```
##      5%     95%
##  66.000 104.525
```

3. Calcular el rango, el rango IQR, la varianza, la desviación estándar, y el coefciente de variación del IMC.

```r
Rango<-max(df.2$diaBP,na.rm=T) - min(df.2$age,na.rm=T)
Rango.IQR<-IQR(df.2$diaBP,na.rm=T)
VAR<-var(df.2$diaBP,na.rm=T)
SD<-sd(df.2$diaBP,na.rm=T)
CV<-(sd(df.2$diaBP,na.rm=T) / abs(mean(df.2$age,na.rm=T)))


print(paste("Rango:",Rango,sep=""))
```

```
## [1] "Rango:110.5"
```

```r
print(paste("IQR:",Rango.IQR,sep=""))
```

```
## [1] "IQR:15"
```

```r
print(paste("Varianza:",round(VAR,2),sep=""))
```

```
## [1] "Varianza:141.86"
```

```r
print(paste("Desv.Estándar:",round(SD,2),sep=""))
```

```
## [1] "Desv.Estándar:11.91"
```

```r
print(paste("Coef. Varición:",round(CV,2),sep=""))
```

```
## [1] "Coef. Varición:0.24"
```

4. Calcular el tamaño muestral según el género.

El conjunto contiene una columna llamada "male" que tiene valor de 0 y 1. Donde 1 significa "male", es decir, el paciente es masculino. Para poder realizar el proyecto vamos a convertir la variable "male" a categórica:

```r
# Convertir variable categórica
df.2$sex <- factor(df.2$male, levels = c(0, 1), labels = c("Femenino", "Masculino"))
```

```r
table(df.2$sex)
```

```
##
##  Femenino Masculino
##      2420      1820
```

5. Realizar un resumen estadístico con la media, el mínimo, los cuartiles y el máximo de las siguites variables: Presion diastólica (diaBP), presión sistolica (sysBP), colesterol total (totChol), IMC (BMI).

```r
column_list <- c("diaBP","sysBP", "totChol", "BMI")
df.2.select<- df.2 %>% select(all_of(column_list))
```

método 1: Usando el paquete base de R.

```r
summary(df.2.select)
```

```
##      diaBP           sysBP          totChol           BMI
##  Min.   : 48.0   Min.   : 83.5   Min.   :107.0   Min.   :15.54
##  1st Qu.: 75.0   1st Qu.:117.0   1st Qu.:206.0   1st Qu.:23.07
##  Median : 82.0   Median :128.0   Median :234.0   Median :25.40
##  Mean   : 82.9   Mean   :132.4   Mean   :236.7   Mean   :25.80
##  3rd Qu.: 90.0   3rd Qu.:144.0   3rd Qu.:263.0   3rd Qu.:28.04
##  Max.   :142.5   Max.   :295.0   Max.   :696.0   Max.   :56.80
##                                  NA's   :50      NA's   :19
```

Alternativamente, usando las funciones `descr` del paquete `summarytools`.

```r
library(summarytools)
descr(df.2.select)
```

```
## Descriptive Statistics
## df.2.select
## N: 4240
##
##                        BMI     diaBP     sysBP   totChol
## ----------------- --------- --------- --------- ---------
##              Mean     25.80     82.90    132.35    236.70
##           Std.Dev      4.08     11.91     22.03     44.59
```

```
##                Min      15.54     48.00     83.50    107.00
##                 Q1      23.07     75.00    117.00    206.00
##             Median      25.40     82.00    128.00    234.00
##                 Q3      28.04     90.00    144.00    263.00
##                Max      56.80    142.50    295.00    696.00
##                MAD       3.69     11.12     19.27     43.00
##                IQR       4.97     15.00     27.00     57.00
##                 CV       0.16      0.14      0.17      0.19
##           Skewness       0.98      0.71      1.14      0.87
##       SE.Skewness       0.04      0.04      0.04      0.04
##           Kurtosis       2.65      1.27      2.15      4.12
##            N.Valid    4221.00   4240.00   4240.00   4190.00
##          Pct.Valid      99.55    100.00    100.00     98.82
```

Y luego dar formato a la tabla con `kable`

```
descr(df.2.select) |>
kable(col.names = c("IMC","Presion diastólica", "Presión sistólica", "Colesterol total")) |>
kable_styling()
```

	IMC	Presion diastólica	Presión sistólica	Colesterol total
Mean	25.8008008	82.8977594	132.3545991	236.6995227
Std.Dev	4.0798402	11.9103945	22.0332996	44.5912839
Min	15.5400000	48.0000000	83.5000000	107.0000000
Q1	23.0700000	75.0000000	117.0000000	206.0000000
Median	25.4000000	82.0000000	128.0000000	234.0000000
Q3	28.0400000	90.0000000	144.0000000	263.0000000
Max	56.8000000	142.5000000	295.0000000	696.0000000
MAD	3.6916740	11.1195000	19.2738000	42.9954000
IQR	4.9700000	15.0000000	27.0000000	57.0000000
CV	0.1581284	0.1436757	0.1664717	0.1883877
Skewness	0.9814853	0.7127456	1.1444748	0.8712564
SE.Skewness	0.0376889	0.0376044	0.0376044	0.0378280
Kurtosis	2.6500637	1.2703811	2.1502363	4.1201313
N.Valid	4221.0000000	4240.0000000	4240.0000000	4190.0000000
Pct.Valid	99.5518868	100.0000000	100.0000000	98.8207547

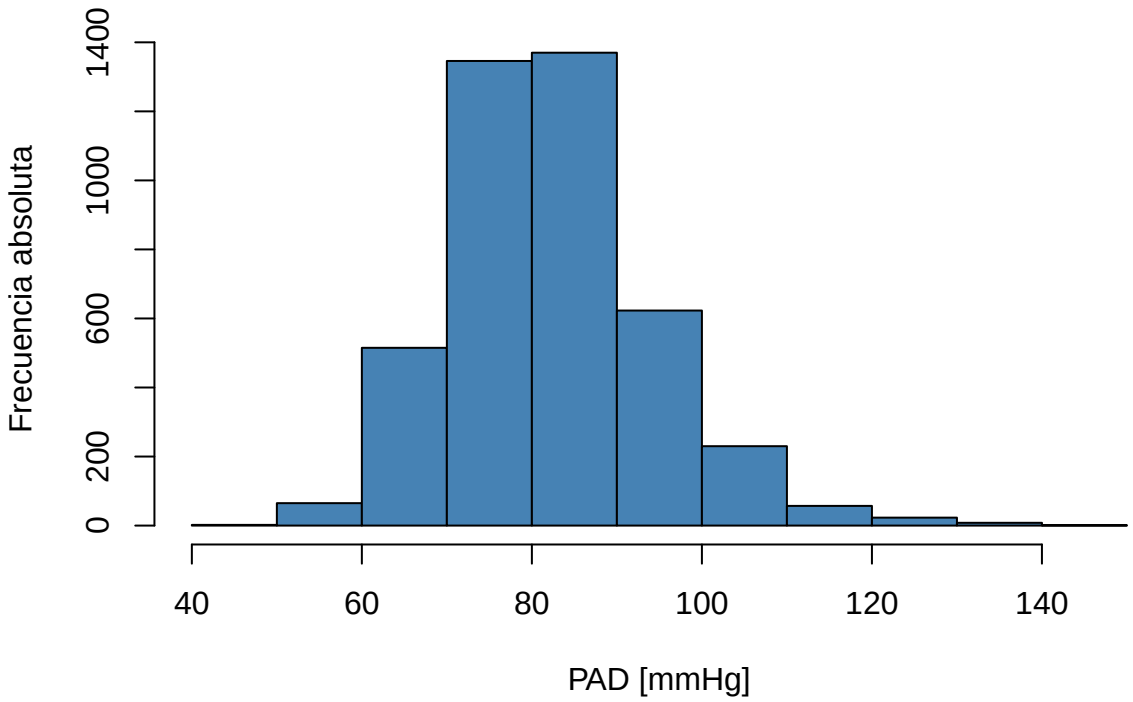
Al igual que con las variables categóricas, el primer objetivo de las estadísticas descriptivas para datos cuantitativos es comprender la distribución de los posibles valores de la variable y la frecuencia con la que se producen. Los datos cuantitativos se pueden obtener de dos maneras: gráficamente con histogramas y diagramas de caja y mediante resúmenes numéricos.

5 Graficar el histograma de frecuencias absolutas, relativas, absolutas acumuladas y relativas acumuladas correspondiente a la Presión diastólica (PAD).

Un histograma de frecuencia relativa es semejante a una gráfica de barras, pero se usa para graficar cantidades en lugar de datos cualitativos. Se puede usar un histograma de frecuencia relativa para describir la distribución de un conjunto de datos en términos de su ubicación y forma, y ver si hay resultados atípicos.

Para graficar el histograma usamos la función `hist`

```
# Histograma de frecuencias absolutas.
histo <- hist(df.2$diaBP, breaks = seq(40, 150, 10), col = "steelblue", main="", xlab="PAD [mmHg]", ylab
```



Los histogramas dividen los datos en intervalos y luego cuentan cuántos puntos de datos caen en cada intervalo. El argumento `breaks=` permite controlar cómo se definen estos intervalos. En la gráfica generada, hemos indicado que el histograma vaya de los valores de PAD de 40 a 150 mmHg, y el ancho de intervalo sea de 10 mmHg.

7. Generar un Diagrama de cajas y bigotes (boxplot) para la varianle PAD

Una representación gráfica popular para una variable continua es un diagrama de caja y bigotes. Los valores extremos o atípicos también se pueden evaluar gráficamente con diagramas de caja y bigotes. Para los datos de participantes de Framingham que consideramos anteriormente, calculamos las siguientes estadísticas resumidas sobre la presión arterial diastólica:

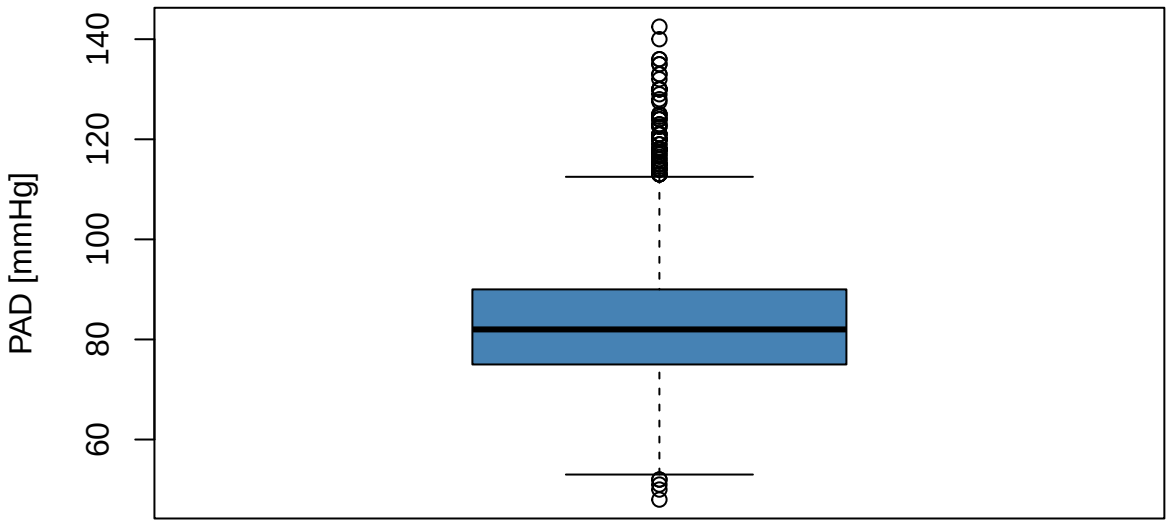
```
summary(df.2$diaBP)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    48.0    75.0    82.0    82.9    90.0   142.5
```

A veces se los denomina cuantiles o percentiles de la distribución. Un cuartil o percentil específico es un valor en el conjunto de datos que contiene un porcentaje particular de los valores que lo igualan o lo dejan por debajo. El primer cuartil, por ejemplo, es el percentil 25, lo que significa que tiene el 25 % de los valores que lo igualan o son inferiores. La mediana es el percentil 50, el tercer cuartil es el percentil 75 y el máximo es el percentil 100 (es decir, el 100 % de los valores lo igualan o son inferiores.

Para dibujar un diagrama de cajas se puede usar la función `boxplot` del paquete `graphics`.

```
boxplot(df.2$diaBP, col = "steelblue", main="", horizontal = F, ylab="PAD [mmHg]")
```

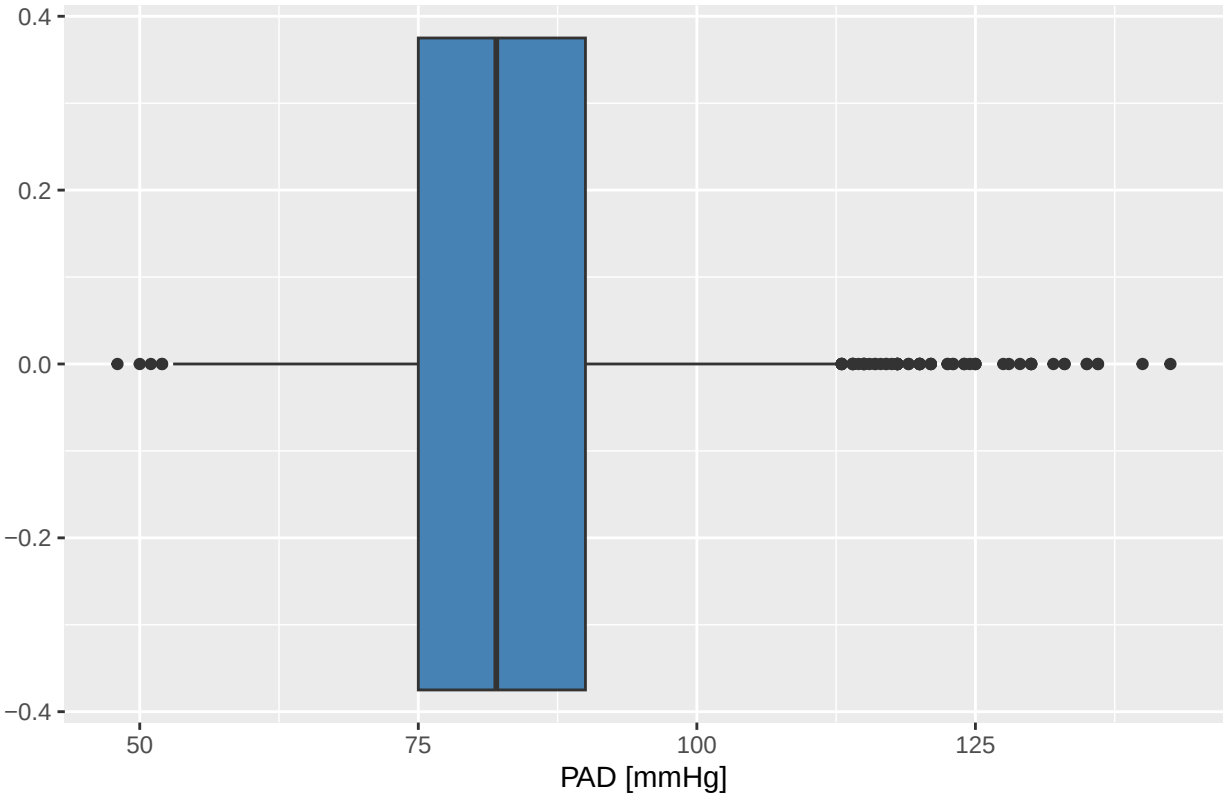


La función `boxplot` muestra la distribución de una variable mediante un diagrama de cajas. Vemos que se utilizan las mismas instrucciones para especificar el título del gráfico y la etiqueta del eje x. Note que el uso de `horizontal = TRUE` orienta el diagrama de caja horizontalmente en lugar de verticalmente.

La Figura generada, es un diagrama de caja y bigotes de las presiones arteriales diastólicas medidas. Las líneas horizontales representan (desde arriba) el máximo, el tercer cuartil, la mediana (también indicada por el punto), el primer cuartil y el mínimo. El cuadro sombreado representa el 50 % de la distribución (entre el primer y el tercer cuartil). Un gráfico de caja y bigotes tiene como objetivo transmitir la distribución de una variable a simple vista. Los valores atípicos se muestran como círculos en la parte superior e inferior de la distribución.

Otra alternativa para generar una gráfica de cajas, es usar la función la función `geom_boxplot` del paquete `ggplot2`.

```
ggplot(df.2, aes(x = diaBP)) +
    geom_boxplot(fill = "steelblue") +
    labs(title = "", x = "PAD [mmHg]")
```



8. Generar un diagrama de cajas para comparar las distribuciones de la variable PAD para los diferentes estados de Obesidad.

Los diagramas de caja y bigotes son muy útiles para comparar distribuciones. La siguiente figura, muestra diagramas de caja y bigotes en paralelo de las distribuciones de PAD, para estado de Obesidad en el Estudio de Framingham.

Primero, agregamos a nuestro conjunto de datos una nueva variable llamada "Obesidad" que contenga las categorías del IMC definidas en la tabla que se nuestra a continuación:.

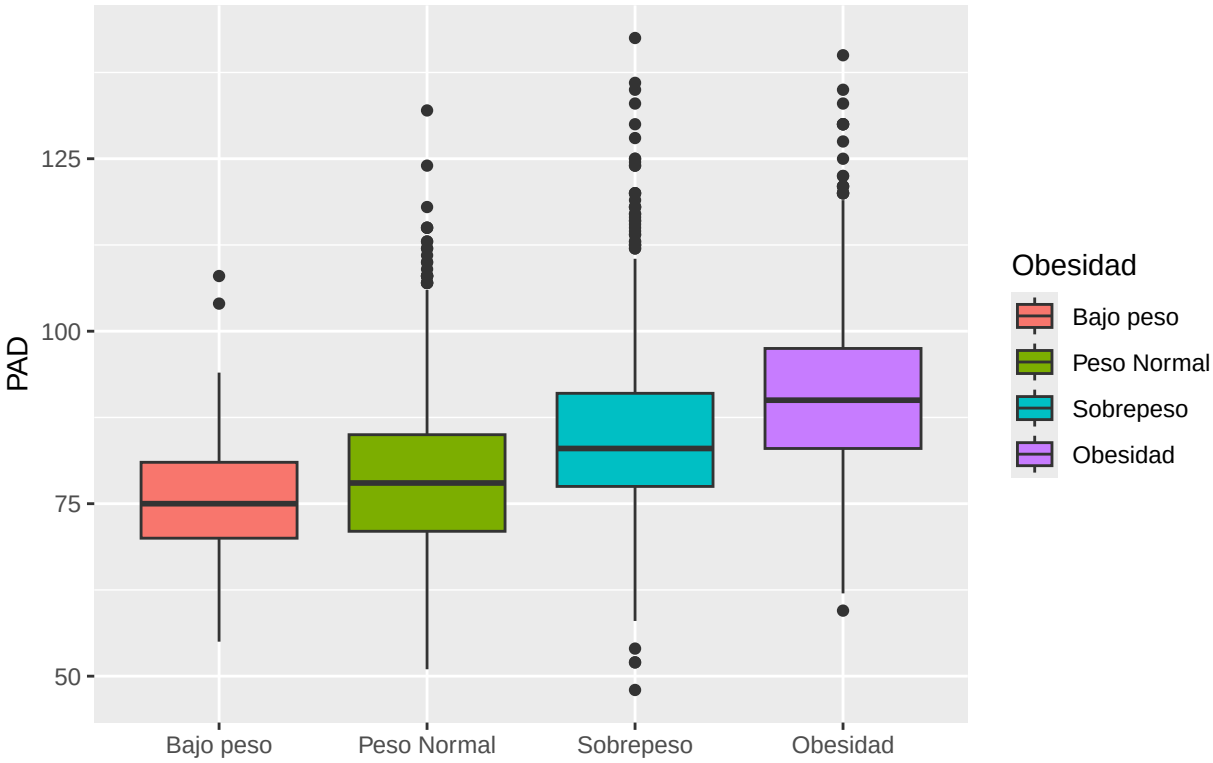
`Tabla de Clasificación del IMC`

IMC	Estado
Por debajo de 18.5	Bajo peso
18.5–24.9	Peso normal
25.0–29.9	Sobrepeso
Por encima de 30.0	Obesidad

```
df.2.ob <- na.omit (df.2) |>
    mutate(Obesidad = cut(BMI, breaks = c(0, 18.5, 24.5, 30, Inf), labels = c("Bajo peso", "Peso Normal
```

Luego, generamos el diagrama de cajas y bigotes para los grupos de Obesidad ("Bajo peso", "Peso Normal", "Sobrepeso", "Obesidad") usando `ggplot`:

```
ggplot(df.2.ob, aes(x=Obesidad, y = diaBP, fill = Obesidad)) +
    geom_boxplot()+
   labs(title = "", x = "", y="PAD")
```

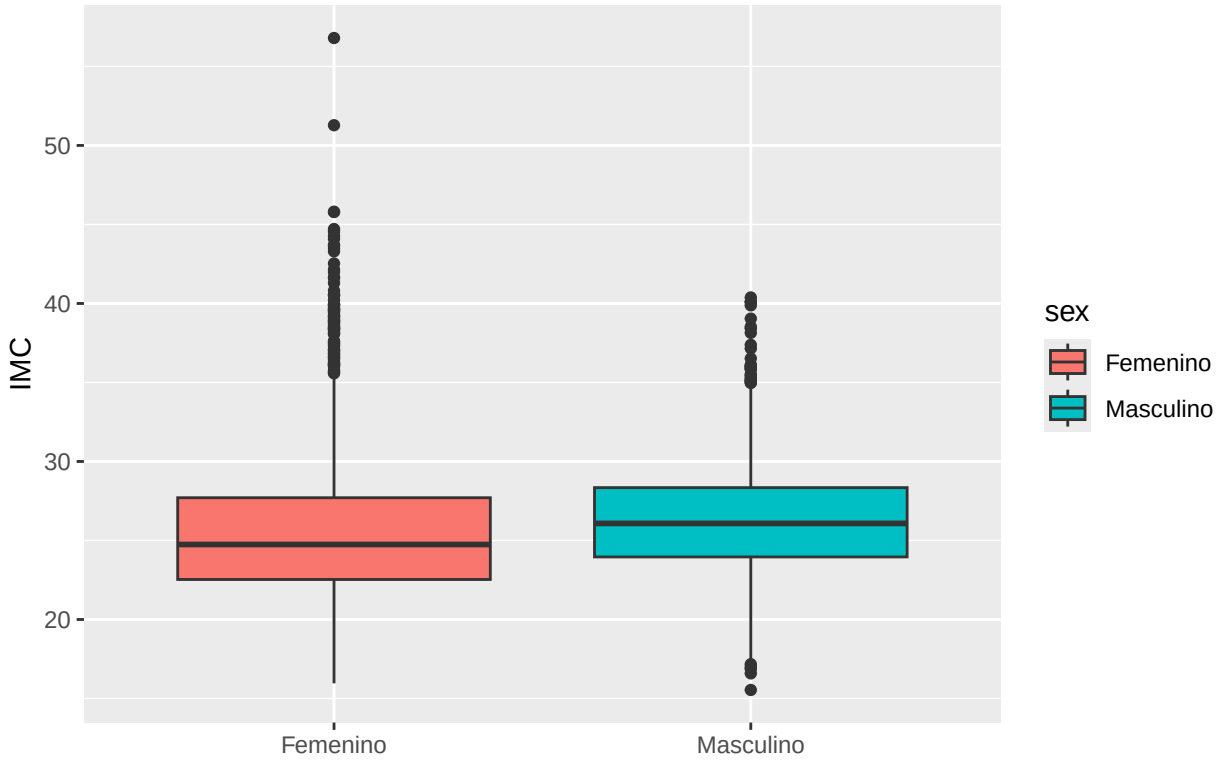


La figura muestra un cambio en la distribución, ya que los pacientes obesos tienen PAD mucho más altas. Se puede observar por ejemplo, que existen valores atípicos entre pacientes de bajo peso en el extremo superior de la distribución.

9. Generar un diagrama de cajas para comparar las distribuciones del IMC por género.

De manera similar al punto anterior, podemos comparar el IMC por género (masculino, femenino) mediante un diagrama de cajas y bigotes:

```
ggplot(df.2
       , aes(x=sex, y = BMI, fill = sex)) +
    geom_boxplot()+
   labs(title = "", x = "", y="IMC")
```



Podemos observar que las distribuciones del índice de masa corporal son similares para hombres y mujeres. Se observa también, que hay muchos valores atípicos en las distribuciones tanto de hombres como de mujeres.

En los diagramas de caja y bigotes, los valores atípicos son valores que superan Q3 + 1.5 IQR o están por debajo de Q1- 1.5 IQR.

