Estadística Descriptiva

Edgar Ramírez Galeano

Proyectos en RStudio

- Un proyecto es una carpeta que contiene todos los scripts, archivos .RData y .Rhistory
- Permite tener nuestros análisis ordenados.
- Al abrir un proyecto RStudio con las pestañas que se tenían activas
- Se sugiere tener una estructura interior, por ejemplo: Scripts, Data, Graficas etc. . .

- La **Estadística descriptiva** registra los datos en tablas y los representa en gráficos. Calcula los parámetros estadísticos (medidas de centralización, medidas de posición y de dispersión), que describen el conjunto estudiado.
- La **distribución de frecuencias** o **tabla de frecuencias** es una ordenación en forma de tabla de los datos estadísticos, asignando a cada dato su frecuencia correspondiente.
- La **frecuencia absoluta** es el número de veces que aparece un determinado valor en un estudio estadístico.
- La **suma de las frecuencias absolutas** es igual al número total de datos, que se representa por N.
- La **frecuencia relativa** es el cociente entre la frecuencia absoluta de un determinado valor y el número total de datos. Se puede expresar en tantos por ciento.

Datos no agrupados

En el diseño de un plato dosificador para siembra de maíz con una sembradora de precisión, se observó la longitud, ancho y grosor de 27 granos de maíz oloton y se obtuvieron los siguientes datos (en mm):

Longitud	Ancho	Grosor	Longitud	Ancho	Grosor	Longitud	Ancho	Grosor
11	9	4	10	11	6	11	11	6
11	11	4	11	11	4	12	12	4
11	9	5	10	11	3	10	11	4
11	10	3	11	11	5	9	10	5
12	10	5	10	10	5	11	10	6
10	12	6	10	9	11	9	10	5
11	11	3	10	11	4	11	10	4
11	10	4	12	10	3	12	11	3
8	9	3	10	11	5	11	10	4

Tablas de frecuencia absoluta

La función ***table*** sirve para construir tablas de frecuencia de una vía, a continuación la estructura de la función.

Código en R

table(..., exclude, useNA)

Los parámetros de la función son:

- **...** espacio para ubicar los nombres de los objetos (variables o vectores) para los cuales se quiere construir la tabla.
- ***exclude***: vector con los niveles a remover de la tabla.
- ***useNA***: instrucción de lo que se desea con los NA. Hay tres posibles valores para este parámetro: **'no'** si no se desean usar, **'ifany'** y **'always'** si se desean incluir.

Ejemplo tablas de frecuencia absoluta

```
> table(MaizOloton$Longitud)
8 9 10 11 12
1 2 8 12 4
 > table(MaizOloton$Longitud,exclude =c(8,12))
9 10 11
2 8 12
 > table(MaizOloton$Ancho)
9 10 11 12
4 10 11 2

addmargins(table(MaizOloton$Longitud))
```

Tablas de Frecuencia Relativa

La función ***prop.table*** se utiliza para crear tablas de frecuencia relativa a partir de tablas de frecuencia absoluta, la estructura de la función se muestra a continuación.

Código en R

prop.table(x, margin=NULL)

- ***x:*** tabla de frecuencia.
- ***margin:*** valor de 1 si se desean proporciones por filas, 2 si se desean por columnas, NULL si se desean frecuencias globales.

Ejemplo de Tablas de Frecuencia Relativa

```
> tf<-table(MaizOloton$Longitud)
 > prop.table(tf)

8 9 10 11 12
0.03703704 0.07407407 0.29629630 0.44444444 0.14814815

> prop.table(table(MaizOloton$Grosor))

3 4 5 6 11
0.22222222 0.33333333 0.25925926 0.14814815 0.03703704
```

Frecuecias Acumuladas

Frecuencia Absoluta Acumulada

```
> cumsum(tf)
8 9 10 11 12
1 3 11 23 27
```

Frecuencia Relativa Acumulada

```
> cumsum(prop.table(table(MaizOloton$Longitud)))
8 9 10 11 12
0.03703704 0.11111111 0.40740741 0.85185185 1.00000000
```

Tabla de Frecuencia de Dos Vías

Se miden dos o más características (variables) en cada individuo

Código en R

```
> table(MaizOloton$Longitud,MaizOloton$Ancho)
9 10 11 12
8 1 0 0 0
9 0 2 0 0
10 1 1 5 1
11 2 5 5 0
12 0 2 1 1
```

Reto Tablas de Frecuecia Datos No Agrupados

Utiliza el Dataframe “mtcars” el cual contiene los datos que se extrajeron de la revista Motor Trend de EE. UU. De 1974, y comprenden el consumo de combustible y 10 aspectos del diseño y rendimiento del automóvil para 32 automóviles (modelos 1973–74).

Para cargar los datos a R utiliza el comando: data("mtcars")

Revise la ayuda del dataframe mtcars el cual describe el significado de cada una de las variables de la tabla.(?mtcars)

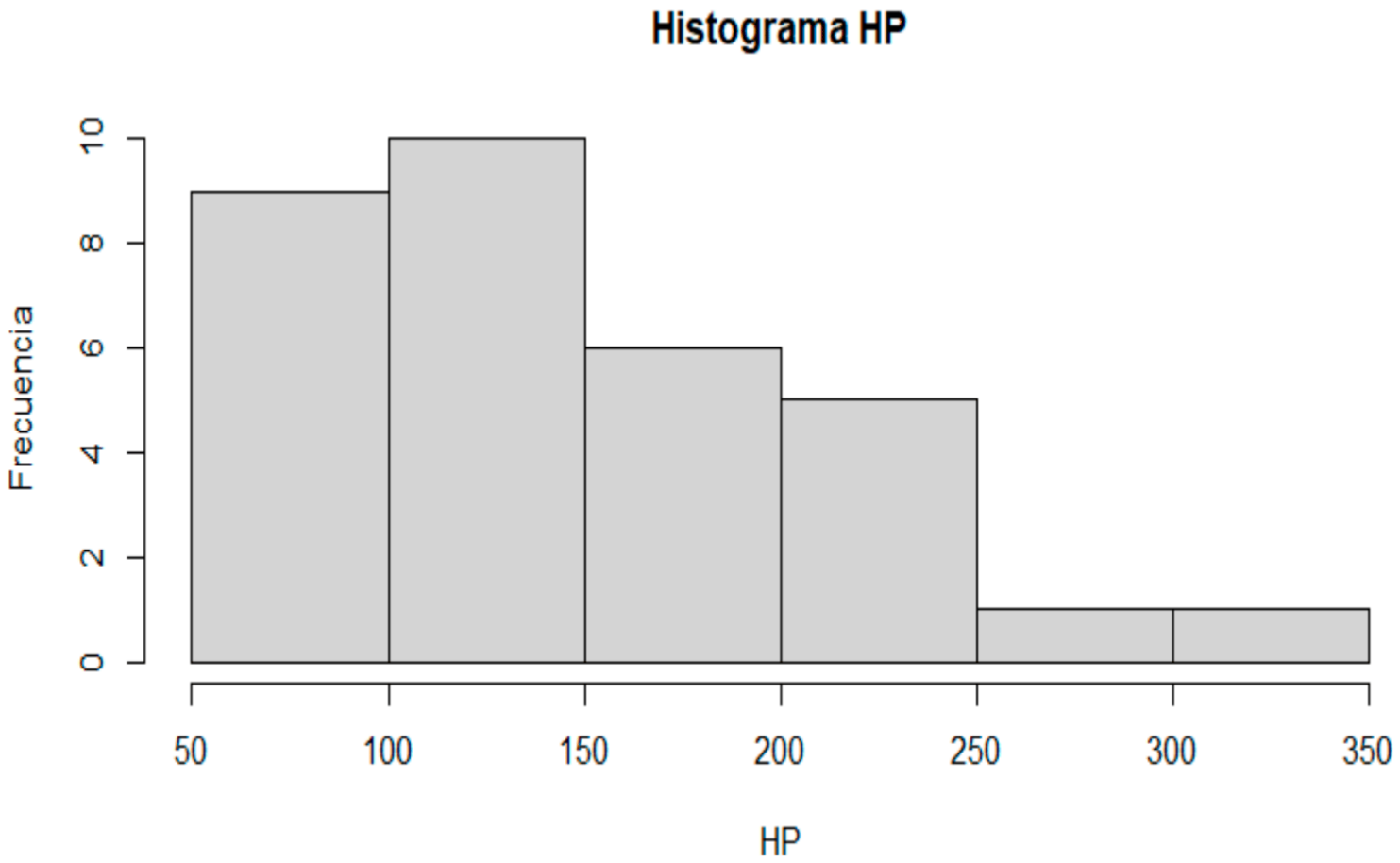
1. Responde las siguientes preguntas:
   - ¿Frecuencia Absoluta de autos con 8 cilindros(cyl)?.
   - ¿Frecuencia Relativa de autos con 3 carburadores(carb)?
   - ¿Cantidad de autos con transmisión automática(am)?
   - ¿Pocentaje de autos con transmisión manual(am)?
   - ¿Cantidad de autos con 4 o menos carburadores(carb)?
2. Construya una tabla de dos vías para obtener la frecuencia absolutas y relativas utilizando las variables transmisión(am) y número de cilindros(cyl). Explique los resultados obtenidos. Transmission (0 = automatic, 1 = manual)