Ejercicios

1.Una universidad desea analizar el rendimiento académico de sus estudiantes en una asignatura clave. Para ello, se ha recopilado la calificación final (en una escala de 0 a 100) de 60 estudiantes. Adicionalmente, se tiene el género de cada estudiante (Masculino/Femenino) para un análisis comparativo.

Datos: Los datos se encuentran en el archivo 02_calificaciones.xlsx *(https://github.com/ghebrer82/EstadisticaAplicada/blob/main/02_calificaciones.xlsx)*

a) Calcular la media, mediana y moda de la variable "Calificación_Final".
b) Calcular los cuartiles (Q1, Q2, Q3) y el percentil 90 de la variable "Calificación_Final".
c) Calcular el rango, el rango intercuartílico (IQR), la varianza, la desviación estándar y el coeficiente de variación de la variable "Calificación_Final".
d) Graficar el histograma de la variable "Calificación_Final", mostrando las frecuencias absolutas, relativas, absolutas acumuladas y relativas acumuladas.
e) Generar un Diagrama de cajas y bigotes (boxplot) para la variable "Calificación_Final".
f) Generar un diagrama de cajas para comparar las distribuciones de la "Calificación_Final" por "Género".

Bibliografía

- Estadística, Mario Triola, 12va Edición, Pearson, 2018.
- Probabilidad y Estadística para Ingeniería y Ciencias, Ronald Walpole, Pearson Educación, 2012.

3. Estimación: Intervalos de Confianza (una muestra)

Introducción

Un objetivo clave en la estadística aplicada es hacer inferencias sobre parámetros poblacionales desconocidos a partir de estadísticas de muestra. Existen dos grandes áreas de inferencia estadística: estimación y prueba de hipótesis. La estimación es el proceso de determinar un valor probable para un parámetro poblacional (por ejemplo, la media o proporción poblacional real) a partir de una muestra aleatoria. En la práctica, seleccionamos una muestra de la población objetivo y utilizamos estadísticas de muestra (por ejemplo, la media o proporción de la muestra) como estimaciones del parámetro desconocido. La muestra debe ser representativa de la población, con participantes seleccionados al azar de la población. Al generar estimaciones, también es importante cuantificar la precisión de las estimaciones a partir de diferentes muestras.

Intervalos de confianza

Existen dos tipos de estimaciones para cada parámetro de población: la estimación puntual y la estimación del intervalo de confianza (IC). Primero se calcula la estimación puntual a partir de una muestra.

La estimación del intervalo de confianza (IC) es un rango de valores probables para el parámetro de población basado en:

- la estimación puntual (por ejemplo, la media de la muestra)
- el nivel de confianza deseado por el investigador (más comúnmente 95%, pero se puede seleccionar cualquier nivel entre 0 y 100%)
- y la variabilidad del muestreo o el error estándar de la estimación puntual.

En términos prácticos, un intervalo de confianza del 95% significa que si tomáramos 100 muestras diferentes y calculáramos un intervalo de confianza del 95% para cada muestra, aproximadamente 95 de los 100 intervalos de confianza contendrán el valor medio real ( ). Sin embargo, en la práctica, seleccionamos una muestra aleatoria y generamos un intervalo de confianza, que puede contener o no la media verdadera. El intervalo observado puede sobrestimar o subestimar . En consecuencia, el IC del 95% es el rango probable del parámetro verdadero, desconocido. El intervalo de confianza no refleja la variabilidad del parámetro desconocido. Más bien, refleja la cantidad de error aleatorio en la muestra y proporciona un rango de valores que es probable que incluyan el parámetro desconocido. Otra forma de pensar en un intervalo de confianza es que es el rango de valores probables del parámetro (definido como la estimación puntual + margen de error) con un nivel específico de confianza (que es similar a una probabilidad).

La ecuación para calcular el intervalo de confianza para la media de una muestra es:

$$\bar{x} \pm z\left(\frac{\sigma}{\sqrt{n}}\right)$$

Objetivos de aprendizaje

Después de completar este proyecto, el estudiante podrá:

- Definir estimación puntual, error estándar, nivel de confianza y margen de error.
- Comparar y contrastar el error estándar y el margen de error.
- Calcular e interpretar intervalos de confianza para medias y proporciones.
- Identificar la ecuación de intervalo de confianza adecuada según el tipo de variable de resultado y la cantidad de muestras.

Material

- Computadora con el software R instalado o acceso a R Studio Cloud (https://posit.cloud/)
- Para la realización de esta práctica se requieren los siguientes paquetes:

```r
library(datasets)
library(ggplot2)
library(readxl)
library(tidyverse)
library(broom)
library(knitr)
library(samplingbook)
library(tidyverse)
library(broom)
library(kableExtra)
```

Ejemplo 3: Consumo de Combustible

Para ilustrar el cálculo de los intervalos de confianza en R, utilizaremos un conjunto de datos del mundo real con R. Para este ejemplo, utilizaremos el conjunto de datos `mtcars`, que contiene datos sobre diversos aspectos del diseño y el rendimiento de los automóviles.

```r
data("mtcars")
head(mtcars)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```r
summary(mtcars)
```

```
##       mpg             cyl             disp             hp
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0
##       drat             wt             qsec             vs
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000
##        am              gear            carb
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
##  Median :0.0000   Median :4.000   Median :2.000
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

Nos interesa estimar el consumo de combustible promedio (mpg) de los autos en este conjunto de datos. A continuación, se muestra cómo calcular un intervalo de confianza del 95 % para el consumo promedio de combustible:

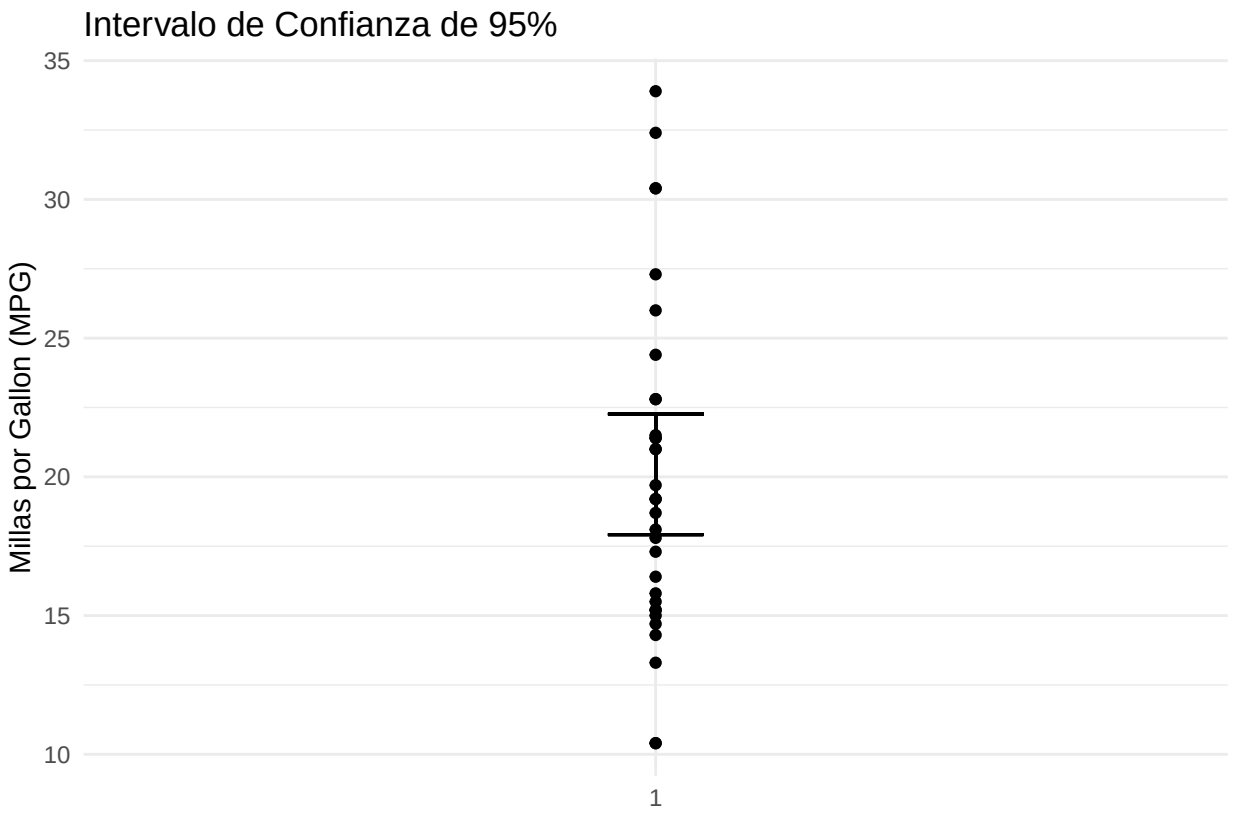
```r
mean_mpg <- mean(mtcars$mpg)
se_mpg <- sd(mtcars$mpg) / sqrt(nrow(mtcars))
ci_mpg <- mean_mpg + c(-1, 1) * qt(0.975, df = nrow(mtcars) - 1) * se_mpg
ci_mpg
```

```
## [1] 17.91768 22.26357
```

Este código calcula la media de mpg y el error estándar de la media (se_mpg) y luego los utiliza para calcular el intervalo de confianza (ci_mpg). La función qt encuentra el valor crítico para la distribución t, lo cual es apropiado aquí debido al tamaño de la muestra y al hecho de que estamos estimando una media.

La visualización ayuda a la comprensión. Creemos un gráfico simple para mostrar este intervalo de confianza:

```r
ggplot(mtcars, aes(x = factor(1), y = mpg)) +
  geom_point() +
  geom_errorbar(aes(ymin = ci_mpg[1], ymax = ci_mpg[2]), width = 0.1) +
  theme_minimal() +
  labs(title = "Intervalo de Confianza de 95% ",
       x = "",
       y = "Millas por Gallon (MPG)")
```



Este código produce un gráfico con la media de mpg y barras de error que representan el intervalo de confianza.

Ejemplo 4: Concentración en Medicamentos

Se sabe que para que un fármaco sea efectivo, la concentración de su principio activo debe ser de al menos 16 mg/mm3. Una farmacia va a comprar un lote de este medicamento, pero antes quiere asegurarse de que los medicamentos del lote son efectivos y para ello analiza la concentración de principio activo en una muestra aleatoria de 50 envases tomados del lote, obteniendo los siguientes resultados en mg/mm3:

Datos: *https://github.com/ghebrer82/EstadisticaAplicada/blob/Ejemplos/Medicamentos.xlsx*

Leemos el archivo "Medicamentos", que contiene las concentración de 50 medicamentos.

```
df.3 <- read_excel('Medicamentos.xlsx')
```

1. Calcular la concentración media de principio activo de la muestra. ¿Puede afirmarse que los medicamentos del lote son efectivos?

```
df.3$concentracion<-as.numeric(df.3$`Concentración (mg/mm³)`)

mean(df.3$concentracion)
```

```
## [1] 17.408
```

A pesar de que la concentración media está por encima de 16 mg/mm3, se trata de una estimación puntual, y por tanto, no podemos garantizar que la media poblacional esté por encima de 16 mg/mm3.

2. Calcular el intervalo de confianza para la media de la concentración del lote con nivel de confianza del 95%. ¿Puede afirmarse ahora que los medicamentos del lote son efectivos?

```
t1 <- t.test(df.3$concentracion)
t1$conf.int
```

```
## [1] 17.03396 17.78204
## attr(,"conf.level")
## [1] 0.95
```

Ponemos los resultado en formato de tabla:

```
tidy(t1) |>
    select(estimate, conf.low, conf.high) |>
    kable() |>
kable_styling()
```

estimate	conf.low	conf.high
17.408	17.03396	17.78204

Como el intervalo entero está por encima de 16 mg/mm3, podemos afirmar con una confianza del 95% que la concentración media de principio activo del lote está por encima de 16 mg/mm3 y, por lo tanto, podemos concluir que los medicamentos del lote son efectivos.

3. ¿Puede afirmarse que los medicamentos del lote son efectivos con un 99% de confianza?

```
t2 <- t.test(df.3$concentracion, conf.level = 0.99)
```

```
tidy(t2) |>
    select(estimate, conf.low, conf.high) |>
    kable() |>
kable_styling()
```

estimate	conf.low	conf.high
17.408	16.90919	17.90681

Como el intervalo entero sigue estando por encima de 16 mg/mm3, podemos afirmar con una confianza del 99% que los medicamentos del lote son efectivos.

4. Qué tamaño muestral sería necesario para obtener una estimación del contenido medio de principio

activo con un margen de error de +-0.5 mg/mm3 y una confianza del 95%?

usamos la función `sample.size.mean` de la librería `samplingbook`. Donde el argumento "e", es el margen de error (0.5) y "level" es el nivel de confianza.

```
sample.size.mean(e = 0.5, S = sd(df.3$concentracion), level = 0.95)
```

```
##
## sample.size.mean object: Sample size for mean estimate
## Without finite population correction: N=Inf, precision e=0.5 and standard deviation S=1.3161
##
## Sample size needed: 27
```

Para obtener un margen de error de +- 0.5 mg/mm3 y un nivel de confianza de 95%, necesitamos una tamaño de muestra de 27.

Ejemplo 5: Intervalo de Confianza para Proporciones

El conjunto de datos Elevador.csv contiene los resultados de una encuesta realizada en una universidad, sobre si el alumnado utiliza habitualmente el elevador del edificio.

Datos: *https://github.com/ghebrer82/EstadisticaAplicada/blob/Ejemplos/Elevador.csv*

```
df.3p <- read_csv("Elevador.csv",show_col_types = FALSE)
```

1. Calcular el intervalo de confianza con 99% para la proporción del alumnado que utiliza habitualmente el elevador.

Para calcular el intervalo de confianza para la proporción de una población podemos utilizar la función `prop.test` del paquete `stats`.

Si queremos mostrar la salida del test en formato de tabla podemos utilizar la función `tidy` del paquete `broom`.

```
frec <- table(df.3p$Respuesta)
tidy(prop.test(frec["Sí"], nrow(df.3p), conf.level = 0.99)) |>
select(estimate, conf.low, conf.high) |>
kable() |>
kable_styling()
```

estimate	conf.low	conf.high
0.62	0.4296721	0.7807223

Se trata de un intervalo poco preciso, ya que su amplitud es bastante grande.

2. Qué tamaño muestral sería necesario para obtener una estimación de la proporción de alumnos que utilizan regularmente elevador con un margen de error de un 1% y una confianza del 95%?

El tamaño muestral necesario para construir un intervalo de confianza para la media depende del nivel de confianza deseado (95% en este caso), del error o semiamplitud del intervalo deseado (0.01 en este caso) y de proporción poblacional, que no se conoce, pero se puede estimar mediante la proporción muestral.

Usamos la función `sample.size.prop` donde le margen de error es e=0.01 y el nivel de confianza es level=0.95:

```
sample.size.prop(e = 0.01, P = frec["Sí"]/nrow(df.3p), level = 0.95)
```

```
##
## sample.size.prop object: Sample size for proportion estimate
```

```
## Without finite population correction: N=Inf, precision e=0.01 and expected proportion P=0.62
##
## Sample size needed: 9051
```

Para un margen de error de 1% y un nivel de confianza del 95% se requiere un tamaño muestral de 9051 elementos.

3. Calcular los intervalo de confianza para las proporciones de chicas y chicos que utilizan regularmente elevador. ¿Existe una diferencia estadísticamente significativa entre la proporción de chicas y chicos que lo utilizan regularmente? En tal caso, ¿quiénes lo utilizan más?

```
df.3p |>
    group_by(Género) |>
    count(Respuesta) |>
    mutate(test = map(n, \(x) tidy(prop.test(x, sum(n))))) |>
    unnest(test) |>
    filter(Respuesta == "Sí") |>
    select(Género, Respuesta, n, estimate, conf.low, conf.high) |>
    kable()
```

Género	Respuesta	n	estimate	conf.low	conf.high
Femenino	Sí	22	0.8461538	0.6427297	0.9495302
Masculino	Sí	9	0.3750000	0.1955019	0.5924241

Como los intervalos de confianza para las proporciones de chicos y chicas que utilizan regularmente el elevador no se traslapan, es decir, no tienen valores en común, podemos concluir que existe una diferencia estadísticamente significativa entre ambas proporciones con un 95% de confianza. Como además el intervalo de confianza para la proporción de chicas está claramente por encima del de chicos, se concluye que en esa universidad hay más chicas utilizan regularmente el elevador.

Ejercicios

1.Una empresa fabricante de teléfonos móviles quiere estimar la duración promedio de la batería de su nuevo modelo en condiciones de uso normal. Para ello, selecciona una muestra aleatoria de 40 teléfonos y registra la duración de su batería en horas hasta que se agota completamente.

Datos: Los datos se encuentran en el archivo Excel adjunto: duración_bateria.xlsx. *(https://github.com/ghebrer82/EstadisticaAplicada/blob/main/03_duracion_bateria.xlsx)*

a) Calcular el intervalo de confianza para la media de la duración de la batería con un nivel de confianza del 95%. Asumir que la desviación estándar poblacional es desconocida y se debe estimar a partir de la muestra.
b) Determinar qué tamaño muestral sería necesario para obtener una estimación de la media de la duración de la batería con un margen de error de ± 0.5 horas y un nivel de confianza del 95%. Utilizar la desviación estándar muestral obtenida en el punto (a) como una estimación de la desviación estándar poblacional.

Bibliografía

- Estadística, Mario Triola, 12va Edición, Pearson, 2018.
- Probabilidad y Estadística para Ingeniería y Ciencias, Ronald Walpole, Pearson Educación, 2012.

4 Inferencia Estadística: Pruebas de Hipótesis

Introducción

Una prueba de hipótesis estadística nos permite sopesar la evidencia a favor o en contra de una afirmación existente. Cuando hacemos una inferencia estadística sobre una media, nuestra hipótesis nula es que la media de la población es igual a un valor particular. Este valor puede haber sido determinado a partir de otro estudio o ser una suposición convencional o fundamentada. Luego planteamos una hipótesis alternativa: que la media es mayor que, menor que o simplemente no es igual al valor propuesto en la hipótesis nula. El proceso de prueba de hipótesis implica establecer dos hipótesis en competencia, la hipótesis nula y la hipótesis alternativa. Se selecciona una muestra aleatoria (o varias muestras con más grupos de comparación), se calculan estadísticas de resumen y luego se evalúa la probabilidad de que los datos de la muestra respalden la investigación o la hipótesis alternativa. De manera similar a la estimación, el proceso de prueba de hipótesis se basa en la teoría de la probabilidad y el teorema del límite central.

Objetivos de aprendizaje

Después de completar esta practica, el estudiante podrá:

- Definir hipótesis nula e hipótesis de investigación, estadística de prueba, nivel de significancia y regla de decisión.
- Explicar la diferencia entre pruebas de hipótesis de una y dos colas.
- Estimar e interpretar valores p.
- Entender la importancia de diferenciar los procedimientos de prueba de hipótesis según el tipo de variable de resultado y el número de muestras.

Material

- Computadora con el software R instalado o acceso a R Studio Cloud (https://posit.cloud/)
- Para la realización de esta práctica se requieren los siguientes paquetes:

```
library(tidyverse)
library(broom)
library(tidymodels)
library(pwr)
library(knitr)
library(readxl)
library(kableExtra)
```

Ejemplo 6: Concentración de medicamentos (Contraste de hipótesis)

Se sabe que para que un fármaco sea efectivo, la concentración de su principio activo debe ser de al menos 16 mg/mm3. Una farmacia va a comprar un lote de este medicamento, pero antes quiere asegurarse de que los medicamentos del lote son efectivos y para ello analiza la concentración de principio activo en una muestra aleatoria de 50 envases tomados del lote, obteniendo los siguientes resultados en mg/mm3:

Datos: *https://github.com/ghebrer82/EstadisticaAplicada/blob/Ejemplos/Medicamentos.xlsx*

Leemos el archivo excel con las concentraciones de 50 medicamentos:

```
df.4 <- read_excel('Medicamentos.xlsx')
```

1. Calcular la concentración media de principio activo de la muestra. ¿Puede afirmarse que los medicamentos del lote son efectivos?

```
df.4$concentracion<-as.numeric(df.4$`Concentración (mg/mm³)`)

mean(df.4$concentracion)
```

```
## [1] 17.408
```

2. Realizar un contraste de hipótesis para ver si la concentración media de principio activo es diferente de 18 mg/mm3:

Tenemos que realizar el contraste bilateral h0: mu=18, h1: mu diferente a 18

```
t.test(df.4$concentracion, mu = 18)
```

```
##
##  One Sample t-test
##
## data:  df.4$concentracion
## t = -3.1806, df = 49, p-value = 0.00255
## alternative hypothesis: true mean is not equal to 18
## 95 percent confidence interval:
##  17.03396 17.78204
## sample estimates:
## mean of x
##    17.408
```

Ponemos los resultados en una tabla:

```
tidy(t.test(df.4$concentracion, mu = 18)) |>
    kable()
```

estimate	statistic	p.value	parameter	conf.low	conf.high	method	alternative
17.408	-3.180609	0.0025495	49	17.03396	17.78204	One Sample t-test	two.sided

Como el p-valor del contraste es 0.00255 que es menor que el riesgo 0.05 ($p<0.05$), rechazamos la hipótesis nula y concluimos que la concentración media es significativamente diferente de 18 mg/mm3

3. Si el fabricante del lote asegura haber aumentado la concentración de principio activo con respecto a anteriores lotes, en los que la media era de 17 mg/mm3 , ¿podemos aceptar la afirmación del fabricante?

Ahora tenemos que realizar el contraste unilateral

h0: mu =17, H1: mu > 17

```
t.test(df.4$concentracion, mu = 17, alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  df.4$concentracion
## t = 2.192, df = 49, p-value = 0.01658
## alternative hypothesis: true mean is greater than 17
## 95 percent confidence interval:
##  17.09595      Inf
## sample estimates:
## mean of x
##    17.408
```

```
tidy(t.test(df.4$concentracion, mu = 17, alternative = "greater")) |>
    kable()
```

estimate	statistic	p.value	parameter	conf.low	conf.high	method	alternative
17.408	2.192041	0.0165787	49	17.09595	Inf	One Sample t-test	greater

Como el p-valor del contraste es 0.01 que es menor que el riesgo 0.05 (p<0.05), podemos rechazar la hipótesis nula y concluimos que con esta muestra hay pruebas significativas de que la afirmación del fabricante sea cierta.

#Ejemplo 7: Dietas para bajar de Peso

Para ver si una dieta ha influido en el peso, se tomó una muestra de 78 personas, antes y después de someterse a 3 diferentes dietas

Datos: *(https://www.kaggle.com/datasets/tombenny/foodhabbits)*

1. Realizar un contraste de hipótesis para ver si hay la media del peso ha disminuido significativamente.

tenemos que realizar el contraste de hipótesis unilateral

h0: mu_inicial = mu_final , H1: mu_inicial diferente a mu_final

leer el archivo:

```
df.5 <- read_csv("foodDiet.csv")
```

Renombramos las variables “pre.weight” y “weight6weeks” a “antes” y “despues”

```
#rename column by name
df.5 <- df.5 %>% rename_at('pre.weight', ~'antes')
df.5 <- df.5 %>% rename_at('weight6weeks', ~'despues')
```

```
t.test(df.5$antes, df.5$despues, paired = TRUE, alternative = "greater")
```

```
##
##  Paired t-test
##
## data:  df.5$antes and df.5$despues
## t = 13.309, df = 77, p-value < 2.2e-16
## alternative hypothesis: true mean difference is greater than 0
## 95 percent confidence interval:
##  3.36389     Inf
## sample estimates:
## mean difference
##        3.844872
```

```
tidy(t.test(df.5$antes, df.5$despues, paired = TRUE, alternative = "greater")) |>
    kable()
```

estimate	statistic	p.value	parameter	conf.low	conf.high	method	alternative
3.844872	13.30875	0	77	3.36389	Inf	Paired t-test	greater

Como el p-valor del contraste es prácticamente 0, que es mucho menor que el riesgo 0.05 (p<0.05), podemos rechazar la hipótesis nula y se concluye que existe una diferencia estadísticamente significativa entre las medias del peso antes y después de la dieta.

2. Realizar el mismo contraste de antes, pero para cada dieta por separado.

```
df.5 |>
    nest(data = -Diet) |>
```

```
    mutate(test = map(data, ~ tidy(t.test(.x$antes, .x$despues, paired = TRUE, alternative = "greater"))
    unnest(test) |>
    select(-data) |>
    kable()
```

Diet	estimate	statistic	p.value	parameter	conf.low	conf.high	method	alternative
2	3.025926	6.231030	7e-07	26	2.197640	Inf	Paired t-test	greater
1	3.300000	7.216771	1e-07	23	2.516301	Inf	Paired t-test	greater
3	5.148148	11.166688	0e+00	26	4.361812	Inf	Paired t-test	greater

Nota: Nesting (anidación) es implícitamente una operación de resumen: se obtiene una fila para cada grupo definido por las columnas no anidadas

Como se puede observar, todos los p-valores son menores que el nivel de significación 0.05 , por lo que se puede concluir que existe una diferencia estadísticamente significativa entre las medias del peso antes y después de cada dieta. Si observamos los intervalos de confianza, se observa que la mayor diferencia entre se da para la dieta 3.

Ejemplo 8: Alumnos Aprobados (Prueba de Hipótesis de Proporciones)

Un profesor imparte clase a dos grupos de instituciones diferentes. En la institución "A" de un total de 40 alumnos, han aprobado 27; y la institución "B", sobre un total de 50 alumnos, han aprobado 15. ¿Se puede afirmar que hay diferencias significativas entre los porcentajes de aprobados en ambos grupos?

Tenemos que realizar el contraste bilateral (proporciones)

H0: pA=pB, H1: pA diferente a pB

```
# Aplicamos el test de comparación de proporciones.
tidy(prop.test(c(55, 32), c(80, 90))) |>
    # Multiplicamos por 100 todas las columnas para obtener porcentajes.
    mutate(across(c(estimate1, estimate2, conf.low, conf.high), ~ .x * 100)) |>
    # Mostramos por pantalla en formato tabla.
    kable()
```

estimate	estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high	method	alternative
68.75	35.55556	17.37244	3.07e-05	1	17.83764	48.55125	2-sample test for equality of proportions with continuity correction		two.sided

Como el p-valor del contraste es prácticamente 0, que es mucho menor que el riesgo 0.05 (p<0.05), podemos rechazar la hipótesis nula y se concluye que existe una diferencia estadísticamente significativa entre los porcentajes de aprobados por Institución.

Ejercicios

1.Una empresa agrícola ha desarrollado un nuevo fertilizante y afirma que el peso promedio de las mazorcas de maíz cultivadas con este fertilizante es superior a los 300 gramos, que es el promedio histórico con el fertilizante estándar. Para probar esta afirmación, se seleccionó una muestra aleatoria de 50 mazorcas de maíz cultivadas con el nuevo fertilizante y se registró su peso.

Datos: Los datos se encuentran en el archivo 04_pesos.xlsx *(https://github.com/ghebrer82/EstadisticaAplicada/blob/main/04_pesos.xlsx)*

Realice lo siguiente (con un nivel de significancia =0.05):

a) Formular las hipótesis nula y alternativa.
b) Realizar la prueba de hipótesis para determinar si el peso promedio de las mazorcas de maíz cultivadas con el nuevo fertilizante es significativamente mayor que 300 gramos.
c) Calcular el valor del estadístico de prueba.
d) Determinar el p-valor de la prueba.
e) Tomar una decisión (rechazar o no rechazar la hipótesis nula) y escribir una conclusión en el contexto del problema.

2.Una empresa de videojuegos quiere determinar si un nuevo programa de entrenamiento "A" mejora los tiempos de reacción de los jugadores más que un programa de entrenamiento "B" existente. Para ello, seleccionaron aleatoriamente a 35 jugadores para el programa "A" y a 30 jugadores para el programa "B". Al finalizar un período de entrenamiento, se registró el tiempo de reacción de cada jugador en un test estandarizado (medido en milisegundos).

Datos: Los datos se encuentran en el archivo 04b_tiempos.xlsx *(https://github.com/ghebrer82/Estadistica Aplicada/blob/main/04b_tiempos.xlsx)*

Realice lo siguiente (con un nivel de significancia =0.05): a) Formular las hipótesis nula y alternativa. b) Realizar la prueba de hipótesis para determinar si el tiempo promedio de reacción del Grupo A es significativamente menor que el del Grupo B. c) Calcular el valor del estadístico de prueba. d) Determinar el p-valor de la prueba. e) Tomar una decisión (rechazar o no rechazar la hipótesis nula) y escribir una conclusión en el contexto del problema.

Bilbiografía

- Estadística, Mario Triola, 12va Edición, Pearson, 2018.
- Probabilidad y Estadística para Ingeniería y Ciencias, Ronald Walpole, Pearson Educación, 2012.

5 Estimación: Intervalos de Confianza (dos muestras)

Introducción

Ahora que hemos aprendido a probar hipótesis y calcular intervalos de confianza para un único parámetro de población, queremos ampliar nuestros métodos de inferencia para poder comparar dos grupos. En esta práctica, nos centraremos en comparar medias utilizando dos diseños de estudio diferentes: dos muestras independientes y pares emparejados.

Objetivos de aprendizaje

- Diferenciar muestras independientes y pareadas o emparejadas.
- Calcular intervalos de confianza para la diferencia de medias y proporciones en muestras independientes y para la diferencia de medias en muestras pareadas.
- Identificar la fórmula de intervalo de confianza adecuada en función del tipo de variable de resultado y la cantidad de muestras.

Material

- Computadora con el software R instalado o acceso a R Studio Cloud (https://posit.cloud/)
- Para la realización de esta práctica se requieren los siguientes paquetes:

```
library(tidyverse)
library(broom)
library(knitr)
library(readxl)
library(kableExtra)
```

Ejemplo 9: Intervalos de confianza para la comparación de medias y proporciones de dos poblaciones

Para ver si una dieta ha influido en el peso, se tomó una muestra de 78 personas, antes y después de someterse a diferentes dietas. Estos datos se incluyen en el archivo "foodDiet01.csv". *https://github.com/ghebrer82/EstadisticaAplicada/blob/Ejemplos/foodDiet01.csv*

Primero, preparamos los datos:

leer el archivo:

```
df.5 <- read_csv("foodDiet01.csv")
```

Renombramos las variables "pre.weight" y "weight6weeks" a "antes" y "despues"

```
#rename column by name
df.5 <- df.5 %>% rename_at('pre.weight', ~'antes')
df.5 <- df.5 %>% rename_at('weight6weeks', ~'despues')
```

1. Calcular los pesos medios antes y después de la dieta. ¿Ha disminuido el peso? ¿Crees que los resultados son estadísticamente significativos?

```
df.5 |>
    # Calculamos la media de las columnas antes y despues del data frame.
    summarize(across(c(antes,despues), ~ mean(.x, na.rm = TRUE))) |>
    # Mostramos por pantalla en formato tabla.
    kbl() |>
kable_styling()
```

antes	despues
72.52564	68.68077

A pesar de que el peso medio después de la dieta ha disminuido en la muestra, no podemos concluir que las medias poblacionales han disminuido significativamente ya que se trata de estimaciones puntuales que no tienen en cuenta el error en la estimación.

2. Calcular el intervalo de confianza para la media de la diferencias entre los pesos antes y despues de someterse a la dieta. ¿Existen pruebas suficientes para afirmar con un 95% de confianza que la dieta ha disminuido el peso de la personas?

```
# Añadimos al data frame una nueva variable con la diferencia entre las ventas de después y antes
df.5$dif <- df.5$despues - df.5$antes
# Aplicamos el test de la t de student para una muestra.
tidy(t.test(df.5$dif)) |>
    # Obtenemos la estimación de la media de la diferencia de pesos y el intervalo de confianza del 95%
    select(estimate, conf.low, conf.high) |>
    # Mostramos por pantalla en formato tabla.
    kable() |>
  kable_styling()
```

estimate	conf.low	conf.high
-3.844872	-4.420141	-3.269602

Podemos llegar a este mismo intervalo de confianza sin necesitad de calcular previamente la diferencia entre las ventas de después y antes, pasándole directamente las dos variables a comparar a la función t.test añadiendo el parámetro paired = TRUE.

```
# Aplicamos el test de la t de student para muestras pareadas.
tidy(t.test(df.5$despues, df.5$antes, paired = TRUE)) |>
    # Obtenemos la estimación de la media de la diferencia entre las ventas de después y antes y el int
    select(estimate, conf.low, conf.high) |>
    kable() |>
  kable_styling()
```

estimate	conf.low	conf.high
-3.844872	-4.420141	-3.269602

podemos afirmar con un 95% de confianza que la media de la diferencia entre los pesos antes y despues de someterse a la dieta es negativa, es decir ha habido una perdida estadísticamente significativa en el peso después de someterse a la dieta y la perdida de peso media estará entre 3.3 y 4.4 kg.

Ejemplo 10: IC de Proporciones

Un profesor imparte clase a dos grupos de instituciones diferentes. En la institución "A" de un total de 40 alumnos, han aprobado 27; y la institución "B", sobre un total de 50 alumnos, han aprobado 15.

1. Existen diferencias significativas en el porcentaje de aprobados de los dos grupos? En tal caso, ¿en qué Institución hay un porcentaje mayor de aprobados y cuánto mayor es?

```
# Aplicamos el test de comparación de proporciones.
tidy(prop.test(c(27, 15), c(40, 50))) |>
    select(estimate1, estimate2, conf.low, conf.high) |>
    # Multiplicamos por 100 todas las columnas para obtener porcentajes.
    mutate(across(everything(), ~ .x * 100)) |>
    # Mostramos por pantalla en formato tabla.
    kable() |> kable_styling()
```

estimate1	estimate2	conf.low	conf.high
67.5	30	15.96215	59.03785

podemos afirmar con un 95% de confianza que existen diferencias estadísticamente significativas entre el porcentaje de aprobados en la institución "A" y "B". Como además el intervalo de confianza es para la diferencia entre el porcentaje de aprobados en A y el porcentaje de aprobados en B, se puede concluir que la proporción de aprobados en A es significativamente mayor que en B, en particular, entre un 16% y un 59% mayor.

Ejercicios:

1. Una clínica de investigación está probando un nuevo tratamiento para reducir los niveles de glucosa en sangre en pacientes con prediabetes. Se seleccionó una muestra aleatoria de 30 pacientes y se registraron sus niveles de glucosa en sangre (en mg/dL) antes de iniciar el tratamiento y después de 3 meses de recibirlo.

Datos: Los datos se encuentran en el archivo 05_glucosa.xlsx *(https://github.com/ghebrer82/EstadisticaAplicada/blob/main/05_glucosa.xlsx)*

Realice lo siguiente: a) Calcular los niveles de glucosa medios antes y después del tratamiento. ¿Han disminuido los niveles de glucosa? (Esto es una observación inicial, no una conclusión estadística formal aquí). b) Calcular el intervalo de confianza para la media de las diferencias entre los niveles de glucosa (Antes - Después) con un nivel de confianza del 95%. ¿Existen pruebas suficientes para afirmar con un 95% de confianza que el tratamiento ha disminuido los niveles de glucosa en las personas? (Utiliza el intervalo de confianza para responder a esta pregunta).

Bibliografía

- Estadística, Mario Triola, 12va Edición, Pearson, 2018.
- Probabilidad y Estadística para Ingeniería y Ciencias, Ronald Walpole, Pearson Educación, 2012.

6 Análisis de Varianza (ANOVA)

Introducción

El análisis de varianza (ANOVA) es una prueba de hipótesis adecuada para comparar las medias de una variable continua en dos o más grupos de comparación independientes. Por ejemplo, en algunos ensayos clínicos hay más de dos grupos de comparación. En un ensayo clínico para evaluar un nuevo medicamento, los investigadores podrían comparar un medicamento experimental con un placebo y un tratamiento estándar. También podría ser interesante comparar la presión arterial media o los niveles medios de colesterol en personas con bajo peso, peso normal, sobrepeso y obesidad.

La técnica ANOVA se aplica cuando hay dos o más de dos grupos independientes. El procedimiento ANOVA se utiliza para comparar las medias de los grupos de comparación y se lleva a cabo utilizando un enfoque similar al de la pruebas de hipótesis aplicadas anteriormente. Sin embargo, debido a que hay más de dos grupos, el cálculo de la estadística de prueba es más complejo. La estadística de prueba debe tener en cuenta los tamaños de muestra, las medias de muestra y las desviaciones estándar de muestra en cada uno de los grupos de comparación.

La estrategia fundamental del ANOVA es examinar sistemáticamente la variabilidad dentro y entre los grupos que se comparan.

Existen pruebas ANOVA de uno o dos factores. En ANOVA de un factor, se tiene un tratamiento o factor de agrupación con k>2 niveles, y deseamos comparar las medias entre las diferentes categorías de este factor. El factor puede representar diferentes dietas, diferentes clasificaciones de enfermedad, diferentes tratamientos médicos, o diferentes grupos de edad. Por otro lado, hay situaciones en las que puede ser de interés comparar las medias entre dos o más factores. Por ejemplo, supongamos que un ensayo clínico está diseñado para comparar cinco tratamientos diferentes para pacientes con una enfermedad crónica. Los investigadores también pueden plantear la hipótesis de que existen diferencias en el resultado según el género. Este es un ejemplo de un ANOVA de dos factores donde los factores son el tratamiento (con cinco niveles) y el género (con dos niveles). En el ANOVA de dos factores, los investigadores pueden evaluar si existen diferencias en las medias debido al tratamiento según el género o si existe una diferencia en los resultados según la combinación o interacción del tratamiento y el genero. Los ANOVA de orden superior se realizan de la misma manera que los ANOVA de un factor

Contrastes Post hoc

Si el valor p del ANOVA es menor que el nivel de significancia, podemos rechazar la hipótesis nula y concluir que tenemos evidencia suficiente para decir que al menos una de las medias de los grupos es diferente de las demás.

Sin embargo, esto no nos dice qué grupos son diferentes entre sí. Simplemente nos dice que no todas las medias de los grupos son iguales.

Para averiguar exactamente qué grupos son diferentes entre sí, debemos realizar una prueba post hoc (también conocida como prueba de comparación múltiple). Un ejemplo de este tipo de pruebas es la prueba de Tukey.

Objetivos de Aprendizaje

Después de completar esta práctica, el estudiante podrá:

- Realizar análisis de varianza utilizando R.
- Interpretar adecuadamente los resultados de las pruebas de análisis de varianza.
- Distinguir entre pruebas de análisis de varianza de uno y dos factores.
- Identificar el procedimiento de prueba de hipótesis adecuado según el tipo de variable de resultado y la cantidad de muestras.
- Realizar prueba post hoc (o de comparación múltiple)

Material

- Computadora con el software R instalado o acceso a R Studio Cloud (https://posit.cloud/)
- Para la realización de esta práctica se requieren los siguientes paquetes:

```
library(tidyverse)
library(broom)
library(rstatix)
library(lme4)
library(knitr)
library(readxl)
library(reshape2)
library(kableExtra)
library(car)
library(ggpubr)
library(ggplot2)
library(ggsignif)
library(multcomp)
```

Ejemplo 11: Pérdida de peso por dieta

Se lleva a cabo un ensayo clínico para comparar programas de pérdida de peso. Los participantes son asignados aleatoriamente a uno de los programas de comparación y se les asesora sobre los detalles del programa asignado. Los participantes siguen el programa asignado durante 8 semanas. El resultado de interés es la pérdida de peso, definida como la diferencia entre el peso medido al inicio del estudio (línea de base) y el peso medido al final (8 semanas), medido en libras.

Se consideran tres dietas populares de pérdida de peso. El primero es una dieta baja en calorías (Low Calorie). El segundo es una dieta baja en grasas (Low Fat) y el tercero es una dieta baja en carbohidratos (Low Carbohydrate). A los efectos de comparación, un cuarto grupo se considera un grupo de control (Control). A los participantes del cuarto grupo se les dice que están participando en un estudio de conductas saludables con pérdida de peso, solo un componente de interés. El grupo de control se incluye aquí para evaluar el efecto placebo (es decir, pérdida de peso debido simplemente a la participación en el estudio). Veinte pacientes aceptan participar en el estudio y son asignados aleatoriamente a uno de los cuatro grupos de dieta. Se mide el peso al inicio del estudio y se aconseja a los pacientes sobre cómo implementar correctamente la dieta asignada (excepto el grupo de control). Después de 8 semanas, se vuelve a medir el peso de cada paciente y se calcula la diferencia de peso restando el peso de las 8 semanas al peso inicial. Las diferencias positivas indican pérdida de peso y las diferencias negativas indican aumento de peso. A los efectos de interpretación, nos referimos a las diferencias de peso como pérdidas de peso y las pérdidas de peso observadas se muestran a continuación.

Datos: *https://github.com/ghebrer82/EstadisticaAplicada/blob/Ejemplos/3diets.xlsx*

Primero, leemos el archivo excel que contiene la información de la pérdida de peso de los participantes.

```
df.6 <- read_excel('3diets.xlsx')
kable(df.6) |> kable_styling()
```

Control	Low Calorie	Low Fat	Low Carbohydrate
2	8	2	3
2	9	4	5
-1	6	3	4
0	7	5	2
3	3	1	3

1. Dibujar el diagrama de cajas con los puntos correspondientes a las perdidas de peso . A la vista del

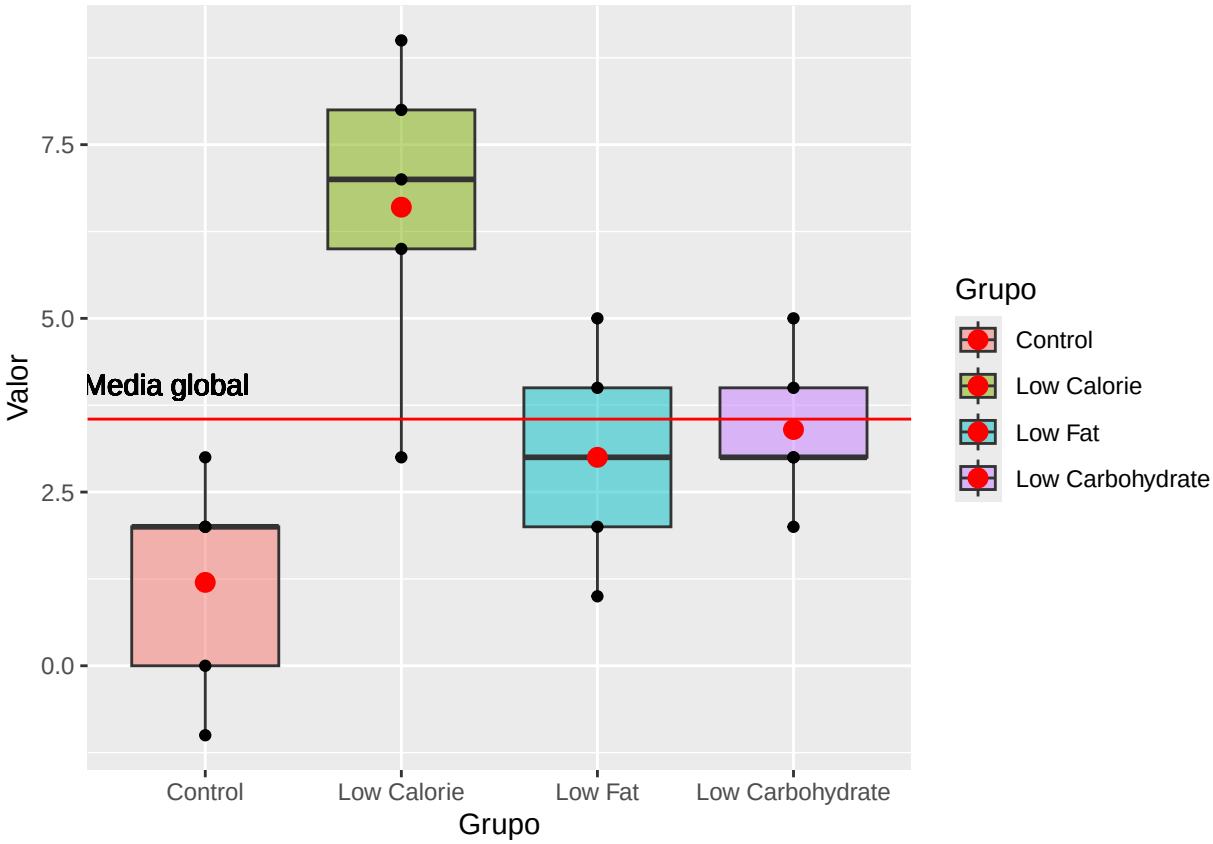
diagrama, ¿crees que existen diferencias significativas entre las perdidas de peso de las tres dietas?

Transformamos a dataframe para realizar ANOVA,usando la función `melt` del paquete `reshape2`:

```
# Transformar la tabla
df.6a <- melt(df.6, variable.name = "Grupo", value.name = "Valor")
```

Generamos el gráfico de cajas para los cuatro grupos (3 diferentes dietas y Control):

```
media<- mean(df.6a$Valor)
df.6a |>
    ggplot(aes(x = Grupo, y = Valor, fill = Grupo)) +
    geom_boxplot(alpha = 0.5) +
    geom_point() +
    stat_summary(fun = mean, geom = "point", size = 3, color = "red") +
    geom_hline(yintercept = media, color = "red") +
    geom_text(aes(x = 0.8, y = media + 0.5, label = "Media global") )
```



Nótese que al diagrama de cajas se le ha agregado la media de cada grupo (círculo rojo) y la media global (línea roja).

Podemos observar que existen diferencias de perdida de peso de las tres dietas, específicamente entre "Low Calorie" y el resto.

2. Realizar un contraste ANOVA para conocer si existen diferencias estadísticamente significativas entre las pérdidas de peso de las tres dietas.

Antes de realizar el contraste de ANOVA hay que comprobar que se cumplen los supuestos del modelo ANOVA.

- Normalidad: Para comprobar la normalidad de la variable dependiente se puede utilizar el test de normalidad de Shapiro-Wilk mediante la función shapiro.test.

- Homogeneidad de la varianza (homocedasticidad): Para comprobar la homogeneidad de las varianzas de los grupos de comparación se puede utilizar el test de Barlett de homogeneidad de varianzas mediante la función bartlett.test.
- Independencia: Para comprobar la independencia de las observaciones se puede utilizar el test de independencia de Durbin-Watson mediante la función durbinWatsonTest del paquete car.

Para realizar un contraste ANOVA podemos usar la función `aov` del paquete `stats`.

Si queremos mostrar la salida en formato de tabla podemos utilizar la función `tidy` del paquete `broom`.

Comprobamos Normalidad usando `shapiro.test`:

```
shapiro.test(df.6a$Valor) |>
    tidy() |>
    kable() |> kable_styling()
```

statistic	p.value	method
0.958491	0.5141629	Shapiro-Wilk normality test

Como el p-valor es mayor que el nivel de significación 0.05, no podemos rechazar la hipótesis nula de normalidad de los datos. (se cumple Normalidad)

Comprobamos Homogeneidad de la varianza usando `barlett.test`:

```
bartlett.test(Valor ~ Grupo, data = df.6a) |>
    tidy() |>
    kable()|> kable_styling()
```

statistic	p.value	parameter	method
1.766426	0.6222682	3	Bartlett test of homogeneity of variances

Como el p-valor es mayor que el nivel de significación 0.05, no podemos rechazar la hipótesis nula de homocedasticidad de las varianzas. (se cumple Homogeneidad)

Comprobamos Independencia, usando `durbinWatsonTest`:

```
durbinWatsonTest(aov(Valor ~ Grupo, data = df.6a)) |>
    tidy() |>
    kable()|> kable_styling()
```

statistic	p.value	autocorrelation	method	alternative
1.916949	0.362	0.0330508	Durbin-Watson Test	two.sided

Como el p-valor es mayor que el nivel de significación 0.05, no podemos rechazar la hipótesis nula de independencia de las observaciones. (se cumple Independencia)

Así pues, se cumplen todas las condiciones del modelo ANOVA, por lo que podemos realizar el contraste de comparación de medias que es:

H0: muA=muB=muC, H1: existen diferencias entre al menos 2 medias

```
library(knitr)
library(broom)
aov1 <- aov(Valor~ Grupo, data = df.6a)
```

```
aov1 |>
    tidy() |>
    kable()|> kable_styling()
```

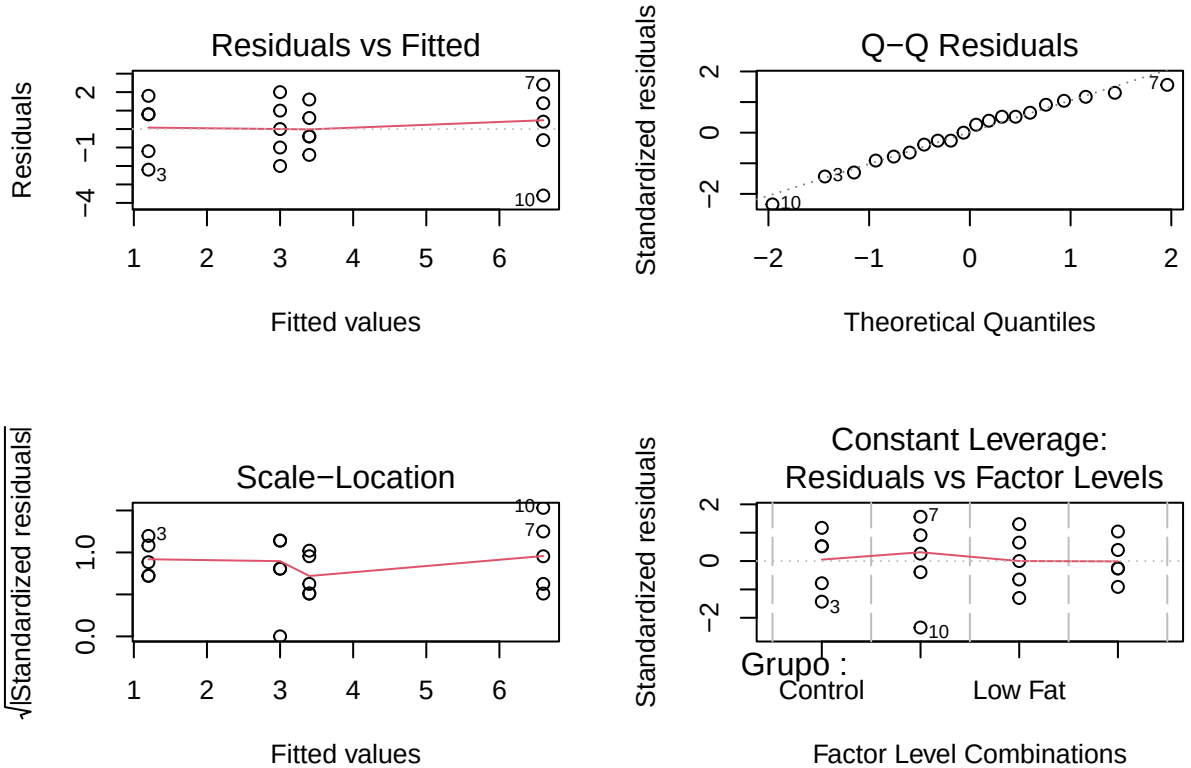
term	df	sumsq	meansq	statistic	p.value
Grupo	3	75.75	25.25	8.559322	0.0012777
Residuals	16	47.20	2.95	NA	NA

Como el p-valor del contraste es 0.001 que es mucho menor que el nivel de significancia de 0.05, rechazamos la hipótesis nula y se concluye que existen diferencias estadísticamente significativas entre las perdidas de peso de al menos dos dietas.

3. Analizar los residuos del modelo ANOVA.

El análisis de los residuos se suele realizar mediante la función plot, pasándole como argumento el modelo ANOVA. Esta función dibuja cuatro diagramas, el de los residuos frente a las predicciones del modelo, el de los cuantiles de los residuos frente a los cuantiles normales (qqplot).