Histogramas

Un histograma es una gráfica que nos permite observar la distribución de datos numéricos usando barras. Cada barra representa el número de veces (frecuencia) que se observaron datos en un rango determinado.

Código en R

```
hist(mtcars$hp)
hist(mtcars$hp,main="Histograma HP",xlab = "HP",ylab =
"Frecuencia")
```


Histograma HP
Frecuencia
0
2
4
6
8
10
50
100
150
200
250
300
350
HP

Datos Agrupados

Máquina seleccionadora de duraznos para almíbar bajo la norma NMX-F-034-1982

Perímetro duraznos

15.2	15.3	15.4	15.5	15.5	15.5	15.6	15.6	15.6	15.7
15.8	15.8	15.9	16.1	16.1	16.1	16.2	16.3	16.3	16.4
16.4	16.4	16.5	16.5	16.6	16.6	16.6	16.8	16.9	17.2
17.3									

Cómo agrupar datos

Proceso de agrupación

1. Se decide el número de intervalos que se van a usar.
   - 3
2. Se calcula su amplitud
   - Máximo: 17.3 Mínimo: 15.2 Recorrido:25816
   - Amplitud: 2.1
3. Se calculan los extremos de los intervalos.
   - 15.2 - 15.9
   - 15.9 - 16.6
   - 16.6 - 17.3
4. Se calcula un valor representativo de cada intervalo. (Punto medio o marca de clase)
   - 15.55
   - 16.25
   - 16.95

Proceso de agrupación en R

Amplitud

```
> rango = max(Durazno$Perimetro) - min(Durazno$Perimetro)
> num_clase = 3
> amplitud= rango / num_clase
[1] 0.7
```

Extremos de los intervalos

```
minimo= min(Durazno$Perimetro)
limites = minimo + amplitud *(0:num_clase)
> limites
[1] 15.2 15.9 16.6 17.3
```

Codificación de datos en R

Al agrupar un conjunto de datos con R, lo que hacemos es codificarlos, convirtiendo la variable cuantitativa en un factor cuyos niveles son las clases en las que hemos agrupado los valores y asignando cada dato a su clase.
La función básica de R para agrupar un vector de datos numéricos y codificar sus valores con las clases a las que pertenecen es

Función

cut(x, breaks=..., labels=..., right=FALSE, include.lowest=TRUE)

- x es el vector numérico que contiene los datos
- El parámetro breaks puede ser un vector numérico formado por los extremos de los intervalos en los que queremos agrupar los datos y que habremos calculado previamente.

- El parámetro right es un parámetro lógico que permite indicar qué tipo de intervalos queremos. Si usamos intervalos cerrados por la izquierda y abiertos por la derecha.

Codificación tractores

```
intervalos = cut(Durazno$Perimetro,breaks = limites,
include.lowest = TRUE)
```

Tabla de Frecuencias Datos Agrupados

Una vez agrupados los datos y codificados con las etiquetas de las clases, ya podemos calcular las tablas de frecuencias absolutas y relativas de los datos agrupados. Una posibilidad es usar las funciones table y prop.table al como lo hacíamos con los datos no agrupados.

Código en R

Frecuencia Absoluta

```
> frecuencia_absoluta<-table(intervalos)
```

Frecuencia Relativa

```
> frecuencia_relativa<-prop.table(frecuencia_absoluta)
```

Frecuencia Absoluta Acumulada

```
> frecuencia_abs_acumulada<-cumsum(table(intervalos))
```

Frecuencia Relativa Acumulada

```
> frecuencia_rel_acumulada<-cumsum(frecuencia_relativa)
```

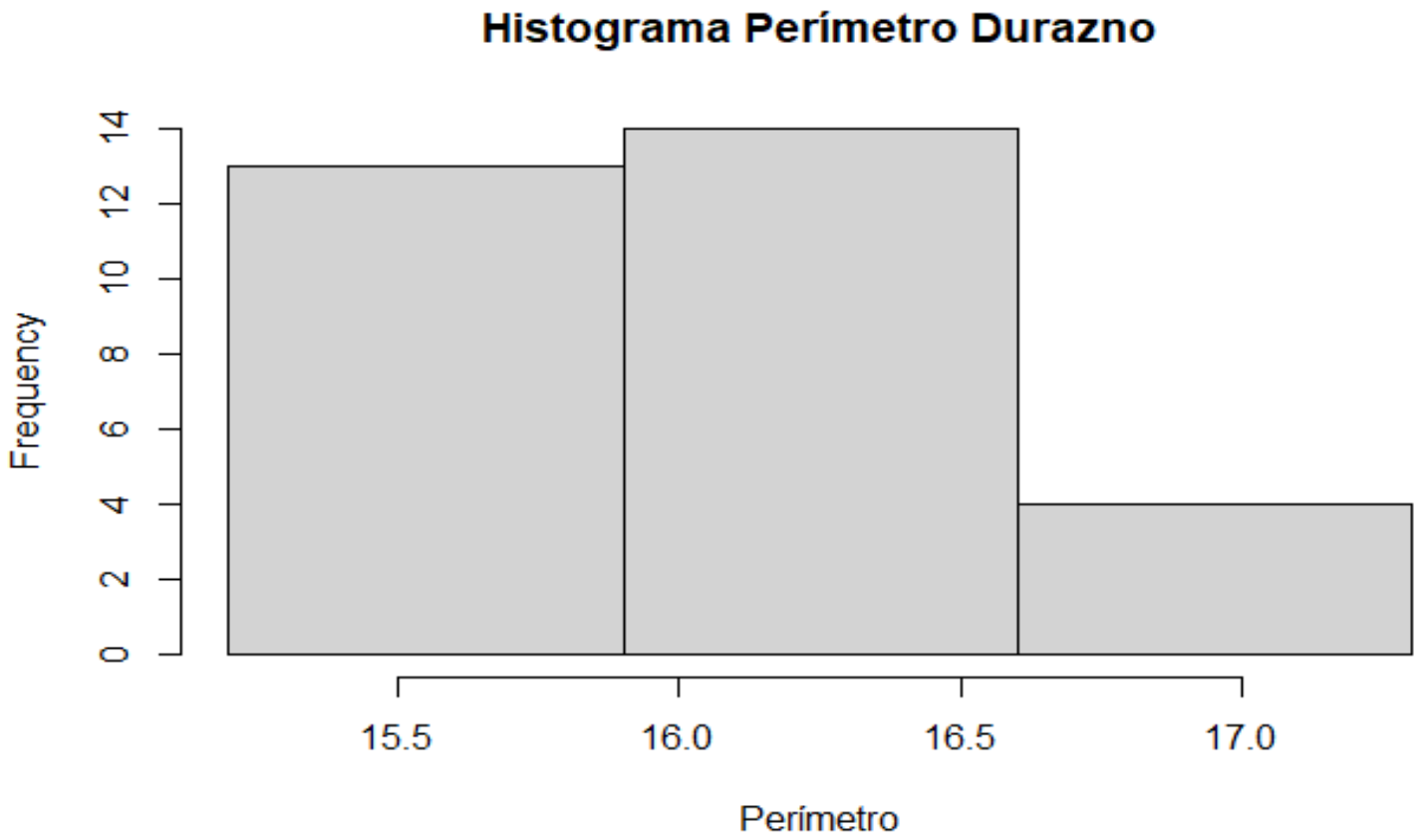
Data Frame que contenga la Tabla de Frecuencias