```
par(mfrow=c(2,2))
plot(aov1)
```



El primer diagrama “Residuals vs Fitted” muestra si los residuos tienen tendencia lineal o no. En este caso, la línea roja que representa las medias es prácticamente horizontal por lo que se puede asumir que la tendencia es lineal.

El segundo diagrama “Normal Q-Q” muestra si los residuos siguen una distribución normal. En este caso los puntos se ajustan bastante bien a la línea recta, por lo que se puede asumir que los residuos siguen una distribución normal.

El tercer diagrama “Scale-Location” muestra si los residuos tienen una varianza constante (homocedasticidad). En este caso, la línea roja que representa la media de los residuos es prácticamente horizontal, por lo

que se puede asumir que la varianza de los residuos es constante.

El cuarto diagrama "Residuals vs Leverage" muestra si hay observaciones influyentes en el modelo. En este caso, no hay observaciones que se salgan de la línea discontinua roja que representa el límite de influencia (distancia de Cook), por lo que no hay datos atípicos que sesguen el modelo.

4. Realizar un contraste post-hoc de comparación de las medias de pérdida de peso por pares. ¿Entre qué lugares existe una diferencia estadísticamente significativa en la perdida de peso media ?

Para realizar un contraste post-hoc de comparación de medias por pares podemos usar la función `TukeyHSD` del paquete `stats`.

Otra opción es utilizar la función `pairwise.t.test` del paquete `stats` que aplica la corrección de Bonferroni a los p-valores.

Tukey:

```
TukeyHSD(aov(Valor ~ Grupo, data = df.6a)) |>
    tidy() |>
    kable()|> kable_styling()
```

term	contrast	null.value	estimate	conf.low	conf.high	adj.p.value
Grupo	Low Calorie-Control	0	5.4	2.292137	8.507863	0.0007219
Grupo	Low Fat-Control	0	1.8	-1.307863	4.907863	0.3769279
Grupo	Low Carbohydrate-Control	0	2.2	-0.907863	5.307863	0.2199271
Grupo	Low Fat-Low Calorie	0	-3.6	-6.707863	-0.492137	0.0205481
Grupo	Low Carbohydrate-Low Calorie	0	-3.2	-6.307863	-0.092137	0.0424600
Grupo	Low Carbohydrate-Low Fat	0	0.4	-2.707863	3.507863	0.9823293

Existe una diferencia muy significativa entre la pérdida de peso de las dietas Low Fat-Low Calorie, Low Carbohydrate-Low Calorie, Control-Low Calorie, pero no entre Low Carbohydrate-Low Fat, Control-Low Fat, Control-Low Carbohydrate

Bonferroni:

```
pairwise.t.test(df.6a$Valor, df.6a$Grupo, p.adjust.method = "bonferroni") |>
    tidy() |>
    kable()|> kable_styling()
```

group1	group2	p.value
Low Calorie	Control	0.0008324
Low Fat	Control	0.7019268
Low Fat	Low Calorie	0.0263239
Low Carbohydrate	Control	0.3591428
Low Carbohydrate	Low Calorie	0.0569541
Low Carbohydrate	Low Fat	1.0000000

Otro método que podemos utilizar para realizar comparaciones múltiples es la corrección de Dunett. Utilizaremos este enfoque cuando queramos comparar la media de cada grupo con una media de control y no nos interesa comparar las medias del tratamiento entre sí.

Por ejemplo, utilizando el código que aparece a continuación, comparamos las medias de las tres dietas con las del grupo control. En este caso, no nos interesan las diferencias entre las dietas.

Usamos la función `glht` del paquete `multcomp`:

```r
dunnet_comparison <- glht(aov1, linfct = mcp(Grupo = "Dunnett"))

summary(dunnet_comparison)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Dunnett Contrasts
##
##
## Fit: aov(formula = Valor ~ Grupo, data = df.6a)
##
## Linear Hypotheses:
##                                Estimate Std. Error t value Pr(>|t|)
## Low Calorie - Control == 0        5.400      1.086   4.971   <0.001 ***
## Low Fat - Control == 0            1.800      1.086   1.657    0.265
## Low Carbohydrate - Control == 0   2.200      1.086   2.025    0.143
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

A partir de los valores p del resultado, podemos ver lo siguiente:

- La diferencia entre la media de Low Calorie y Control es estadísticamente significativa a un nivel de significancia de 0.05. ($p < 0.05$)
- La diferencia entre la media de Low Fat y Control no es estadísticamente significativa a un nivel de significancia de 0.05. ($p > 0.05$)
- La diferencia entre la media de Low Carbohydrate y Control no es estadísticamente significativa a un nivel de significancia de 0.05. ($p > 0.05$)

Como dijimos antes, este enfoque compara la media de todos los demás grupos con la del grupo Control. Observe que no se realizan pruebas para las diferencias entre las dietas, porque no nos interesan las diferencias de esos grupos.

Nota: para usar la función la corrección de Dunnet con la función `glht`, es necesario que en los datos a los que se aplicó ANOVA, la primera columna contenga los datos del grupo Control.

Ejemplo 12: ANOVA dos factores

Consideremos el ensayo clínico descrito anteriormente en el que se comparan tres tratamientos en competencia para el dolor articular en términos de su tiempo medio de alivio del dolor en pacientes con osteoartritis. Debido a que los investigadores plantean la hipótesis de que puede haber una diferencia en el tiempo de alivio del dolor en hombres y mujeres, asignan aleatoriamente a 15 hombres participantes a uno de los tres tratamientos en competencia y asignan aleatoriamente a 15 mujeres participantes a uno de los tres tratamientos en competencia (es decir, asignación aleatoria estratificada). Los hombres y las mujeres que participan no saben qué tratamiento se les asigna. Se les indica que tomen la medicación asignada cuando experimenten dolor articular y que registren el tiempo, en minutos, hasta que el dolor ceda. Los datos (tiempos de alivio del dolor) se muestran a continuación y están organizados por el tratamiento asignado y el género del participante.

Datos: *https://github.com/ghebrer82/EstadisticaAplicada/blob/Ejemplos/anova2.xlsx*

Leemos el archivo que contiene los datos de los participantes:

```r
df.6b <- read_excel("anova2.xlsx")
```

1. Realizar un contraste ANOVA para conocer si el Tiempo depende del Género.

Tenemos que realizar el contraste:

H0: muA=muB=muC H1: existen diferencias en al menos dos medias

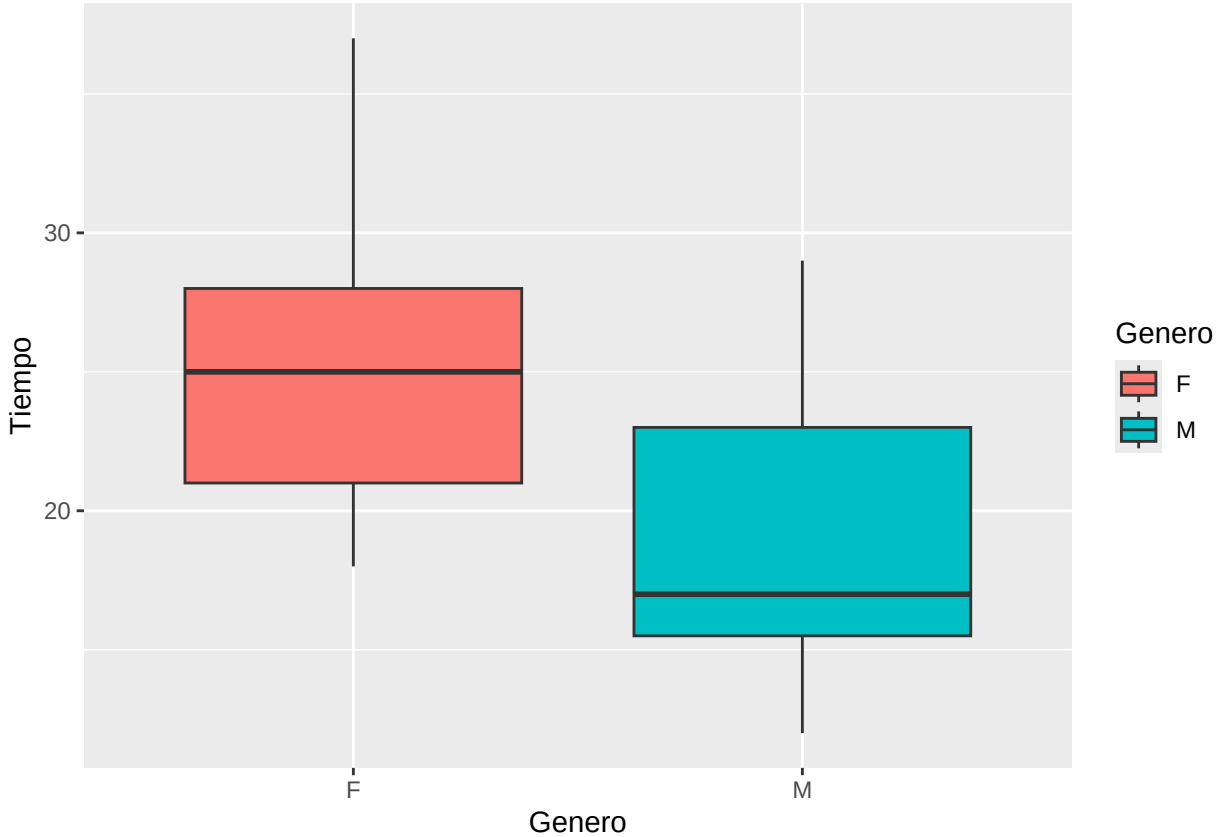
```
library(knitr)
library(broom)
aov(Tiempo ~ Genero, data = df.6b) |>
    tidy() |>
    kable()|> kable_styling()
```

term	df	sumsq	meansq	statistic	p.value
Genero	1	313.6333	313.63333	10.00501	0.003738
Residuals	28	877.7333	31.34762	NA	NA

Dado que el valor p<0.05, rechazamos la hipótesis nula, por lo que existen diferencias significativas en los tiempos de ambos géneros.

Podemos hacer le diagrama de cajas de los tiempos por género:

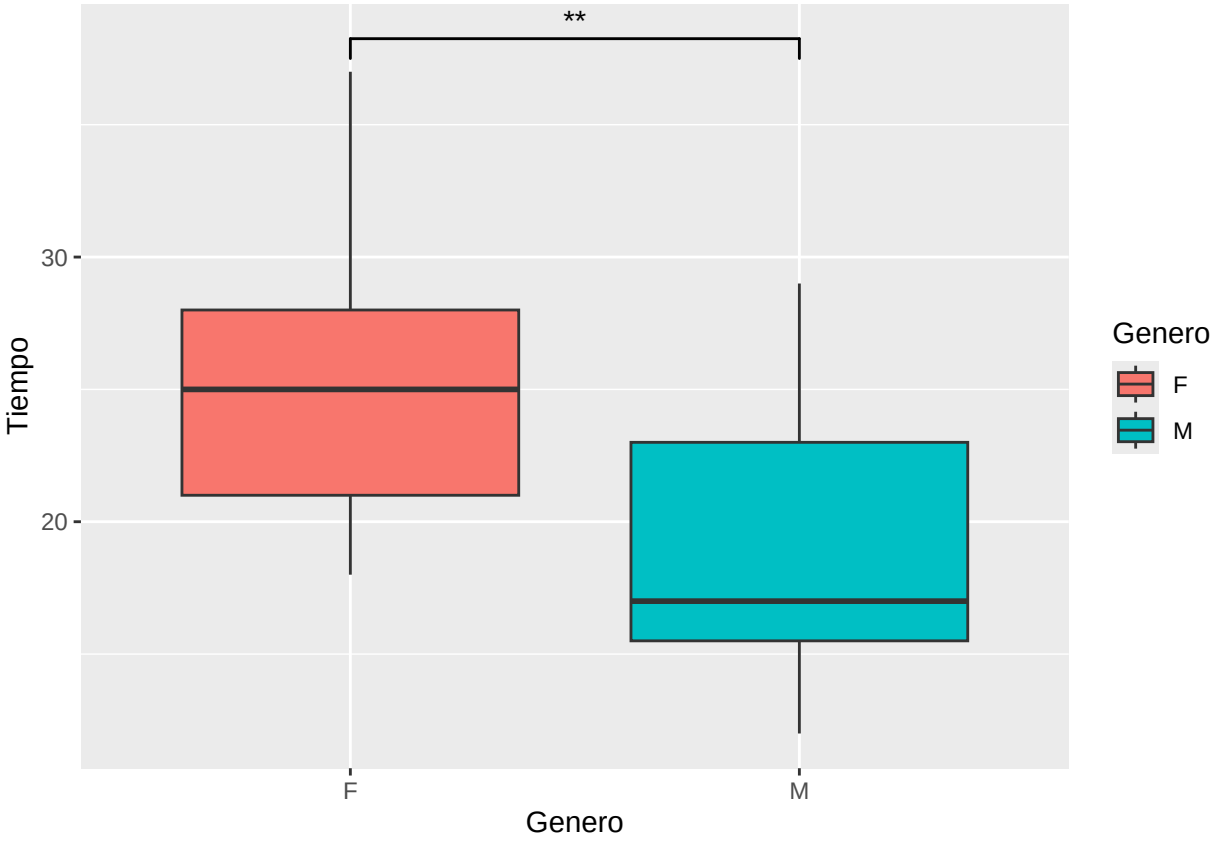
```
my.boxplot.G<-ggplot(df.6b) +
  aes(x = Genero, y = Tiempo, fill = Genero) +
  geom_boxplot()
  my.boxplot.G
```



Ahora, agreguemos el nivel de significancia al diagrama, usando la función `geom_signif` del paquete `ggsignif`:

```
my.boxplot.G+
geom_signif(comparisons = list(c("M", "F")),
```

```
            map_signif_level = c("***"=0.001, "**"=0.01, "*"=0.05)
             )
```



El nivel de significancia, se define de acuerdo a la siguiente tabla:

Symbol	Meaning
ns	P $>$ 0.05
*	P 0.05
**	P 0.01
***	P 0.001
****	P 0.0001

2. Realizar un contraste ANOVA para saber si el tiempo depende del tratamiento.

```
aov(Tiempo ~ Tratamiento, data = df.6b) |>
    tidy() |>
    kable()|> kable_styling()
```

term	df	sumsq	meansq	statistic	p.value
Tratamiento	2	651.4667	325.7333	16.28968	2.29e-05
Residuals	27	539.9000	19.9963	NA	NA

Dado que el valor p<0.05, rechazamos la hipótesis nula, por lo que existen diferencias significativas en los tiempos de al menos 2 tratamientos.

Para saber entre cuales tratamiento existen diferencias significativas, hacemos la prueba post hoc de Tukey,

usando la función `TukeyHSD` del paquete `stats`:

```
TukeyHSD(aov(Tiempo ~ Tratamiento, data = df.6b)) |>
    tidy() |>
    kable()|> kable_styling()
```

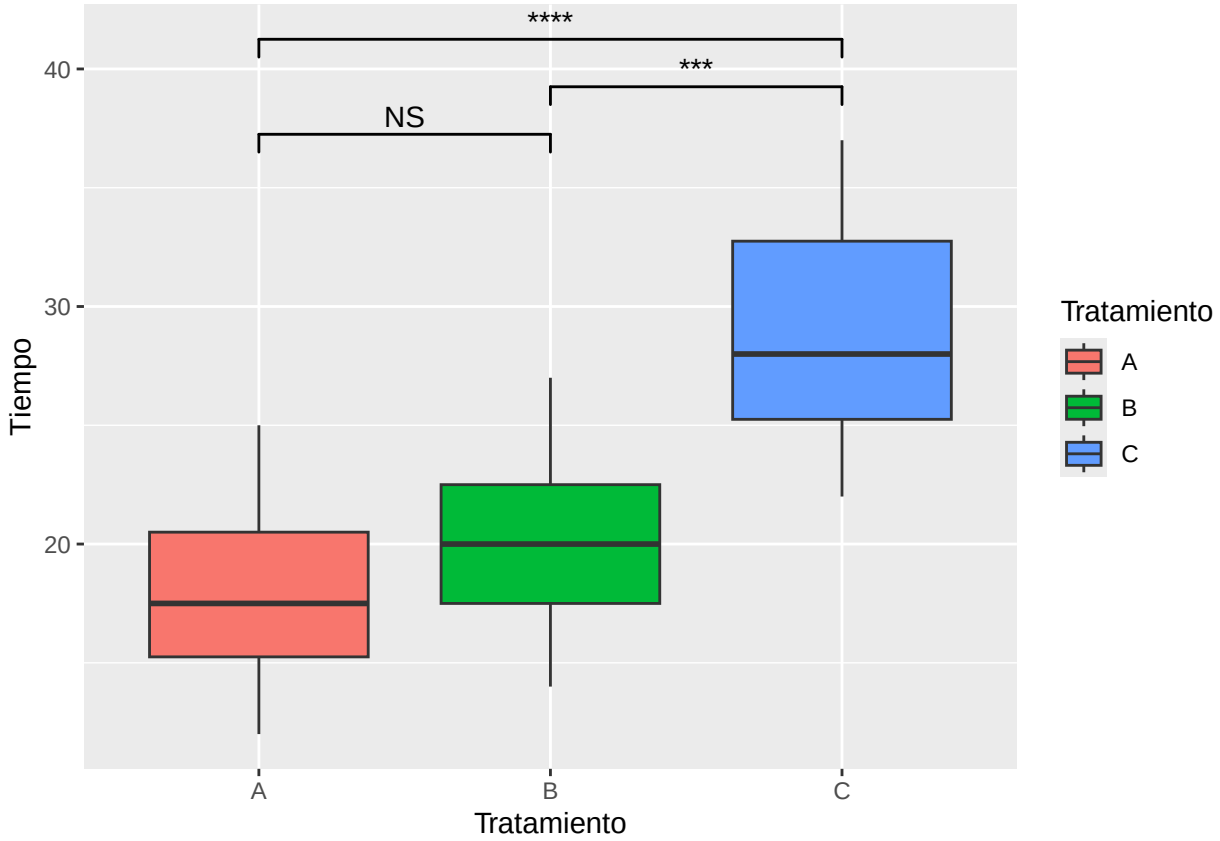
term	contrast	null.value	estimate	conf.low	conf.high	adj.p.value
Tratamiento	B-A	0	2.2	-2.758376	7.158376	0.5222392
Tratamiento	C-A	0	10.8	5.841624	15.758376	0.0000302
Tratamiento	C-B	0	8.6	3.641624	13.558376	0.0005664

Se observa que existe diferencia estadísticamente significativa ($p<0.05$), entre A y C, B y C únicamente.

Graficamos los tiempos por tratamiento:

```
my.boxplot.T<-ggplot(df.6b) +
  aes(x = Tratamiento, y = Tiempo, fill = Tratamiento) +
  geom_boxplot()

my.boxplot.T+
geom_signif(comparisons = list(c("A", "B"), c("A", "C"), c("B","C")),
              map_signif_level = TRUE,
             y_position = c(36, 40, 38), annotation = c("NS","****", "***"))
```



2. Realizar un contraste ANOVA dos factores para saber si el Tiempo depende del Género y del Tratamiento.

Para realizar un contraste ANOVA de dos factores SIN INTERACCIÓN se puede utilizar tanto la función aov

como la función lm pero incluyendo la fórmula del modelo vd ~ f1 + f2, donde vd es la variable dependiente, f1 es el primer factor y f2 el segundo.

```
library(knitr)
library(broom)
aov(Tiempo ~ Genero + Tratamiento, data = df.6b) |>
    tidy() |>
    kable()|> kable_styling()
```

term	df	sumsq	meansq	statistic	p.value
Genero	1	313.6333	313.633333	36.03919	2.4e-06
Tratamiento	2	651.4667	325.733333	37.42958	0.0e+00
Residuals	26	226.2667	8.702564	NA	NA

El p-valor correspondiente al Género es menor al nivel de significancia de 0.05, por lo tanto rechazamos la hipótesis nula y se concluye que el tiempo depende del género. Por otro lado, el p-valor correspondiente la Tratamiento es casi cero, que es menor al nivel de significancia de 0.05, entonces podemos rechazar la hipótesis nula y se concluye que el tiempo depende del tratamiento.

3. Incluir en el modelo anterior también la interacción entre el Género y el Tratamiento. ¿Es significativa la interacción entre los dos factores?

Para realizar un contraste ANOVA de dos factores CON INTERACCIÓN se puede utilizar tanto la función aov como la función lm pero incluyendo la fórmula del modelo vd ~ f1 * f2, donde vd es la variable dependiente, f1 es el primer factor y f2 el segundo.

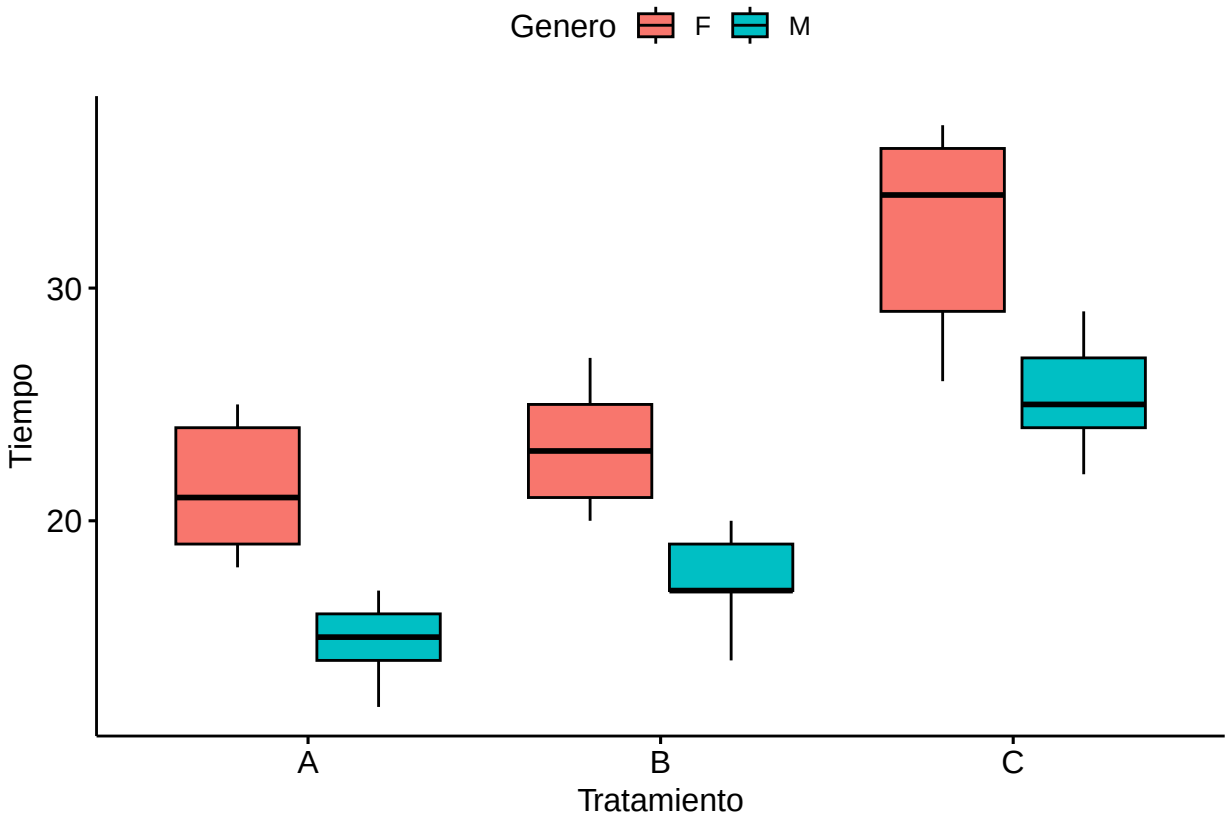
```
aov(Tiempo ~ Genero * Tratamiento, data = df.6b) |>
    tidy() |>
    kable()|> kable_styling()
```

term	df	sumsq	meansq	statistic	p.value
Genero	1	313.633333	313.6333333	33.5436720	0.0000057
Tratamiento	2	651.466667	325.7333333	34.8377897	0.0000001
Genero:Tratamiento	2	1.866667	0.9333333	0.0998217	0.9053725
Residuals	24	224.400000	9.3500000	NA	NA

La interacción entre los 2 factores no es significativa dado que su p-valor es 0.90 que es mayor al nivel de significancia de 0.05.