Código en R

```
tabla_de_frecuencias<-data.frame(Intervalos=levels(intervalos),
f=as.numeric(frecuencia_absoluta),
fa=as.numeric(frecuencia_abs_acumulada),
fr=as.numeric(frecuencia_relativa),
fra=as.numeric(frecuencia_rel_acumulada)
)
```

intervalos f fa fr fra
1 [15.2,15.9] 13 13 0.419 0.419
2 (15.9,16.6] 14 27 0.451 0.870
3 (16.6,17.3] 4 31 0.129 1.00

```
> hist(Durazno$Perimetro,breaks = limites,include.lowest =
TRUE,main = "Histograma Perímetro Durazno",xlab =
"Perímetro")
```

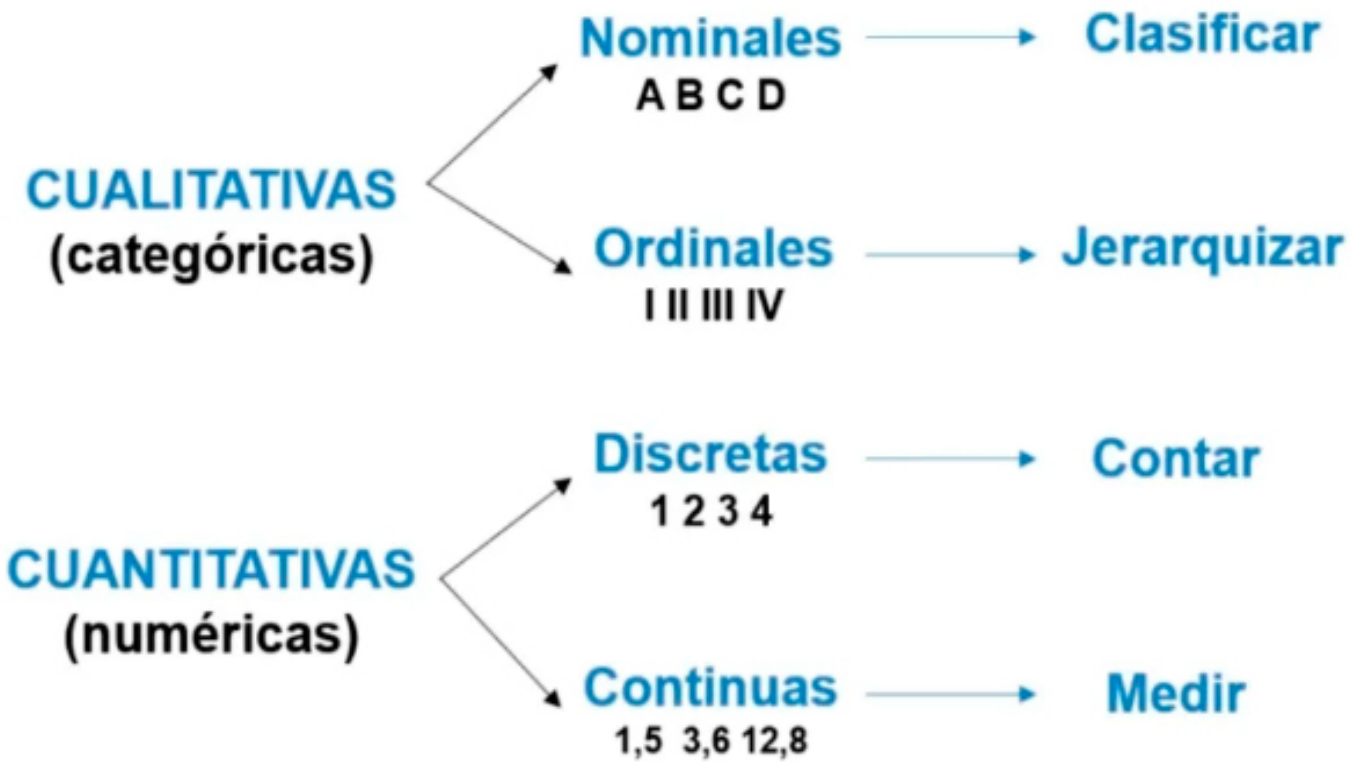


Reto Tablas de Frecuencias Datos Agrupados

Realizar una tabla de frecuencias e histograma con cinco intervalos del peso del durazno para el diseño de una máquina seleccionadora requerida en la agroindustria.

Peso	Peso	Peso	Peso
58.3	64.9	70.7	64.4
59.3	65.1	71.4	64.9
60.5	66.1	71.5	70.1
61	66.5	71.8	70.5
61.4	66.6	72.5	
62	66.8	72.6	
63.4	67.5	74	
63.5	67.5	75.8	
64.2	69.6	76.8	

Tipos de variables



Medidas de Tendencia Central

Las medidas de Tendencia Central ubican el valor alrededor del cual se concentra un conjunto de datos.

- **Media aritmética:** Es el valor promedio de la distribución.
- **Mediana:** Es la puntación de la escala que separa la mitad superior de la distribución y la inferior, es decir divide la serie de datos en dos partes iguales.
- **Moda:** Es el valor que más se repite en una distribución.

Media aritmética:

```
> mean(MaizOloton$Longitud)
[1] 10.59259
```

Mediana:

```
> median(MaizOloton$Longitud)
[1] 11
```

Moda:

```
> tf<-table(MaizOloton$Longitud)
> names(sort(tf,decreasing = T)[1])
[1] 11
```