Podemos complementar estos resultados con diagramas de caja:

```
bxp.TG <- ggboxplot(df.6b, x = "Tratamiento", y = "Tiempo", fill = "Genero",
                    )
bxp.TG
```



Observe que existe el mismo patrón de tiempo en todos los tratamientos, tanto en hombres como en mujeres (efecto del tratamiento). También hay un efecto del género: específicamente, los tiempos de las mujeres son mas altos comparados con los de los hombres para todos los tratamientos.

4. Hacer ANOVA y prueba de Tukey para conocer si existen diferencias significativas entre los tiempos por género, para cada tratamiento:

Usamos la función `tukey_hsd` del paquete `rstatix`. Esta función realiza ANOVA usando la función `aov` seguido de una prueba de Tukey.

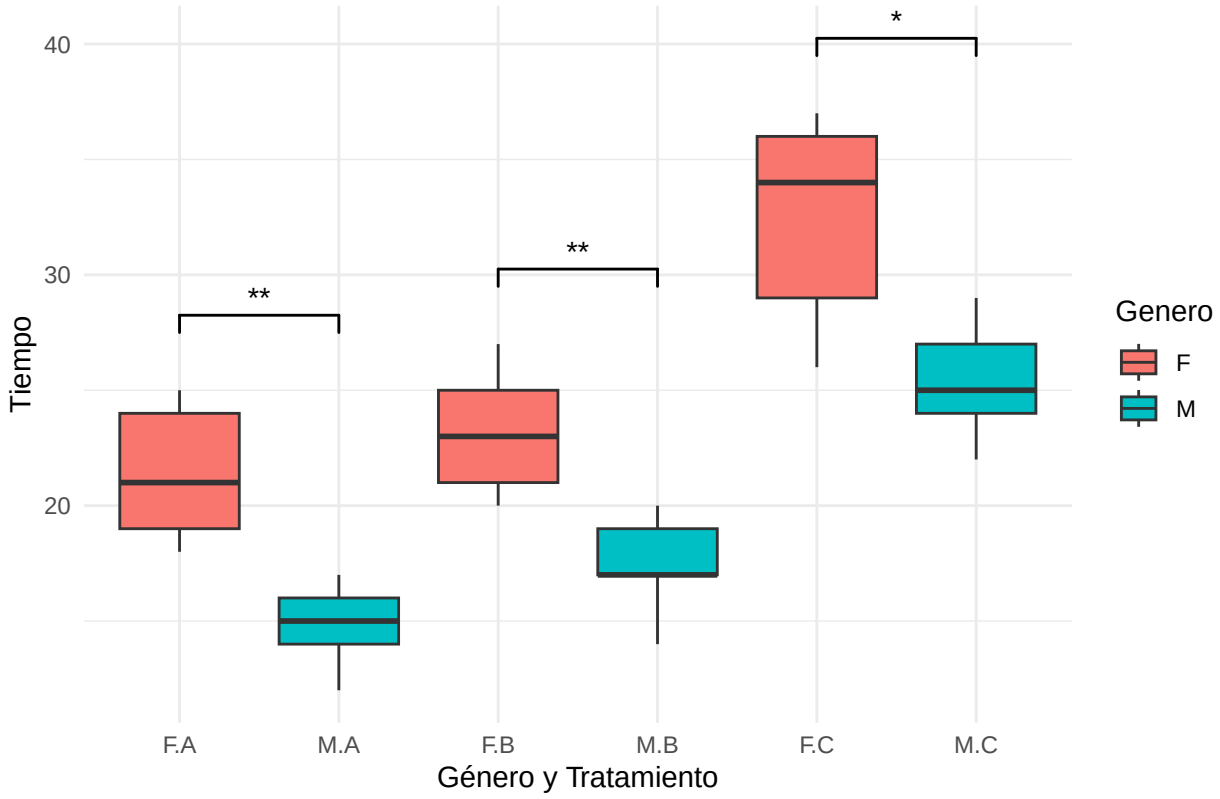
```
df.6b |>
  group_by(Tratamiento) |>
  tukey_hsd(Tiempo ~ Genero)|>
       kable()|> kable_styling()
```

Tratamiento	term	group1	group2	null.value	estimate	conf.low	conf.high	p.adj	p.adj.signif
A	Genero	F	M	0	-6.6	-10.318320	-2.881680	0.00347	**
B	Genero	F	M	0	-5.8	-9.589151	-2.010848	0.00773	**
C	Genero	F	M	0	-7.0	-12.610750	-1.389250	0.02060	*

A continuación, generamos el diagrama de caja, con los niveles de significancia:

```
# Create the boxplot
ggplot(df.6b, aes(x = interaction(Genero,Tratamiento), y = Tiempo, fill = Genero)) +
  geom_boxplot() +
  labs(title = "",
       x = "Género y Tratamiento",
       y = "Tiempo") +
  theme_minimal() +
```

```
  geom_signif(comparisons = list(c("M.A", "F.A"), c("M.B", "F.B"), c("M.C", "F.C")),
              map_signif_level = TRUE,
              y_position = c(27, 29, 39),
              test="t.test")   # Adjust y_position as needed
```



Ejemplo 13: ANOVA de 2 factores

Supongamos que el mismo ensayo clínico del ejemplo anterior se replica en un segundo sitio clínico y se observan los siguientes datos contenidos en el archivo `anova2_02.xlsx`:*https://github.com/ghebrer82/EstadisticaAplicada/blob/Ejemplos/anova2_02.xlsx*

```
df.6c <- read_excel("anova2_02.xlsx")
```

1. Hacer ANOVA de dos factores

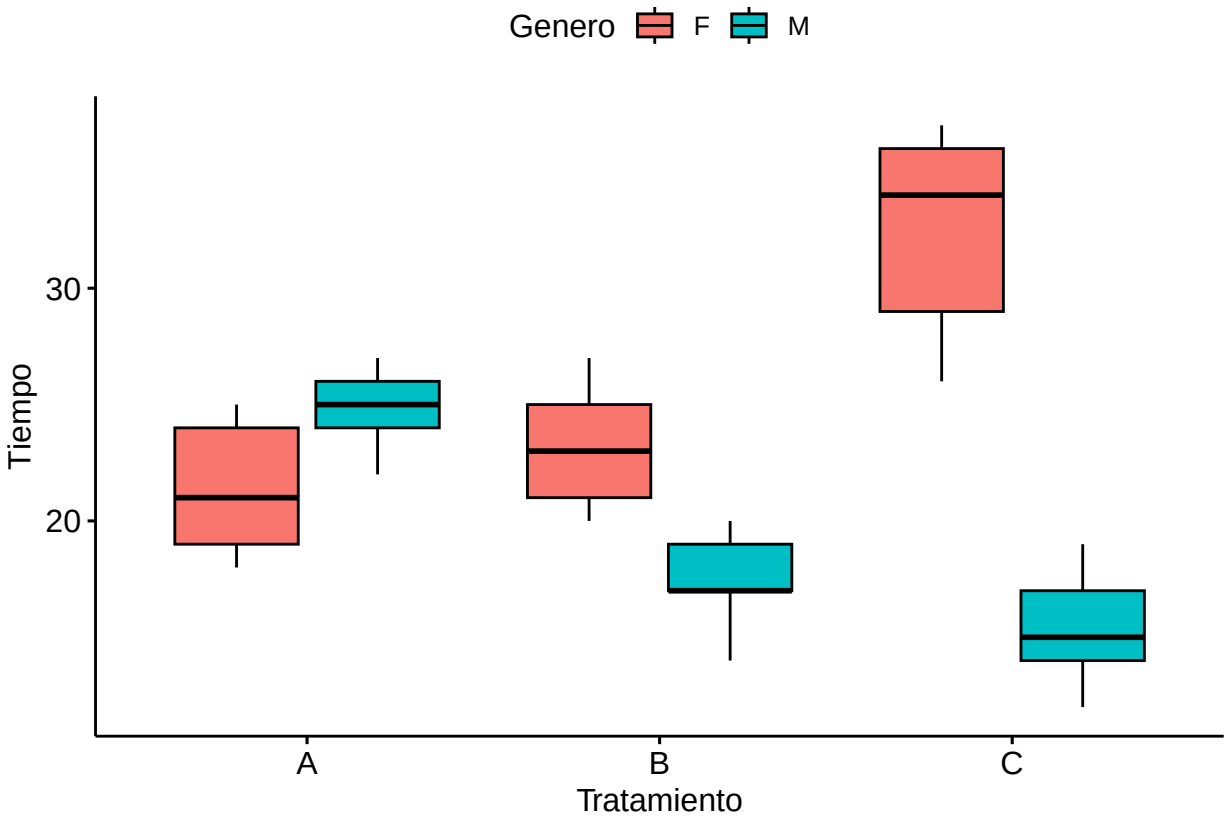
```
aov(Tiempo ~ Genero * Tratamiento, data = df.6c) |>
    tidy() |>
    kable()|>  kable_styling()
```

term	df	sumsq	meansq	statistic	p.value
Genero	1	313.63333	313.63333	33.543672	0.0000057
Tratamiento	2	71.46667	35.73333	3.821747	0.0362345
Genero:Tratamiento	2	521.86667	260.93333	27.907308	0.0000005
Residuals	24	224.40000	9.35000	NA	NA

Observamos que hay un efecto significativo del tratamiento ($p<0.05$) , y un efecto altamente significativo del género y de la interacción entre los dos factores ($p<0.0001$).

2. Genere los diagramas de caja de los tiempos de cada género para cada tratamiento:

```
bxp.TG.3 <- ggboxplot(df.6c, x = "Tratamiento", y = "Tiempo", fill = "Genero",
                      )
bxp.TG.3
```



Observe que el efecto del tratamiento varía según el género. En el tratamiento A el tiempo es mayor en los hombres, mientras que en B y C el tiempo es mayor en las mujeres. Por lo tanto, no podemos generalizar el efecto del tratamiento.

Ejercicios

1. Un gimnasio desea evaluar la eficacia de tres rutinas de ejercicio diferentes (Rutina A, Rutina B, Rutina C) en la pérdida de peso de sus miembros. Para ello, se asignó aleatoriamente a 45 participantes (15 por rutina) a seguir una de las rutinas durante un período de 8 semanas. Al final del período, se registró la pérdida de peso de cada participante en kilogramos.

Datos: Los datos se encuentran en el archivo 06_perdida_peso.xlsx *(https://github.com/ghebrer82/EstadisticaAplicada/blob/main/06_perdida_peso.xlsx)*

Realice lo siguiente (con un nivel de significancia =0.05):

a) Dibujar el diagrama de cajas (boxplot) con los puntos individuales (jittered points) correspondientes a las pérdidas de peso para cada una de las tres rutinas de ejercicio. A la vista del diagrama, ¿crees que existen diferencias significativas entre las pérdidas de peso de las tres rutinas de ejercicio? (Esta es una pregunta de observación visual, no una conclusión estadística formal).

b) Realizar un contraste ANOVA de un factor para conocer si existen diferencias estadísticamente significativas entre las pérdidas de peso medias de las tres rutinas de ejercicio. Formular las hipótesis nula y alternativa. Mostrar la tabla ANOVA. Indicar el p-value. Tomar una decisión (rechazar o no rechazar la hipótesis nula) y escribir una conclusión en el contexto del problema.

c) Realizar un contraste post-hoc de comparación de las medias de pérdida de peso por pares (utilizando la corrección de Bonferroni o Tukey).¿Entre qué rutinas de ejercicio existe una diferencia estadísticamente significativa en la pérdida de peso media?

2. Una investigadora está interesada en determinar el efecto de dos factores (tipo de dieta y tipo de ejercicio) sobre la pérdida de peso en un grupo de individuos. Se seleccionaron 40 participantes y se asignaron aleatoriamente a una de cuatro combinaciones posibles de dieta y ejercicio.

Factores:

Dieta (Factor A): Dieta Baja en Carbohidratos (D1) Dieta Mediterránea (D2) Ejercicio (Factor B): Ejercicio Aeróbico (E1) Ejercicio de Fuerza (E2)

Cada combinación de dieta y ejercicio tiene 10 participantes, y la variable de respuesta es la pérdida de peso (en kg) después de 3 meses.

Datos: Los datos se encuentran en el archivo 06b_perdida_peso_2.xlsx *(https://github.com/ghebrer82/EstadisticaAplicada/blob/main/06B_perdida_peso_2.xlsx)*

Responder lo siguiente: a) Dibujar el diagrama de cajas para visualizar la distribución de la pérdida de peso por cada combinación de dieta y ejercicio. b) Realizar un contraste ANOVA de dos factores SIN interacción para evaluar el efecto principal de la dieta y el ejercicio en la pérdida de peso. c) Realizar un contraste ANOVA de dos factores CON interacción para evaluar el efecto principal de la dieta, el ejercicio y la posible interacción entre ellos en la pérdida de peso. d) Hacer ANOVA y prueba de Tukey para conocer si existen diferencias significativas entre los niveles de los factores (si el ANOVA lo justifica).

Bibliografía

- Estadística, Mario Triola, 12va Edición, Pearson, 2018.
- Probabilidad y Estadística para Ingeniería y Ciencias, Ronald Walpole, Pearson Educación, 2012.

7 Regresión lineal Simple

Introducción

La regresión lineal es una de las técnicas estadísticas más fundamentales. Evalúa la relación entre la variable dependiente y una o más variables independientes. En el análisis de regresión, la variable dependiente se denota como "Y" y las variables independientes como "X".

Existen condiciones para modelar los datos mediante regresión lineal. En primer lugar, las dos variables deben ser cuantitativas. En segundo lugar, la relación entre las dos variables debe ser lineal. En tercer lugar, no debe haber valores atípicos en los datos.

La forma más directa de comprobar las condiciones de la regresión lineal es mediante la creación de un diagrama de dispersión. El diagrama de dispersión representa gráficamente cada par ordenado (variable independiente, variable dependiente), lo que permite ver si los puntos de datos se encuentran más o menos en línea recta (con una tendencia positiva o negativa), si tienen una relación curva (no lineal) o si no tienen relación y parecen una nube de puntos.

Una vez que haya determinado que la regresión lineal es apropiada para modelar sus datos, se deben calcular la pendiente e intersección y predecir la respuesta media para un valor particular de la variable independiente.

El coeficiente de correlación (también conocido como correlación o r, indistintamente) es un concepto clave en estadística. Mide la fuerza de la relación lineal entre dos variables cuantitativas. El coeficiente de determinación (también llamado R^2) está matemáticamente relacionado con el coeficiente de correlación (R^2 es el cuadrado de la correlación) y proporciona información adicional sobre la relación. La coeficiente de correlación r, mide la relación entre las dos variables en una escala de -1 a 1, y 0 es la ausencia de relación. R^2 se mide en una escala de 0 a 1 y generalmente se expresa como un porcentaje del 0% al 100%. R^2 nos dice el porcentaje de variación en la variable de respuesta que se explica por las diferencias en la variable explicativa. Si R^2 está cerca del 100%, entonces saber cuál es la variable independiente nos dice casi todo lo que necesitamos saber para estimar el valor de la variable dependiente. Por el contrario, si R^2 es pequeño, entonces falta información en nuestro modelo que, si estuviera presente, nos permitiría hacer un mejor trabajo prediciendo la variable dependiente.

Objetivos de aprendizaje

Después de completar esta práctica, el estudiante podrá:

- Generar un modelo de regresión lineal simple.
- Calcular e interpretar coeficientes en un análisis de regresión lineal
- Calcular e interpretar un coeficiente de correlación y de determinación.
- Interpretar diagramas de dispersión, incluyendo la identificación de la distinción entre variables independientes y dependientes y la forma de asociación entre ellas.

Material

- Computadora con el software R instalado o acceso a R Studio Cloud (https://posit.cloud/)
- Para la realización de esta práctica se requieren los siguientes paquetes:

```
library(tidyverse)
library(broom)
library(knitr)
library(kableExtra)
library(ggplot2)
library(readxl)
```

Ejemplo 14: Contenido de alcohol en sangre

En un experimento que llevó a cabo la Ohio State University (OSU), 16 estudiantes se ofrecieron como voluntarios para participar en el experimento. Cada estudiante sopló en un alcoholímetro para indicar que su concentración inicial de alcohol en sangre era cero. El número (entre 1 y 9) de cervezas de 12 onzas que debían beber se asignó a cada uno de los sujetos sacando boletos de un cuenco. Treinta minutos después de consumir su última cerveza, un oficial de policía del departamento de policía de OSU midió la concentración de alcohol en sangre de los estudiantes. El oficial también realizó una prueba de sobriedad antes y después del consumo de alcohol. Esto implicaba realizar cuatro tareas simples, calificadas en una escala del 1 al 10, que demostraban coordinación: mantener el equilibrio sobre un pie, tocarse la punta de la nariz con el dedo índice, colocar la cabeza hacia atrás con los ojos cerrados y caminar en puntas. El oficial de policía no sabía cuánto alcohol había consumido cada sujeto.

Datos: *https://github.com/ghebrer82/EstadisticaAplicada/blob/Ejemplos/BLOODALC.xlsx*

Primero, leemos los datos:

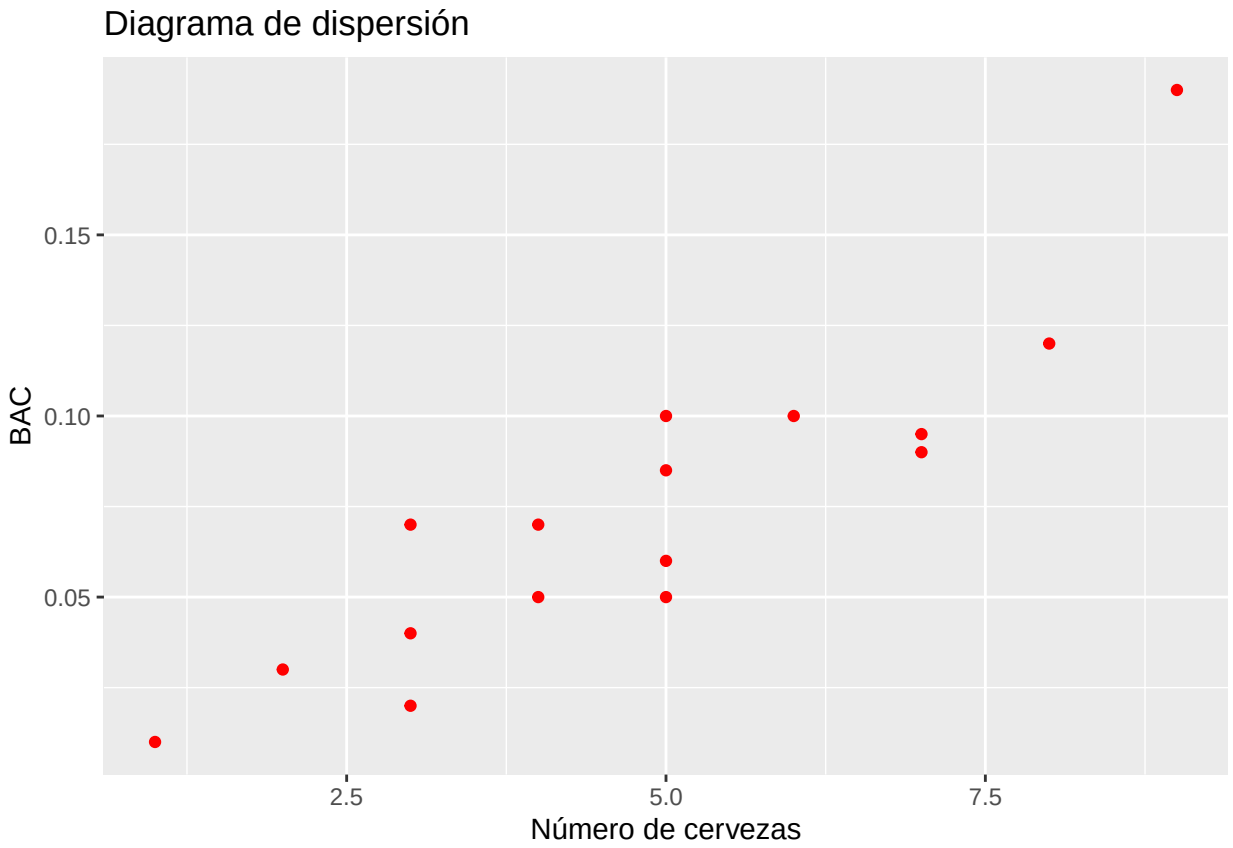
```
df.7b<-read_excel("BLOODALC.xlsx")
```

los datos contienen la siguiente información:

- ID = número de identificación
- Gender= Género
- Weight = Peso de cada sujeto en libras
- Beers = Número de cervezas de 12 onzas consumidas
- BAC = contenido de alcohol en sangre
- 1st-Sobriety = puntuación combinada en las cuatro pruebas de sobriedad antes del consumo de alcohol
- 2nd-Sobriety = puntuación combinada en las cuatro pruebas de sobriedad después del consumo de alcohol

1. Hacer un diagrama de dispersión del nivel de alcohol en sangre (BAC) en función del número de cervezas consumidas. ¿Crees que el número de cervezas consumidas sería un buen predictor del nivel de alcohol en sangre? ¿Por qué?

```
ggplot(df.7b, aes(x = df.7b$Beers, y = df.7b$BAC, na.rm=TRUE)) +
    geom_point(col = "red", na.rm=TRUE) +
    labs(title = "Diagrama de dispersión", x = "Número de cervezas", y = "BAC")
```



La cantidad de cervezas consumidas sería un buen predictor porque tiene una fuerte relación con el nivel de alcohol en sangre. A medida que aumenta la cantidad de cervezas consumidas, también aumenta el nivel de alcohol en sangre.

2. Para cada situación que se muestra a continuación, haga un diagrama de dispersión que muestre la relación entre el nivel de alcohol en la sangre y la diferencia en los puntajes de sobriedad (debe crear esta variable).

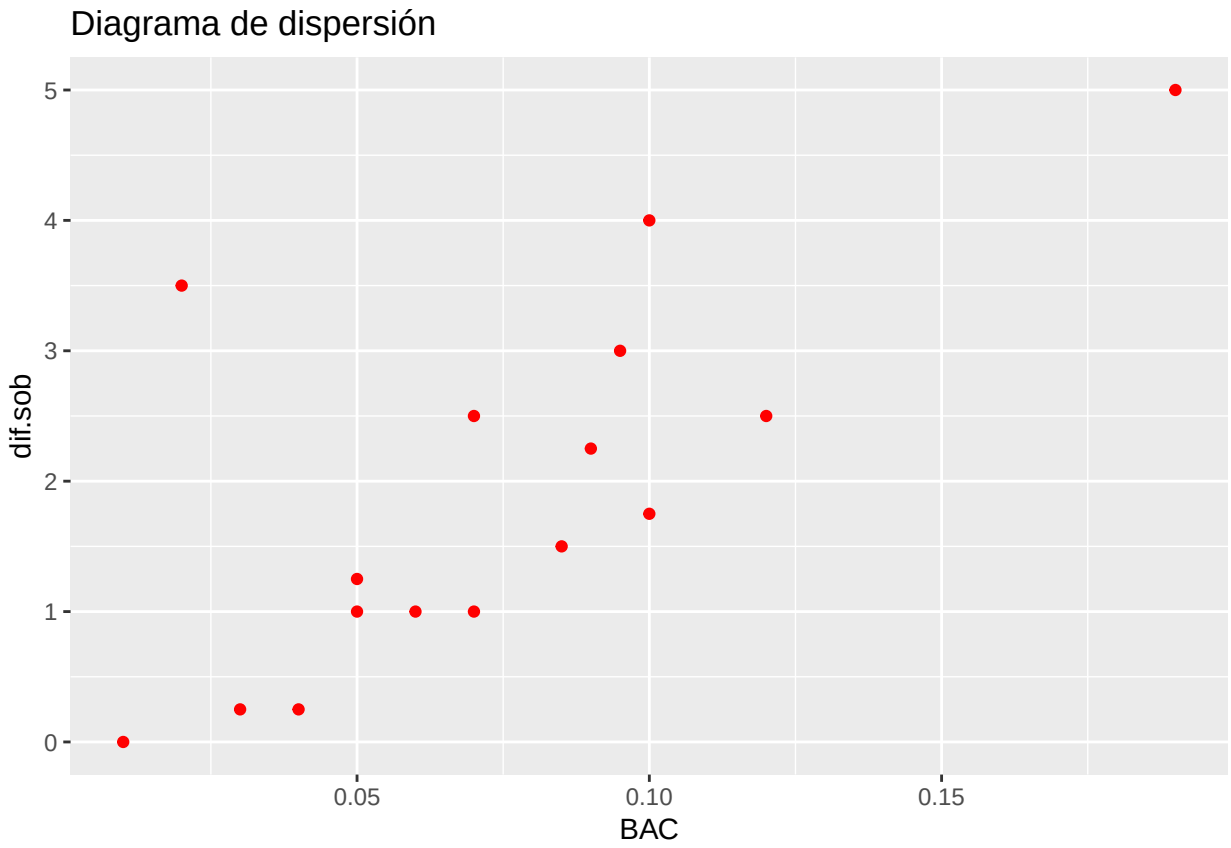
a) Suponga que desea saber cuánto se deteriora la coordinación debido al aumento de alcohol en el cuerpo.
b) Es ilegal conducir con un contenido de alcohol en sangre de 0.1. Suponga que desea saber qué tan confiable es la prueba de sobriedad para determinar el nivel de alcohol en la sangre.

Primero creamos la nueva variable llamada "dif.sob":

```
df.7b$dif.sob <- abs(df.7b$`2nd-Sobr` - df.7b$`1st-Sobr`)
```

Luego, generamos la gráfica de dispersión para para el inciso a):

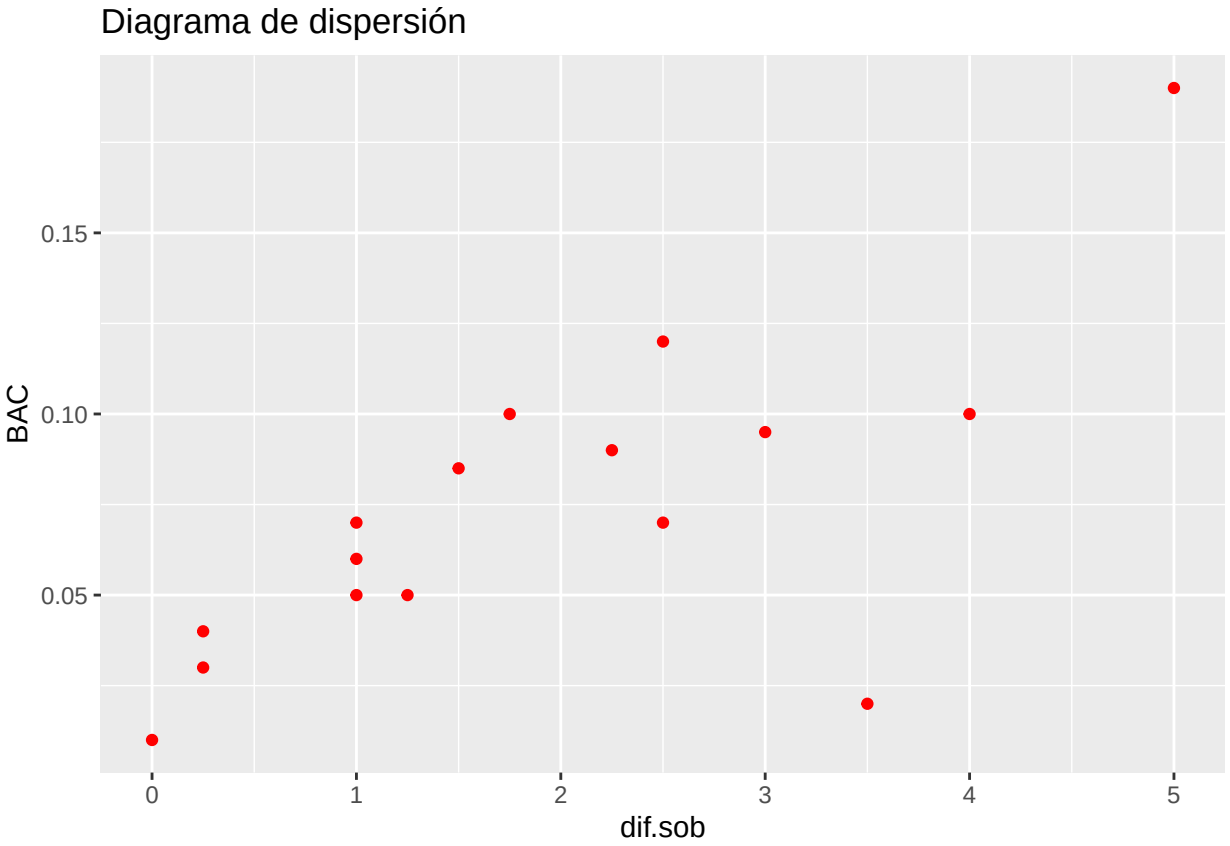
```
ggplot(df.7b, aes(x = BAC, y = dif.sob, na.rm=TRUE)) +
    geom_point(col = "red", na.rm=TRUE) +
    labs(title = "Diagrama de dispersión", x = "BAC", y = "dif.sob")
```



En este caso, La variable independiente debe ser el nivel de alcohol en sangre (BAC), mientras que la variable dependiente debe ser la diferencia en los puntajes de la prueba de sobriedad. En general, el gráfico muestra que a medida que aumenta el nivel de alcohol en sangre, aumenta el deterioro de la coordinación.

A continuación, generamos la gráfica de dispersión para para el inciso b):

```
ggplot(df.7b, aes(x = dif.sob, y = BAC, na.rm=TRUE)) +
    geom_point(col = "red", na.rm=TRUE) +
    labs(title = "Diagrama de dispersión", x = "dif.sob", y = "BAC")
```



En ese caso, la variable independiente debe ser la diferencia entre los puntajes de la prueba de sobriedad, mientras que la variable dependiente debe ser el nivel de alcohol en sangre. En general, el gráfico muestra que cuando la coordinación de un sujeto se deteriora, su nivel de alcohol en sangre se encuentra en un nivel alto.

3. Realice una regresión del BAC sobre la diferencia entre los puntajes de la prueba de sobriedad. ¿Qué valor de BAC se predecirá para un sujeto con una diferencia en los puntajes de la prueba de sobriedad de 2.5? ¿Es confiable esta predicción? Explique.

Para ajustar un modelo de regresión se utiliza la función lm del paquete stats. Esta función requiere que se le pase como parámetro la fórmula del modelo de regresión que debe tener la sintaxis y ~ f(x), donde y es la variable dependiente en el modelo, x es la variable independiente, y f(x) es una expresión matemática que describe el modelo.

```
recta_BAC_dif <- lm(BAC ~ dif.sob, df.7b)
summary(recta_BAC_dif)
```

```
##
## Call:
## lm(formula = BAC ~ dif.sob, data = df.7b)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.088544 -0.010827  0.000295  0.017569  0.048385
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.031377   0.013660   2.297  0.03756 *
## dif.sob     0.022048   0.005766   3.824  0.00186 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03195 on 14 degrees of freedom
## Multiple R-squared:  0.5109, Adjusted R-squared:  0.4759
## F-statistic: 14.62 on 1 and 14 DF,  p-value: 0.001861
```

El resultado de la regresión anterior indica que la ecuación para la línea de mínimos cuadrados en esta configuración es

BAC predicho $= 0.031377 + 0.022048 \times$ Diferencia de sobriedad

Por lo tanto, para una diferencia de puntaje en la prueba de sobriedad de 2.5, predecimos

BAC $= 0.031377 + 0.022048 \times (2.5) = 0.086497$.

O usando la función `predict.lm`:

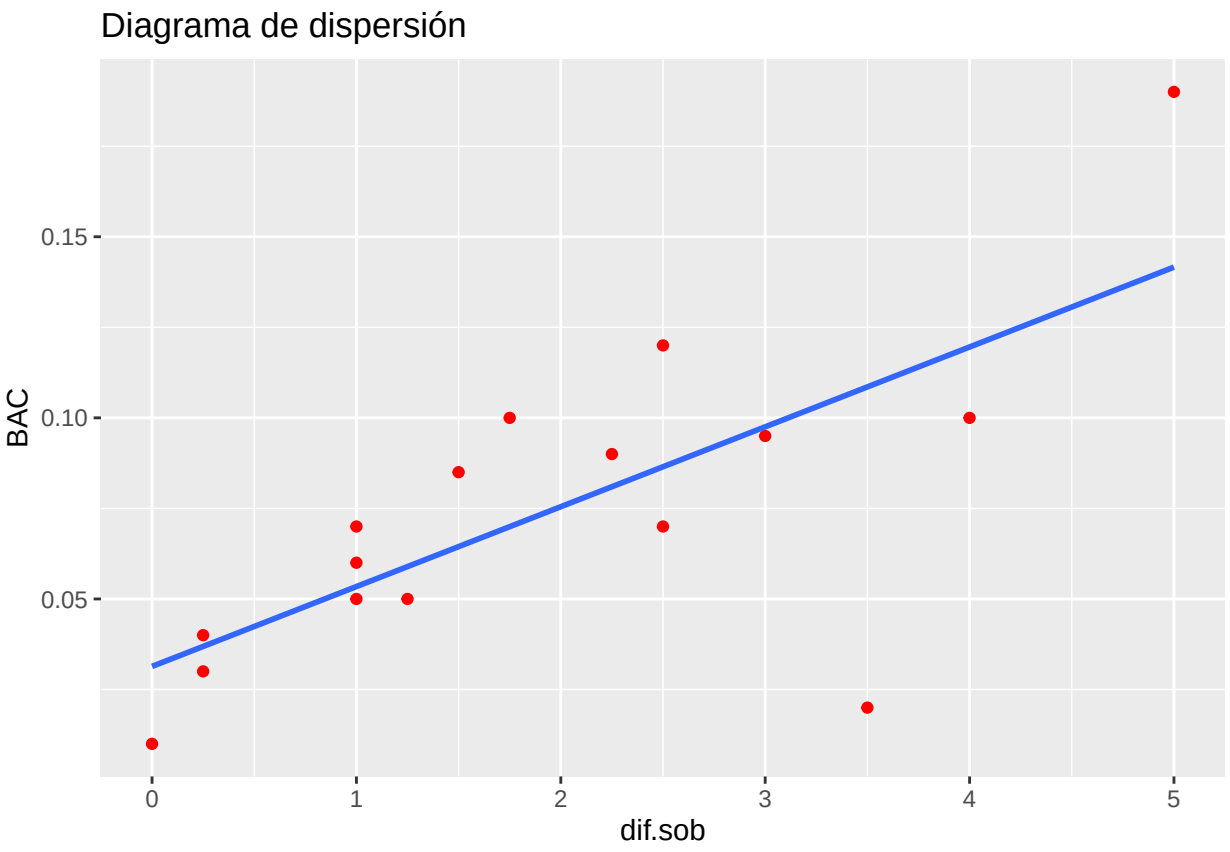
```
as.vector(predict.lm(recta_BAC_dif, newdata = list(dif.sob = 2.5)))
```

```
## [1] 0.08649625
```

El valor de R^2 para esta regresión es solo del 51.1 %. Por lo tanto, hay una gran cantidad de variación en el BAC que no se explica por la diferencia entre los puntajes de la prueba de sobriedad. Esto indica que la predicción puede no ser confiable.

Generamos la gráfica de dispersión con la recta de regresión:

```
ggplot(df.7b, aes(x = dif.sob, y =BAC )) +
    geom_point(col = "red") +
    geom_smooth(method = "lm", se=FALSE) +
    labs(title = "Diagrama de dispersión", x = "dif.sob", y = "BAC")
```



4. Realice una regresión del BAC sobre la cantidad de cervezas consumidas. ¿Qué valor de BAC se predecirá para un sujeto que bebió 4 cervezas? Calcule la desviación estándar de la muestra del BAC y compárela con la estimación de la desviación estándar del BAC sobre la línea de regresión.

```
recta_BAC_beer <- lm(BAC ~ Beers, df.7b)
summary(recta_BAC_beer)
```

```
##
## Call:
## lm(formula = BAC ~ Beers, data = df.7b)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.027118 -0.017350  0.001773  0.008623  0.041027
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.012701   0.012638  -1.005    0.332
## Beers        0.017964   0.002402   7.480 2.97e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02044 on 14 degrees of freedom
## Multiple R-squared:  0.7998, Adjusted R-squared:  0.7855
## F-statistic: 55.94 on 1 and 14 DF,  p-value: 2.969e-06
```

El resultado de la regresión anterior indica que la línea de mínimos cuadrados en esta configuración es

BAC predicho = –0.012701 + 0.017964 × Cervezas.

Por lo tanto, para un sujeto que bebió 4 cervezas, predecimos:

BAC = –0.012701 + 0.017964 × (4) = 0.059155.

usando R:

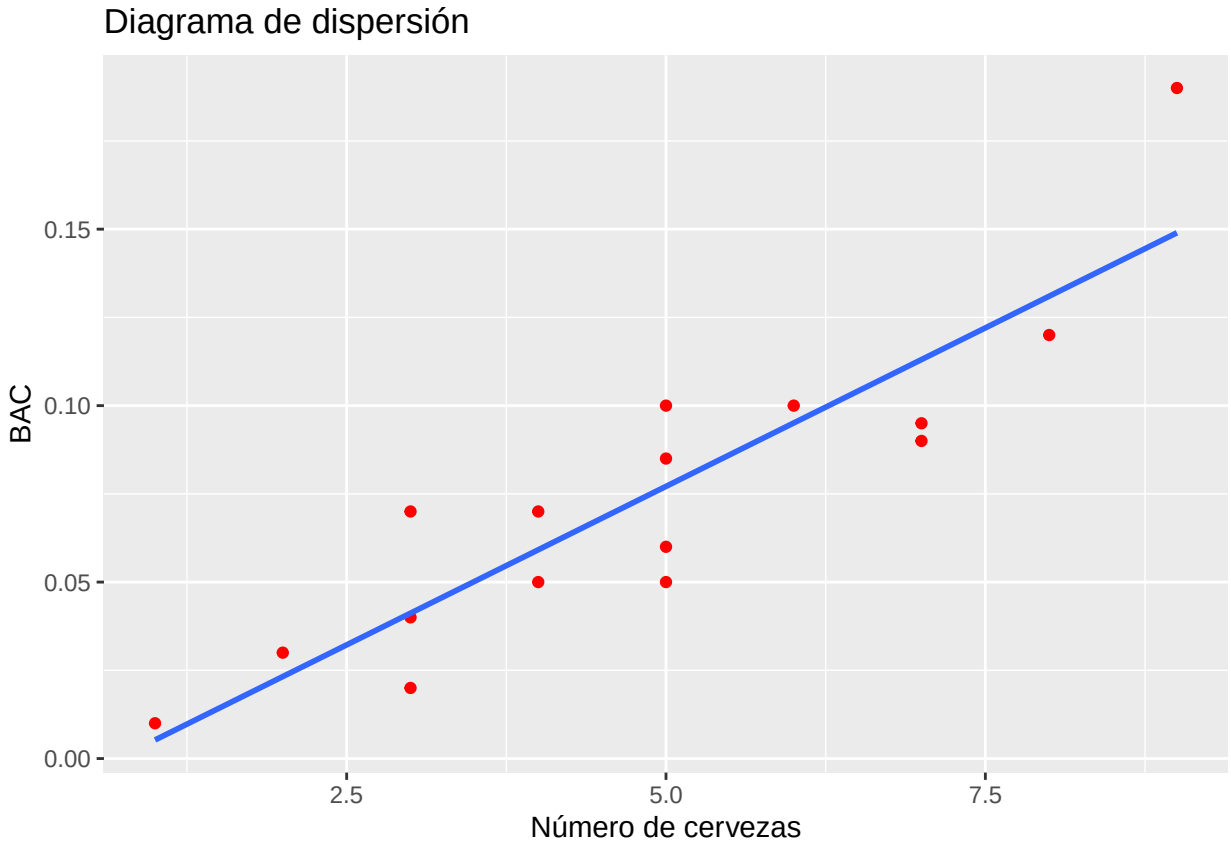
```
as.vector(predict.lm(recta_BAC_beer, newdata = list(Beers = 4)))
```

```
## [1] 0.05915444
```

El valor de R^2 para esta regresión es 79.9 %. Por lo tanto, indica que la predicción puede ser confiable.

Graficamos la regresión de BAC sobre número de cervezas consumidas:

```
BAC_Beers_plot<-ggplot(df.7b, aes(x = Beers, y =BAC )) +
    geom_point(col = "red") +
    geom_smooth(method = "lm", se=FALSE) +
    labs(title = "Diagrama de dispersión", x = "Número de cervezas", y = "BAC")
BAC_Beers_plot
```



5. ¿Cuál de las dos variables independientes (es decir, la cantidad de cervezas consumidas o la diferencia en los puntajes de la prueba de sobriedad) es un mejor predictor del BAC para estos datos?

La cantidad de cervezas consumidas es un mejor predictor del nivel de alcohol en sangre para estos datos. Al comparar los resultados de regresión para los incisos 3 y 4, vemos que:

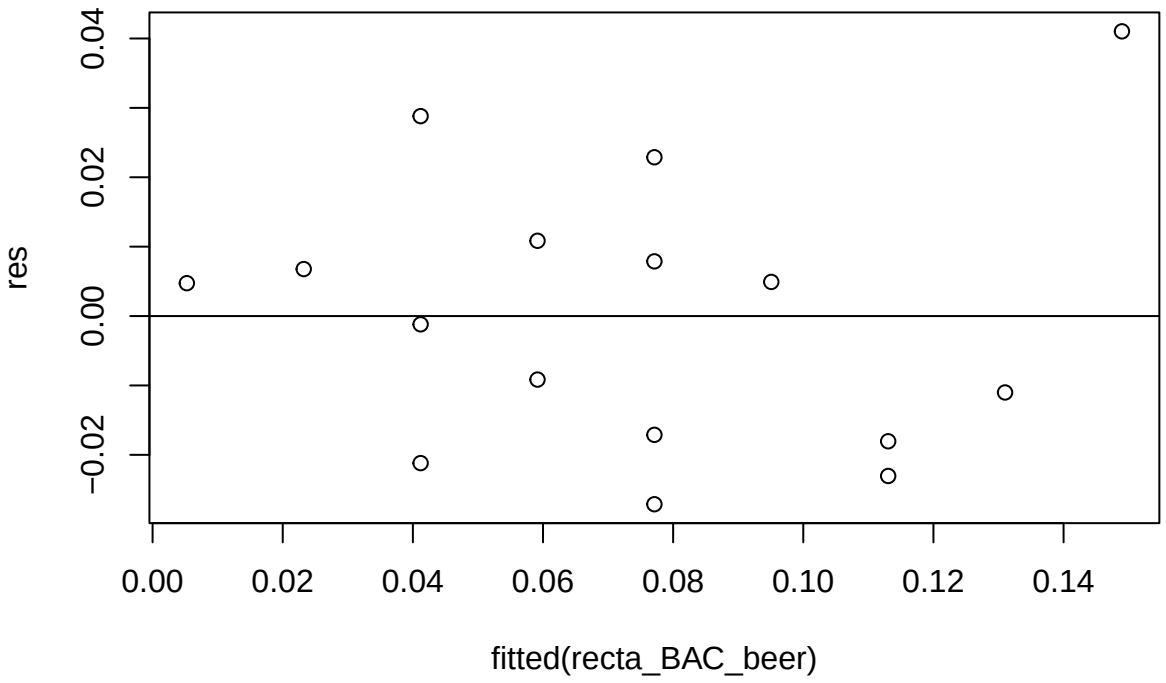
para la regresión del nivel de alcohol en sangre sobre la diferencia entre los puntajes de las pruebas de sobriedad $R^2 = 51.1\ \%$

Mientras que para la regresión del nivel de alcohol en sangre sobre la cantidad de cervezas consumidas $R^2 = 79.9\ \%$

En otras palabras, la cantidad de cervezas consumidas explica más la variación del nivel de alcohol en sangre que la diferencia entre los puntajes de las pruebas de sobriedad. Además, un vistazo a los diagramas de dispersión muestra que los puntos en el diagrama de nivel de alcohol en sangre versus cantidad de cervezas consumidas están más agrupados alrededor de una línea recta que aquellos en el diagrama de nivel de alcohol en sangre versus diferencia entre los puntajes de las pruebas de sobriedad.

6. Elabore un diagrama de dispersión de los valores residuales frente a los valores previstos para la regresión de la Pregunta 5. ¿Este diagrama de dispersión muestra que puede haber problemas con la regresión?

```
res <- resid(recta_BAC_beer)
plot(fitted(recta_BAC_beer), res)
abline(0, 0)  # Adds a horizontal line at 0
```



No hay un patrón en esta gráfica. Todos los puntos se encuentran dentro de un error de 2 RSE (es decir, 2 × (0.0204) = 0.0408) desde cero. Por lo tanto, este gráfico no sugiere ningún problema con la regresión.

Nota: El error estándar residual (RSE) se utiliza para medir el grado de ajuste de un modelo de regresión a un conjunto de datos. En términos simples, mide la desviación estándar de los residuos en un modelo de regresión

7. Haga una regresión del BAC en la relación cerveza-peso. ¿Es la relación cerveza-peso un mejor predictor de BAC que el número de cervezas consumidas?

Primero, deberá crear una nueva variable a partir de una proporción de las variables numero de cervezas (Beers) y el peso (Weight) de la siguiente manera. Sea la relación cerveza-peso (B/W):

Relación B/W = Beers/(Weight + 120)

Creamos la nueva variable de relación cerveza peso:

```
df.7b$b.w<- (df.7b$Beers / (df.7b$Weight_OSU + 120))
```

Hacemos la regresión de BAC sobre la relación cerveza-peso (B/W):

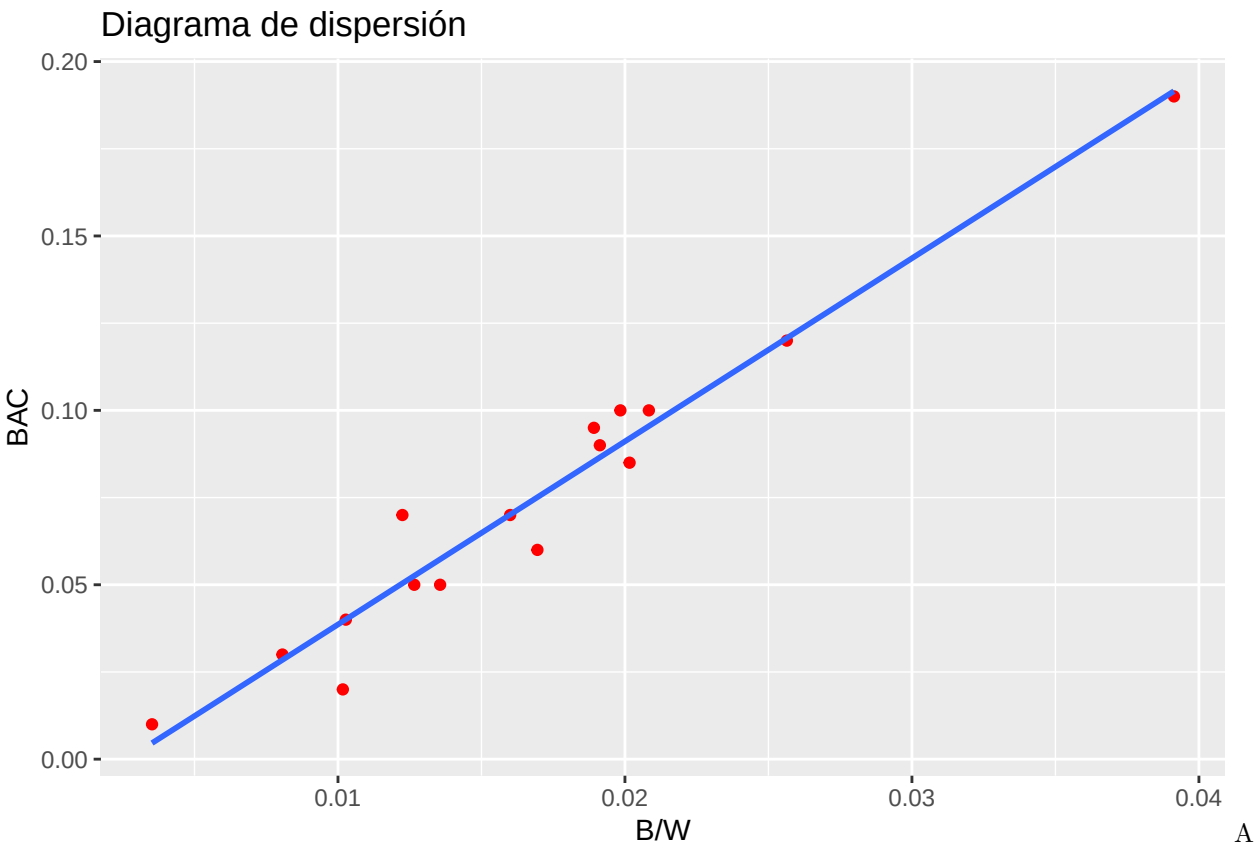
```
recta_BAC_BW <- lm(BAC ~ b.w, df.7b)
summary(recta_BAC_BW)
```

```
##
## Call:
## lm(formula = BAC ~ b.w, data = df.7b)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0195068 -0.0036659 -0.0000831  0.0047369  0.0195995
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.013873   0.005717  -2.427   0.0293 *
## b.w          5.248971   0.309163  16.978 9.78e-11 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.009833 on 14 degrees of freedom
## Multiple R-squared:  0.9537, Adjusted R-squared:  0.9504
## F-statistic: 288.3 on 1 and 14 DF,  p-value: 9.782e-11
```

Y hacemos el gráfico de dispersión con la recta de regresión:

```
BAC_BW_plot<-ggplot(df.7b, aes(x = b.w, y =BAC )) +
    geom_point(col = "red") +
    geom_smooth(method = "lm", se=FALSE) +
    labs(title = "Diagrama de dispersión", x = "B/W", y = "BAC")
BAC_BW_plot
```



A

partir del resultado anterior, la ecuación de regresión dada es:

BAC predicho = –0.013873 + 5.24897 × relación B/W

Esta regresión es incluso mejor que la regresión del BAC sobre el número de cervezas consumidas. El valor R^2 del 95.4% muestra que la predicción es bastante confiable.

8. Haga una predicción del valor de BAC para un sujeto que bebió 3 cervezas y pesa 140 libras.

Primero, calculamos la relación cerveza-peso para alguien que bebió 3 cervezas y pesa 140 libras:

Relación B/W = 3 / (140 + 120) = 0.01154

Por lo tanto, la predicción del BAC para el sujeto dado sería:

```
as.vector(predict.lm(recta_BAC_BW, newdata = list(b.w= 0.01154)))
```

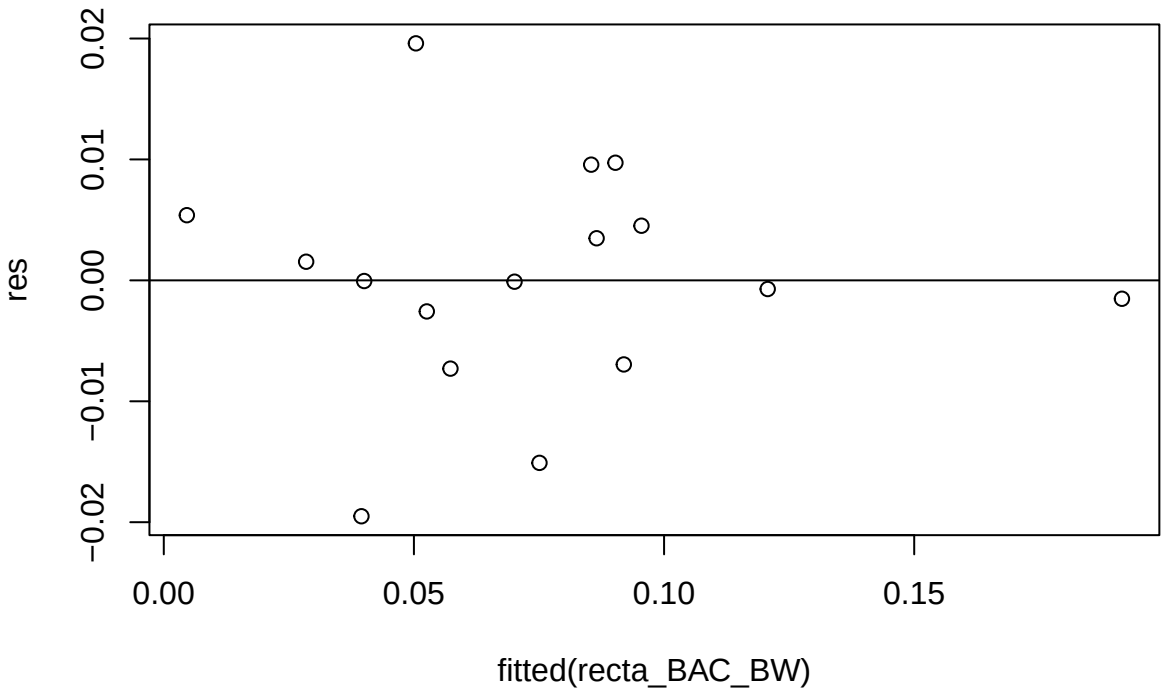
```
## [1] 0.04670052
```

O sustituyendo en la ecuación de la recta de regresión:

BAC = –0.013873 + 5.24897 × (0.01154) = 0.0467

9. Elabore un diagrama de dispersión de los valores residuales frente a los valores previstos para la regresión de la Pregunta 8. ¿Este diagrama de dispersión muestra que puede haber problemas con la regresión?

```
res <- resid(recta_BAC_BW)
plot(fitted(recta_BAC_BW), res)
abline(0, 0)  # Adds a horizontal line at 0
```



No hay un patrón en esta gráfica. Todos los puntos se encuentran dentro de un error de 2 RSE. Por lo tanto, este gráfico no sugiere ningún problema con la regresión.

Ejercicios

1. Un profesor de estadística está interesado en investigar si existe una relación lineal entre el número de horas que un estudiante dedica a estudiar para un examen y la calificación que obtiene en dicho examen. Para ello, ha recopilado datos de 20 estudiantes seleccionados aleatoriamente.

Datos:Los datos se encuentran en el archivo 07_estudio_califica.xlsx *(https://github.com/ghebrer82/EstadisticaAplicada/blob/main/07_estudio_califica.xlsx)*

Responder lo siguiente: a) Generar datos para las 20 observaciones en una hoja de cálculo. b) Realizar la regresión lineal simple para modelar la relación entre las horas de estudio y la calificación del examen. c) Generar la gráfica de dispersión de los datos, incluyendo la recta de regresión ajustada. d) Realizar una predicción: ¿Qué calificación se esperaría obtener si un estudiante estudia 15 horas? e) Calcular e interpretar el valor de R2 (coeficiente de determinación).

Bibliografía

- Estadística, Mario Triola, 12va Edición, Pearson, 2018.
- Probabilidad y Estadística para Ingeniería y Ciencias, Ronald Walpole, Pearson Educación, 2012.

8 Estadística No Paramétrica

Introducción

Todas las pruebas presentadas en las practicas anteriores se denominan pruebas paramétricas y se basan en ciertos supuestos. Por ejemplo, cuando se realizan pruebas de hipótesis para medias de resultados continuos, todas las pruebas paramétricas suponen que el resultado se distribuye de manera aproximadamente normal en la población. Esto no significa que los datos en la muestra observada sigan una distribución normal, sino que el resultado sigue una distribución normal en la población total, lo que no se observa. Para muchos resultados, los investigadores se sienten cómodos con el supuesto de normalidad. Cuando el tamaño de la muestra es pequeño y no se conoce la distribución del resultado y no se puede suponer que se distribuya de manera aproximadamente normal, entonces son adecuadas las pruebas alternativas denominadas pruebas no paramétricas.

Pruebas de normalidad

Para evaluar si una distribución normal modela bien una muestra de datos, se usan pruebas de normalidad de los datos. Algunos ejemplos de los tipos de prueba más comúnmente utilizados para la normalidad de datos incluyen: Kolmogorov–Smirnov (KS), Shapiro–Wilk (S-W), KS corregida de Lilliefors, entre otras.

Cada prueba es esencialmente una prueba de bondad de ajuste y compara los datos observados con los cuantiles de la distribución normal (u otra distribución especificada). La hipótesis nula para cada prueba es H0: los datos siguen una distribución normal frente a H1: los datos no siguen una distribución normal. Si la prueba es estadísticamente significativa ($p < 0.05$), entonces los datos no siguen una distribución normal y se justifica una prueba no paramétrica.

Pruebas paramétricas y no paramétricas

El término paramétrico se refiere a los parámetros de los datos resultantes (distribución), lo que supone que la muestra (media, desviaciones estándar) se distribuye normalmente. Las pruebas paramétricas se basan en la presuposición de que los conjuntos de datos analizados o investigados siguen una distribución normal de valores en forma de "curva de campana" (distribución gaussiana). Por el contrario, los conjuntos de datos no paramétricos tienden a ser sesgados, irregulares o tener algunos valores atípicos. En otras palabras, las muestras o la población pueden parecer no distribuidas normalmente en las pruebas no paramétricas.

Las pruebas no paramétricas son pruebas estadísticas que suponen que los conjuntos de datos disponibles siguen una distribución particular pero no especifica. Las pruebas no paramétricas se aplican principalmente a muestras representadas como datos nominales u ordinales. La mayoría de las pruebas o métodos no paramétricos pueden aplicarse a datos ordinales, datos clasificados, entre otros sin verse totalmente afectados por los valores atípicos. Las pruebas no paramétricas se aplican asumiendo que no requieren ni exigen que se cumpla ninguna condición dada, en particular sobre los parámetros de la población de la que se extrae la muestra. En resumen, se recomienda usar pruebas no paramétricas cuando las observaciones deben extraerse necesariamente de una población que no tenga una distribución normal, cuando los tamaños muestrales sean pequeños, cuando la mediana representa mejor los datos muestreados, o cuando los conjuntos de datos capturados deben medirse esencialmente en una escala nominal u ordinal.

Ejemplos de pruebas no paramétricas incluyen: correlación de Spearman, Chi-cuadrada, Rango con signo de Wilcoxon, U de Mann-Whitney, H de Kruskal–Wallis.

Objetivos de Aprendizaje

Después de completar este práctica, el estudiante podrá:

- Identificar múltiples aplicaciones donde los enfoques no paramétricos son apropiados.
- Realizar e interpretar la prueba U de Mann-Whitney (suma de rangos de Wilcoxon)
- Realizar e interpretar la prueba de Rango con signo de Wilcoxon.
- Realizar e interpretar la prueba H de Kruskal Wallis.
- Identificar el procedimiento de prueba de hipótesis no paramétrica adecuado según el tipo de variable de resultado y la cantidad de muestras.

Material

- Computadora con el software R instalado o acceso a R Studio Cloud (https://posit.cloud/)
- Para la realización de esta práctica se requieren los siguientes paquetes:

```
library(tidyverse)
library(broom)
library(knitr)
library(kableExtra)
library(ggplot2)
library(readxl)
library(dplyr)
library(ggpubr)
library(patchwork)
```

Ejemplo 15: Pruebas de normalidad

La prueba de normalidad la podemos hacer mediante diferentes enfoques.

Enfoque visual: Histograma, gráfico Q-Q Si el histograma tiene aproximadamente forma de "campana", los datos tienen una distribución normal. Si los puntos del gráfico Q-Q caen aproximadamente a lo largo de una línea diagonal recta, entonces los datos están distribuidos normalmente.

Pruebas estadísticas: Shapiro-Wilk, Kolmogorov-Smirnov Si el valor p$>$ 0.05, entonces los datos tienen una distribución normal.

1. Evalúe la normalidad usando a) Histograma, b) Gráfico Q-Q, c)Prueba de Shapiro-Wilk, d) Prueba de Kolmogorov-Smirnov

Generemos dos conjuntos de datos dataset1 y dataset2. Para data set 1 usemos la función `rnom` y para dataset2 usemos `rexp`:

```
set.seed(111)
dataset1 <- data.frame(y=rnorm(200))
dataset2<- data.frame(y=rexp(200, rate=3))
```

a) Histograma

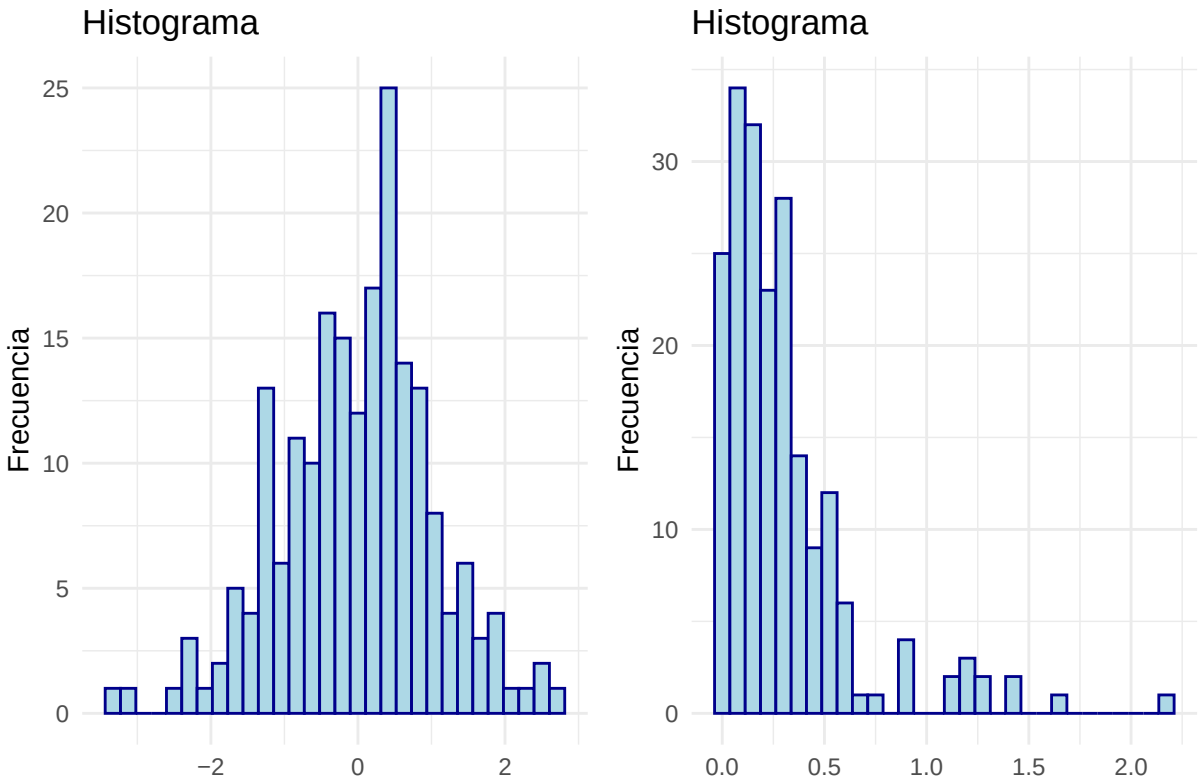
Generamos Histograma para cada conjunto de datos:

```
bar1<-ggplot(dataset1, aes(x=y)) +
  geom_histogram(color="darkblue", fill="lightblue") +
  labs(title = "Histograma",
       x = "",
       y = "Frecuencia") +
  theme_minimal()

bar2<-ggplot(dataset2, aes(x=y)) +
  geom_histogram(color="darkblue", fill="lightblue")+
    labs(title = "Histograma",
       x = "",
       y = "Frecuencia") +
  theme_minimal()
```

Combinamos gráficos en una sola figura:

```
wrap_plots(bar1, bar2, ncol = 2, nrow = 1)
```



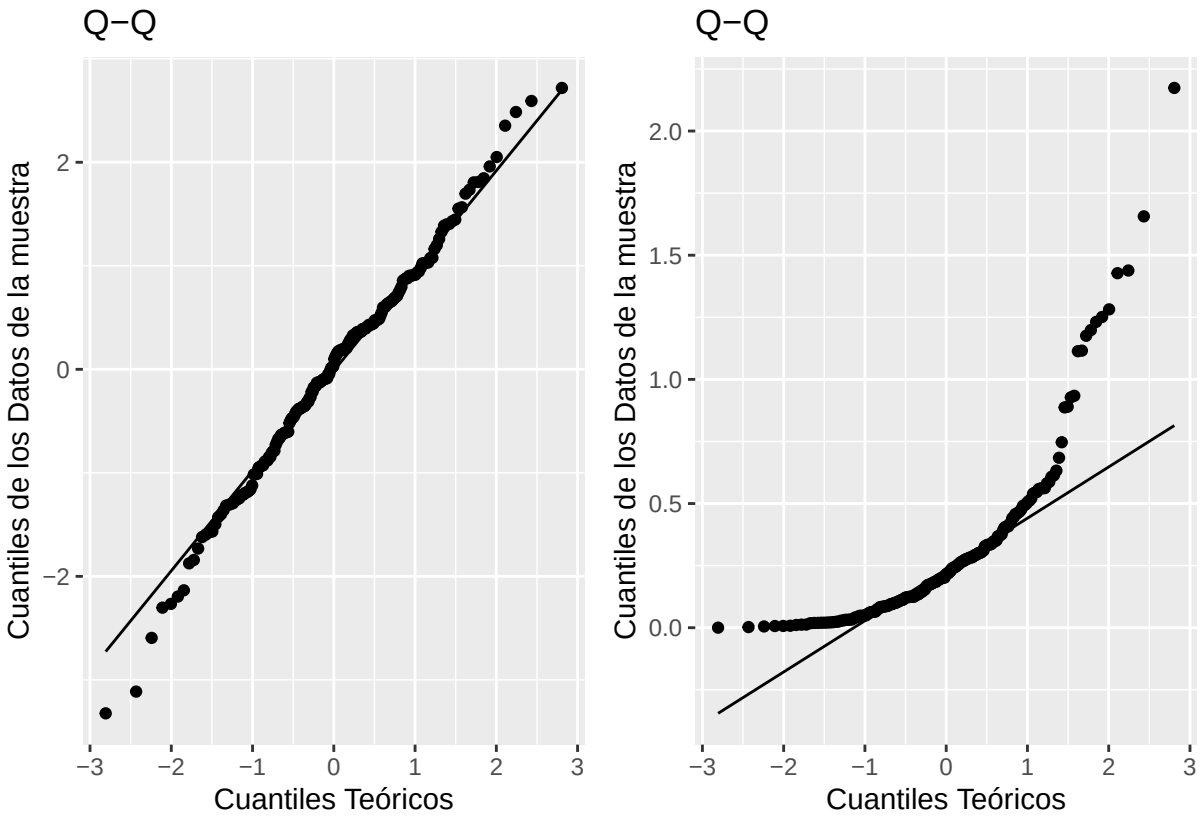
El histograma de la izquierda muestra un conjunto de datos que se distribuye normalmente (aproximadamente en forma de "campana") y el de la derecha muestra un conjunto de datos que no se distribuye normalmente.

b) Gráfico Q-Q

```
qq1<-ggplot(dataset1, aes(sample=y)) +
  labs(title = "Q-Q",
       x = "Cuantiles Teóricos",
       y = "Cuantiles de los Datos de la muestra")+
  stat_qq() +
  stat_qq_line()

qq2<-ggplot(dataset2, aes(sample=y)) +
  labs(title = "Q-Q",
       x = "Cuantiles Teóricos",
       y = "Cuantiles de los Datos de la muestra")+
  stat_qq() +
  stat_qq_line()
```

```
wrap_plots(qq1, qq2, ncol = 2, nrow = 1)
```



El gráfico Q-Q de la izquierda muestra un conjunto de datos que se distribuye normalmente (los puntos caen a lo largo de una línea diagonal recta) y el gráfico Q-Q de la derecha muestra un conjunto de datos que no se distribuye normalmente.

c) Shapiro-Wilk

Usamos la función `shapiro.test`:

```
set.seed(111)
dataset01 <- rnorm(200)
dataset02<- rexp(200, rate=3)
```

Dataset01

```
shapiro.test(dataset01)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dataset01
## W = 0.99279, p-value = 0.4332
```

El valor p del dataset01 no es inferior a 0.05 ($p > 0.05$), lo que indica que los datos se distribuyen normalmente.

Dataset02

```
shapiro.test(dataset02)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dataset02
```

```
## W = 0.7483, p-value < 2.2e-16
```

El valor p del dataset02 es inferior a 0.05 ($p < 0.05$), lo que indica que los datos NO se distribuyen normalmente.

d) Kolmogorov–Smirnov

Usamos las función `ks.test`:

Dataset01

```
ks.test(dataset01,"pnorm")
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  dataset01
## D = 0.049438, p-value = 0.7126
## alternative hypothesis: two-sided
```

El valor p del dataset01 no es inferior a 0.05 ($p > 0.05$), lo que indica que los datos se distribuyen normalmente.

Dataset02

```
ks.test(dataset02,"pnorm")
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  dataset02
## D = 0.50008, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

El valor p del dataset02 es inferior a 0.05 ($p < 0.05$), lo que indica que los datos NO se distribuyen normalmente.

Ejemplo 16: Pruebas de Normalidad

Consideremos un ensayo clínico en el que se pide a los participantes del estudio que califiquen la gravedad de sus síntomas después de 6 semanas de tratamiento asignado. La gravedad de los síntomas podría medirse en una escala ordinal de 5 puntos con opciones de respuesta: los síntomas empeoraron mucho, empeoraron levemente, no cambiaron, mejoraron levemente o mejoraron mucho. Supongamos que hay n=20 participantes en el ensayo, asignados aleatoriamente a un tratamiento experimental o placebo.

Datos: *https://github.com/ghebrer82/EstadisticaAplicada/blob/Ejemplos/clinical_01.csv*

```
df.8<-read.csv("clinical_01.csv")
```

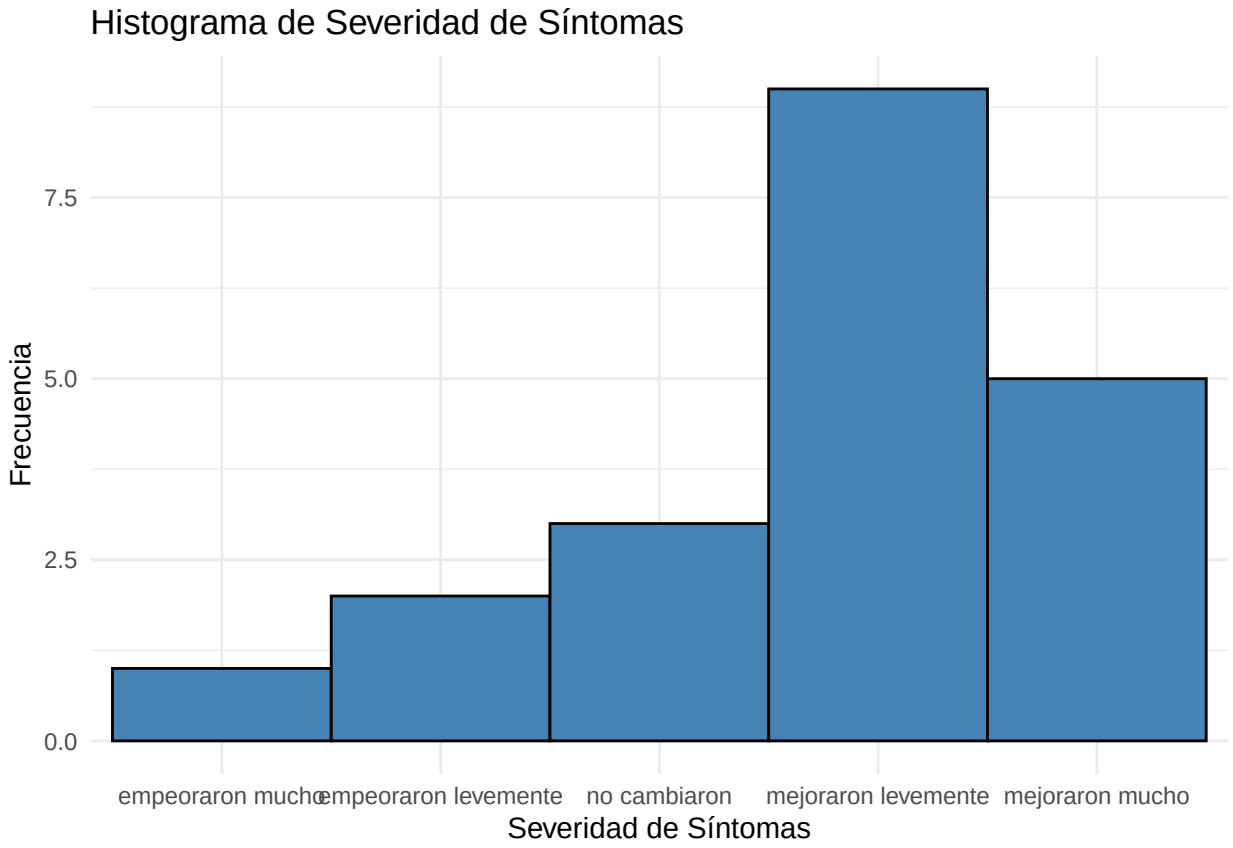
1. Evalúe la normalidad usando a) Histograma, b) Gráfico Q-Q, c)Prueba de Shapiro-Wilk, d) Prueba de Kolmogorov-Smirnov

Primero, cambiamos la variable Severidad de Síntomas a Factor:

```
df.8$Severidad_Síntomas <- factor(df.8$Severidad_Síntomas,
                                  levels = c("empeoraron mucho",
                                             "empeoraron levemente",
                                             "no cambiaron",
                                             "mejoraron levemente",
                                             "mejoraron mucho"))
```

a) Generamos el histograma:

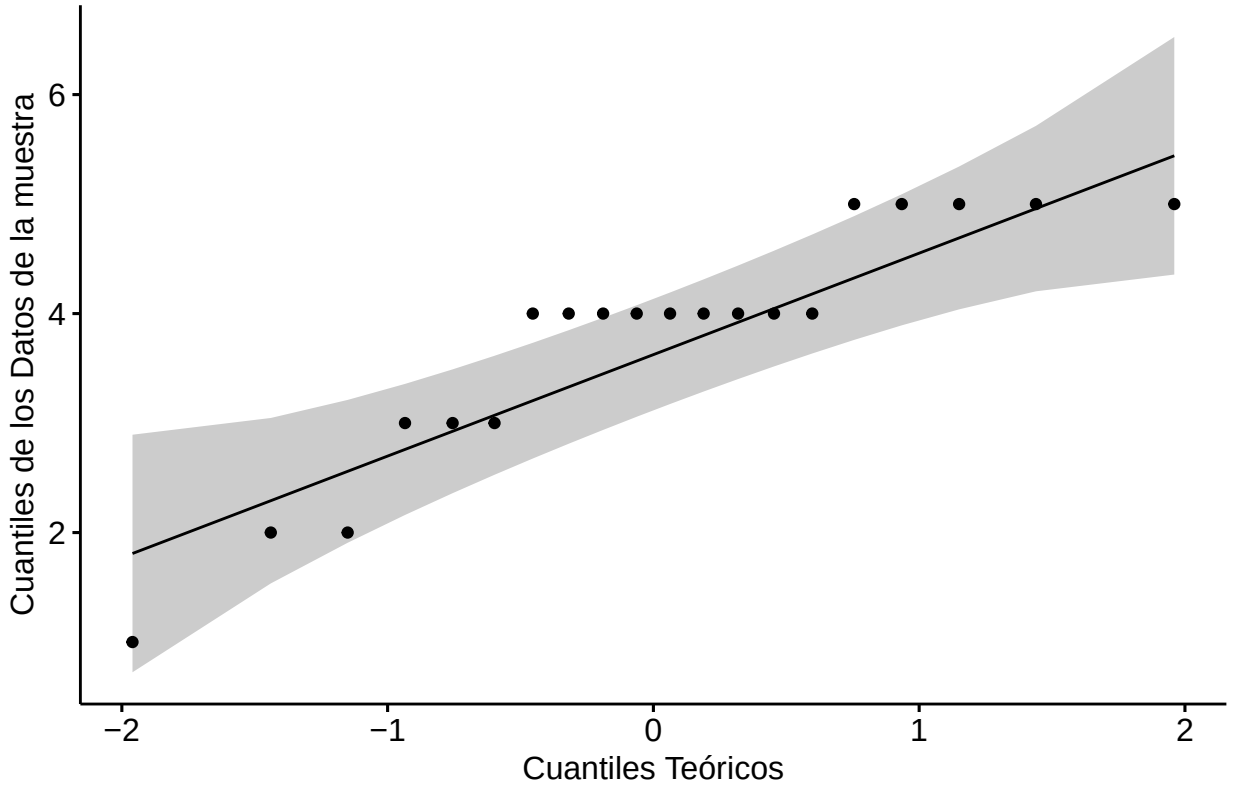
```r
ggplot(df.8, aes(x = Severidad_Síntomas)) +
  geom_bar(fill = "steelblue", color = "black", width = 1) +
  labs(title = "Histograma de Severidad de Síntomas",
       x = "Severidad de Síntomas",
       y = "Frecuencia") +
  theme_minimal()+
   scale_x_discrete(limits = c("empeoraron mucho","empeoraron levemente","no cambiaron",     "mejoraron
```



La distribución de la severidad de los síntomas no parece ser normal, ya que más participantes informan una mejoría en los síntomas en lugar de un empeoramiento de los mismos.

b) Generamos gráfico Q-Q

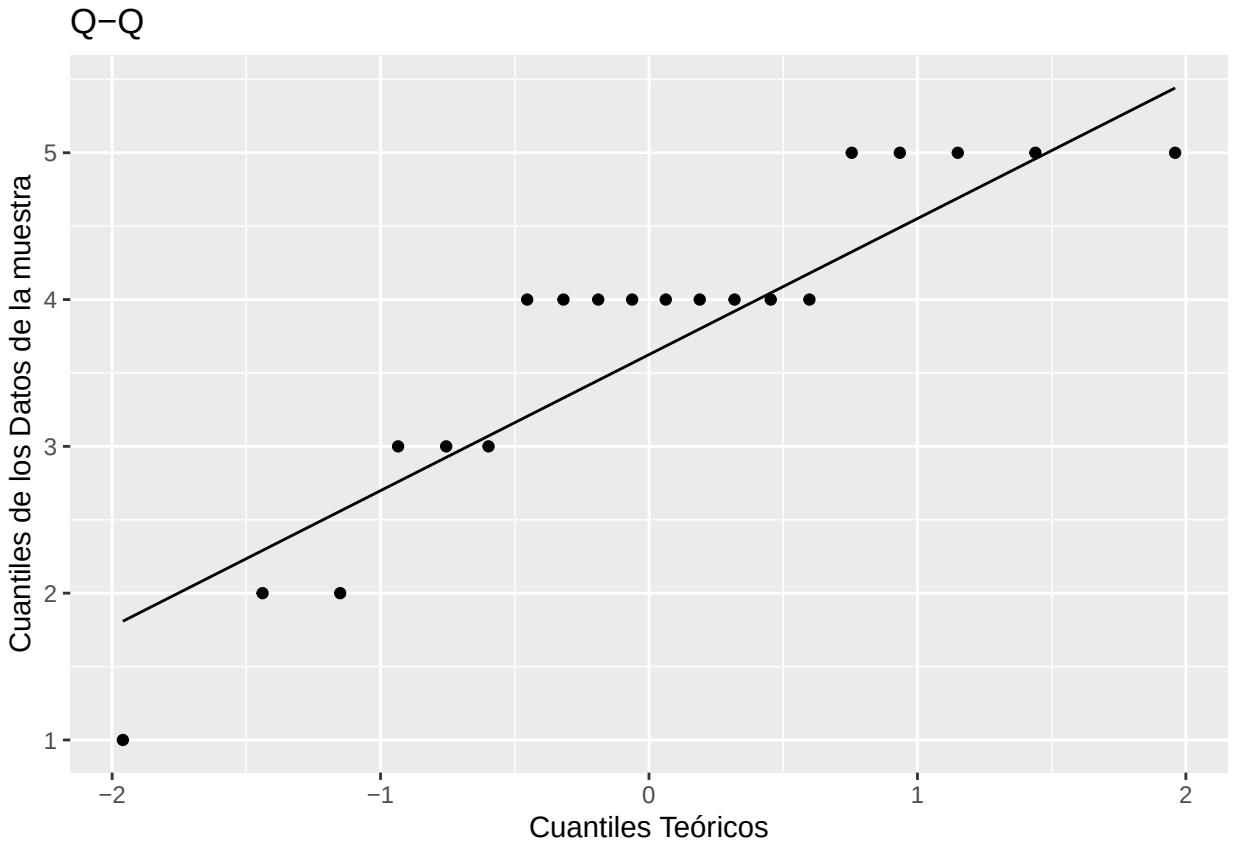
```r
ggqqplot(as.numeric(df.8$Severidad_Síntomas), main = "", xlab = "Cuantiles Teóricos", ylab = "Cuantiles
```



Alternativa:

```
qq1.2<-ggplot(df.8, aes(sample= as.numeric(Severidad_Síntomas))) +
  labs(title = "Q-Q",
       x = "Cuantiles Teóricos",
       y = "Cuantiles de los Datos de la muestra")+
  stat_qq() +
  stat_qq_line()

qq1.2
```



Usamos las funciones `ks.test` para Kolmogorov–Smirnov, y `shapiro.test` para la prueba de Shapiro–Wilk.

```
datos<-as.numeric(as.factor(df.8$Severidad_Síntomas))

ks.test(datos, "pnorm")
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  datos
## D = 0.92725, p-value = 2.317e-15
## alternative hypothesis: two-sided
```

```
shapiro.test(datos)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  datos
## W = 0.85762, p-value = 0.007172
```

En ambas pruebas $p<0.05$, por lo que se rechaza la normalidad de los datos.

Ejemplo 17: Prueba Chi-cuadrada

Usaremos el conjunto de datos Niveles de anemia en Nigeria, que se puede descargar desde Kaggle *(https://www.kaggle.com/datasets/adeolaadesina/factors-affecting-children-anemia-level/data)*. El conjunto de datos proviene de las Encuestas demográficas y de salud de Nigeria. Explora el impacto de la edad de las madres y los factores socioeconómicos en los niveles de anemia entre los niños de 0 a 59 meses.

1. Use la Prueba de Chi-cuadrada para evaluar la relación entre el nivel educativo de la madre y el nivel de anemia del hijo.

```
df.8.2<-read.csv("anemia.csv")
```

Primero creamos una tabla de contingencia, para observar como los valores de la variables se distribuyen en la diferentes categorías

```
selected_data <- df.8.2 %>%
  dplyr::select(Highest.educational.level, Anemia.level)
contingency_table <- table(selected_data$Highest.educational.level, selected_data$Anemia.level)

print(contingency_table)
```

```
##
##                      Mild Moderate Not anemic Severe
##   Higher         1497   296      238        591     14
##   No education 10185  1450     1769       1874    113
##   Primary        3079   608      645        900     42
##   Secondary      6027  1240     1322       1972     62
```

Ahora que tenemos la tabla de contingencia, aplicamos la prueba Chi-cuadrada:

```
chi_square_test <- chisq.test(contingency_table)

print(chi_square_test)
```

```
##
##  Pearson's Chi-squared test
##
## data:  contingency_table
## X-squared = 434.57, df = 12, p-value < 2.2e-16
```

Al realizar una prueba de chi-cuadrado, normalmente establecemos dos hipótesis:

- Hipótesis nula (H0): la hipótesis nula afirma que no existe asociación entre las dos variables categóricas que se están probando. Supone que cualquier diferencia observada en los datos se debe al azar en lugar de a una relación verdadera.
- Hipótesis alternativa (H1): la hipótesis alternativa establece una asociación significativa entre las dos variables. Sugiere que las diferencias observadas no se deben al azar y que existe una relación entre las variables.

En este ejemplo, Rechazamos la hipótesis nula porque el valor p es mucho menor que el nivel de significancia de 0.05 ($p< 0.05$). Esto demuestra que existe una asociación significativa entre el nivel de educación de la madre y el nivel de anemia del hijo. En otras palabras, los resultados de la prueba de chi-cuadrado indican que la probabilidad de que un niño tenga anemia está significativamente asociada con el nivel de educación de la madre.

Ejemplo 18: Prueba U de Mann Whitney

La prueba U de Mann-Whitney es una prueba no paramétrica popular para comparar resultados entre dos grupos independientes. La prueba U de Mann-Whitney, a veces llamada prueba de Mann-Whitney Wilcoxon o prueba de suma de rangos de Wilcoxon, se utiliza para comprobar si es probable que dos muestras deriven de la misma población (es decir, que las dos poblaciones tengan la misma forma). Algunos investigadores interpretan esta prueba como una comparación de las medianas entre las dos poblaciones. Recordemos que la prueba paramétrica compara las medias (H0: 1= 2) entre grupos independientes.

Se propone un nuevo enfoque de la atención prenatal para las mujeres embarazadas que viven en una comunidad rural. El nuevo programa incluye visitas domiciliarias durante el embarazo, además de las visitas

habituales o programadas regularmente. Se diseñó un ensayo piloto aleatorizado con 15 mujeres embarazadas para evaluar si las mujeres que participan en el programa dan a luz bebés más sanos que las mujeres que reciben la atención habitual. El resultado es el indicador de puntuación APGAR que se mide 5 minutos después del nacimiento. Recuerde que las puntuaciones APGAR varían de 0 a 10; las puntuaciones de 7 o más se consideran normales (saludables), de 4 a 6 bajas y de 0 a 3 críticamente bajas. ¿Existe evidencia estadística de una diferencia en los puntajes de APGAR en mujeres que reciben la atención prenatal nueva y mejorada en comparación con la atención prenatal habitual?

Datos: Los datos se incluyen en el archivo Apgar.xlsx:*https://github.com/ghebrer82/EstadisticaAplicada/blob/Ejemplos/Apgar.xlsx*

Realizamos la prueba U de Mann Whitney

Primero, leemos los datos:

```
df.8.3<-read_excel("Apgar.xlsx")
```

Las hipótesis son:

H0: Las dos poblaciones son iguales.

H1: Las dos poblaciones no son iguales.

Realizamos la prueba usamos la función `wilcox.test`

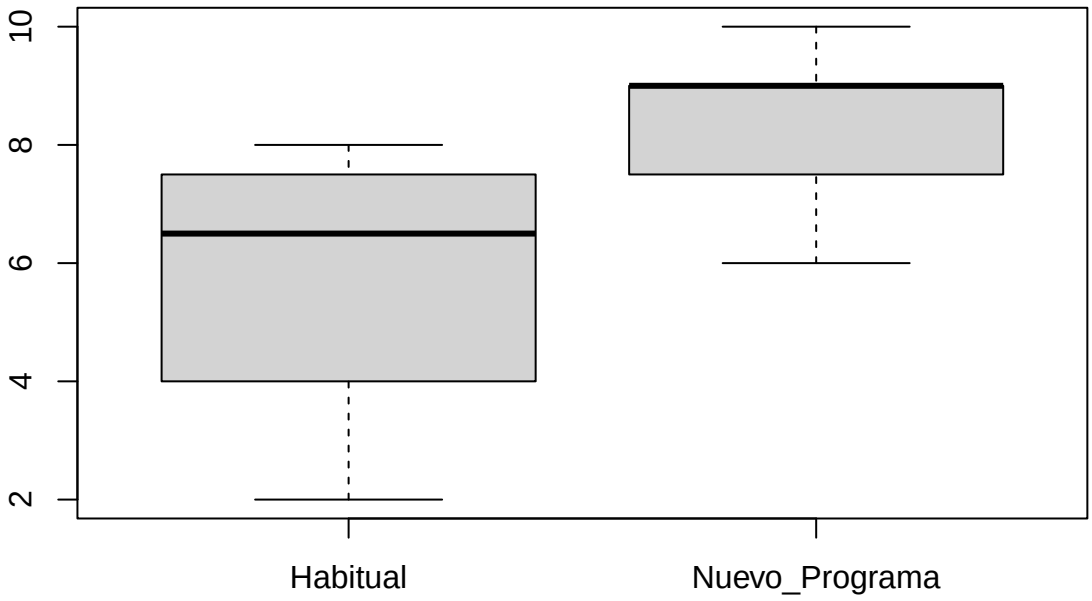
```
Mann.Whitney.Test <-wilcox.test(df.8.3$Habitual, df.8.3$Nuevo_Programa)
Mann.Whitney.Test
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  df.8.3$Habitual and df.8.3$Nuevo_Programa
## W = 8.5, p-value = 0.02609
## alternative hypothesis: true location shift is not equal to 0
```

Tenemos evidencia estadísticamente significativa ($p < 0.05$) para demostrar que las poblaciones de puntajes APGAR no son iguales en mujeres que reciben atención prenatal habitual en comparación con el nuevo programa de atención prenatal.

Visualizamos los datos de ambos programas:

```
boxplot(df.8.3)
```



Ejemplo 19: Prueba U de Mann Whitney

Se lleva a cabo un ensayo clínico para evaluar la eficacia de una nueva terapia antirretroviral para pacientes con influenza. Los pacientes son asignados aleatoriamente para recibir una terapia antirretroviral estándar (atención habitual) o la nueva terapia antirretroviral y son monitoreados durante 3 meses. El resultado primario es la carga viral, que representa el número de copias de virus por mililitro de sangre. Se asignan aleatoriamente 30 participantes y los datos se incluyen en el archivo retroviral.xlsx *https://github.com/gheb rer82/EstadisticaAplicada/blob/Ejemplos/retroviral.xlsx*

```
df.8.4<-read_excel("retroviral.xlsx")
```

Las hipótesis son:

H0: Las dos poblaciones son iguales. H1: Las dos poblaciones no son iguales.

Realizamos la prueba:

```
Mann.Whitney.Test.4 <-wilcox.test(df.8.4$Terapia_Estandar,
                                  df.8.4$Nueva_Terapia)
Mann.Whitney.Test.4
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  df.8.4$Terapia_Estandar and df.8.4$Nueva_Terapia
## W = 125, p-value = 0.6186
## alternative hypothesis: true location shift is not equal to 0
```

No rechazamos H0 porque $p > 0.05$. Por la tanto, No tenemos pruebas suficientes para concluir que los grupos de tratamiento difieren en la carga viral.

Ejemplo 20: Wilcoxon Signed Rank Test

Una prueba no paramétrica popular para datos emparejados o pareados es la prueba de rangos con signo de Wilcoxon.

Consideremos una investigación clínica para evaluar la eficacia de un nuevo fármaco diseñado para reducir las conductas repetitivas en niños afectados por autismo. Si el fármaco es eficaz, los niños mostrarán menos

conductas repetitivas durante el tratamiento en comparación con cuando no reciben tratamiento. Un total de 8 niños con autismo se inscriben en el estudio. El psicólogo del estudio observa a cada niño durante un período de 3 horas, tanto antes del tratamiento como después de tomar el nuevo fármaco durante 1 semana. Se mide el tiempo que cada niño presenta una conducta repetitiva durante cada período de observación de 3 horas. La conducta repetitiva se puntúa en una escala de 0 a 100, que representa el porcentaje del tiempo de observación en el que el niño presenta una conducta repetitiva. Por ejemplo, una puntuación de 0 indica que el niño no presentó una conducta repetitiva durante todo el período de observación, mientras que una puntuación de 100 indica que el niño presentó una conducta repetitiva constantemente. Los datos se muestran en el archivo aut.xlsx *https://github.com/ghebrer82/EstadisticaAplicada/blob/Ejemplos/aut.xlsx*

```
df.8.5<-read_excel("aut.xlsx")
```

Las hipótesis se dan a continuación.

H0: La diferencia de medianas es cero H1: La diferencia de medianas es positiva

Realizamos la prueba usamos la función `wilcox.test` con el argumento `paired=TRUE`, dado que son 2 muestras dependientes y `alternative="greater"`, dado que la H1, menciona que la diferencia es mayor que cero.

```
Wilcoxon.Test.5 <-wilcox.test(df.8.5$Antes_Tr,df.8.5$Despues_Tr,
                              paired=TRUE, alternative="greater")
Wilcoxon.Test.5
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  df.8.5$Antes_Tr and df.8.5$Despues_Tr
## V = 32, p-value = 0.02874
## alternative hypothesis: true location shift is greater than 0
```

Debido a que el valor $p < 0.05$, tenemos evidencia estadísticamente significativa de que las conductas repetitivas mejoran después de tomar el medicamento en comparación con el tratamiento anterior.

Ejemplo 21: Prueba H de Kruskal-Wallis

Una prueba no paramétrica popular para comparar resultados entre más de dos grupos independientes es la prueba de Kruskal-Wallis. La prueba de Kruskal-Wallis compara las medianas entre k grupos de comparación ($k > 2$) y a veces se describe como un ANOVA con los datos reemplazados por sus rangos. Las hipótesis nulas y de investigación para la prueba no paramétrica de Kruskal Wallis se establecen de la siguiente manera:

H0: Las medianas de la población k son iguales

H1: Las medianas de la población k no son todas iguales

Un entrenador personal está interesado en comparar los umbrales anaeróbicos de los atletas de élite. El umbral anaeróbico se define como el punto en el que los músculos no pueden obtener más oxígeno para mantener la actividad o el límite superior del ejercicio aeróbico. También está relacionado con la frecuencia cardíaca máxima. Los siguientes datos son umbrales anaeróbicos para corredores de fondo, ciclistas de fondo, nadadores de fondo y esquiadores de fondo.

¿Existe una diferencia en los umbrales anaeróbicos entre los diferentes grupos de atletas de élite?

Datos: *https://github.com/ghebrer82/EstadisticaAplicada/blob/Ejemplos/umbral.xlsx*

Leemos los datos incluido en el archivo `umbral.xlsx`

```
df.8.6<-read_excel("umbral.xlsx")
```

Antes de realizar la prueba, debemos modificar el formato del conjunto de datos:

```r
# Convert the data frame to a long format
df.8.6a <- stack(df.8.6)
```

Luego, hacemos la prueba usando la función `kruskal.test`:

```r
kruskal_test <- kruskal.test(values ~ ind, data = df.8.6a)
print(kruskal_test)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  values by ind
## Kruskal-Wallis chi-squared = 9.1143, df = 3, p-value = 0.02781
```

Contamos con evidencia estadísticamente significativa ($p < 0.05$) de que existe una diferencia en los umbrales anaeróbicos entre los cuatro grupos diferentes de atletas de élite.

Observe que en este ejemplo, los umbrales anaeróbicos de los corredores de fondo, ciclistas y esquiadores de fondo son comparables (solo con observar los datos). Los nadadores de fondo parecen ser los atletas que difieren de los demás en términos de umbrales anaeróbicos. Recuerde que, de manera similar a las pruebas de ANOVA, rechazamos la hipótesis nula a favor de la hipótesis alternativa si dos de las medianas no son iguales.

Ejercicios

1. Una clínica de salud desea evaluar la distribución de los tiempos de espera de sus pacientes en la sala de espera. Se cree que los tiempos de espera podrían no seguir una distribución normal debido a factores como picos de afluencia o demoras inesperadas. Para investigar esto, se han registrado los tiempos de espera (en minutos) de una muestra de 50 pacientes.

Datos:Los datos se encuentran en el archivo 08_tiempos_nopar.xlsx *(https://github.com/ghebrer82/EstadisticaAplicada/blob/main/08_tiempos_nopar.xlsx)*

Responder lo siguiente:

a) Evaluar la normalidad de los datos usando un Histograma.
b) Evaluar la normalidad de los datos usando un Gráfico Q-Q.
c) Evaluar la normalidad de los datos usando la Prueba de Shapiro-Wilk. d) Evaluar la normalidad de los datos usando la Prueba de Kolmogorov-Smirnov.

2. La dirección de la clínica, tras observar que los tiempos de espera no siguen una distribución normal, quiere comparar si los tiempos de espera promedio (o más apropiadamente, las medianas) son diferentes entre el turno de la mañana y el turno de la tarde. Para esto, se ha tomado una muestra de 50 pacientes, donde 25 corresponden al turno de la mañana y 25 al turno de la tarde.

Datos:Los datos se encuentran en el archivo 08b_tiempos_2.xlsx *(https://github.com/ghebrer82/EstadisticaAplicada/blob/main/08b_tiempos_2.xlsx)*

Realizar lo siguiente: a) Visualizar las distribuciones de los tiempos de espera para cada turno usando diagramas de cajas. b) Realizar la Prueba de Mann-Whitney U (también conocida como Prueba de Wilcoxon Rank-Sum), una prueba no paramétrica equivalente a la prueba t de Student para muestras independientes, para determinar si hay diferencias significativas en las medianas de los tiempos de espera entre el turno de la mañana y el turno de la tarde.

Bibliografía

- Estadística, Mario Triola, 12va Edición, Pearson, 2018.
- Probabilidad y Estadística para Ingeniería y Ciencias, Ronald Walpole, Pearson Educación, 2012.