Medidas de dispersión

Las medidas de dispersión nos informan sobre cuánto se alejan del centro los valores de la distribución.

- **Rango o recorrido:** Es la diferencia entre el mayor y el menor de los datos de una distribución estadística.
- **Varianza:** Es la media aritmética del cuadrado de las desviaciones respecto a la media.
- **Desviación típica :** Es la raíz cuadrada de la varianza.
- **Desviación Media o Desviación Promedio:** Es el promedio de variación de cada observación en valores absolutos con respecto a la media.

Rango o recorrido:

```
diff(range(MaizOloton$Longitud))
[1] 8 12
> max(MaizOloton$Longitud) - min(MaizOloton$Longitud)
[1] 4
```

Varianza:

```
> var(MaizOloton$Longitud)
[1] 0.9430199
```

Desviación típica :

```
> sd(MaizOloton$Longitud)
[1] 0.9710921
```

Coeficiente de variación

```
sd(MaizOloton$Longitud)/mean(MaizOloton$Longitud)
0.09167653
```

Desviación Media

```
mean(abs(MaizOloton$Longitud-mean(MaizOloton$Longitud)))
```

Medidas de posición

Las medidas de posición dividen un conjunto de datos en grupos con el mismo número de individuos.

Para calcular las medidas de posición es necesario que los datos estén ordenados de menor a mayor.

- **Cuartiles :** Dividen la serie de datos en cuatro partes iguales.
- **Deciles:** Dividen la serie de datos en diez partes iguales.
- **Percentiles:** Dividen la serie de datos en cien partes iguales.

Cuartiles

```
> summary(mtcars$hp)
Min. 1st Qu. Median Mean 3rd Qu. Max.
52.0 96.5 123.0 146.7 180.0 335.0

> quantile(mtcars$hp)
0% 25% 50% 75% 100%
52.0 96.5 123.0 180.0 335.0
```

Deciles:

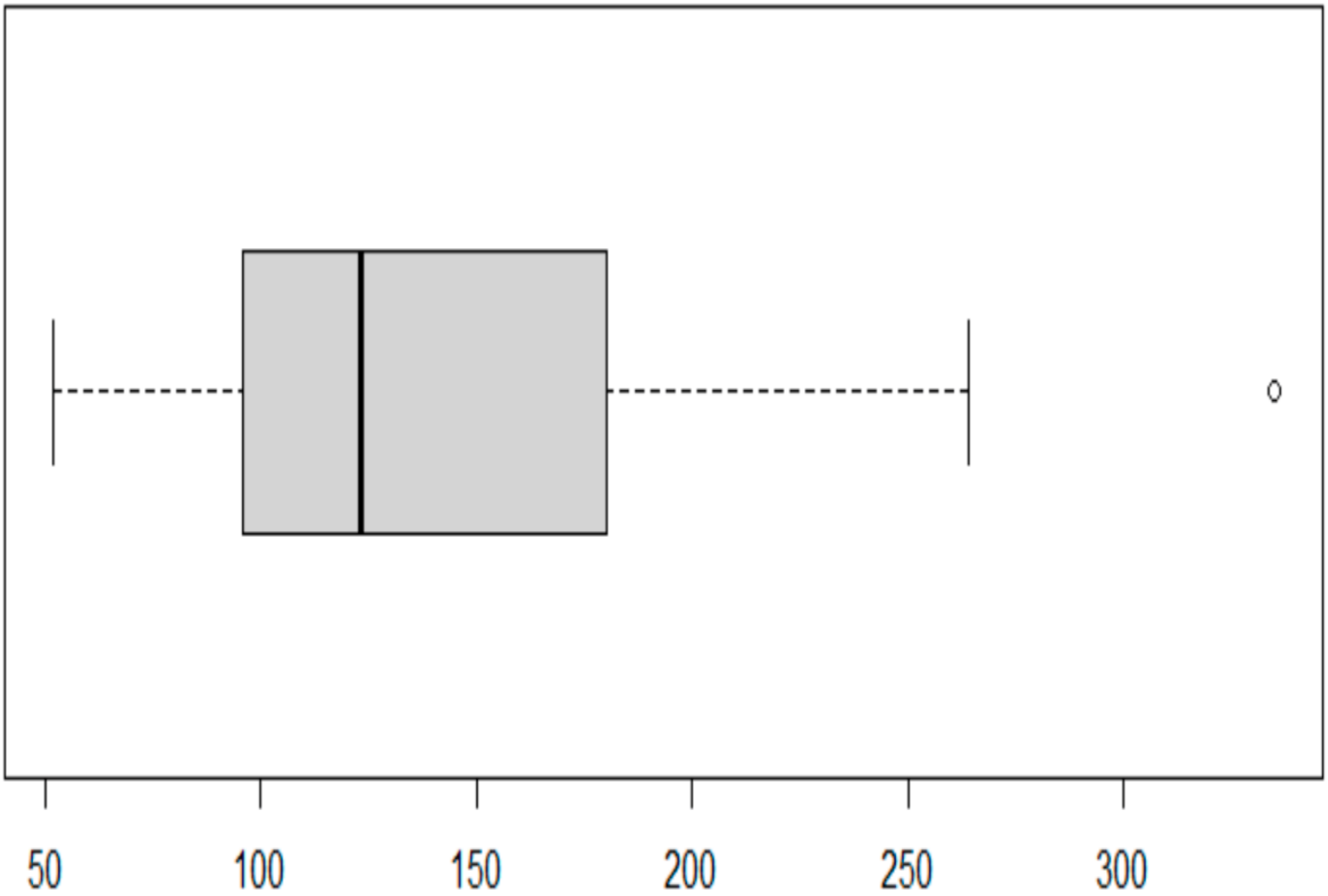
```
> quantile(mtcars$hp, prob = c(0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7,
0.8, 0.9))
10% 20% 30% 40% 50% 60% 70% 80% 90%
66.0 93.4 106.2 110.0 123.0 165.0 178.5 200.0 243.5
```

Diagrama de cajas

Lo que aparece es un diagrama de cajas en el que "la caja" es "lo que parece una caja" (el cuadrilátero :)) y los bigotes son las líneas (punteadas) con unos segmentos finales (superiores e inferiores). También se pueden observar unos puntos por encima o por debajo de los límites de los bigotes, estos se conocen como valores extremos (valores atípicos) u outliers (en inglés).

Código en R

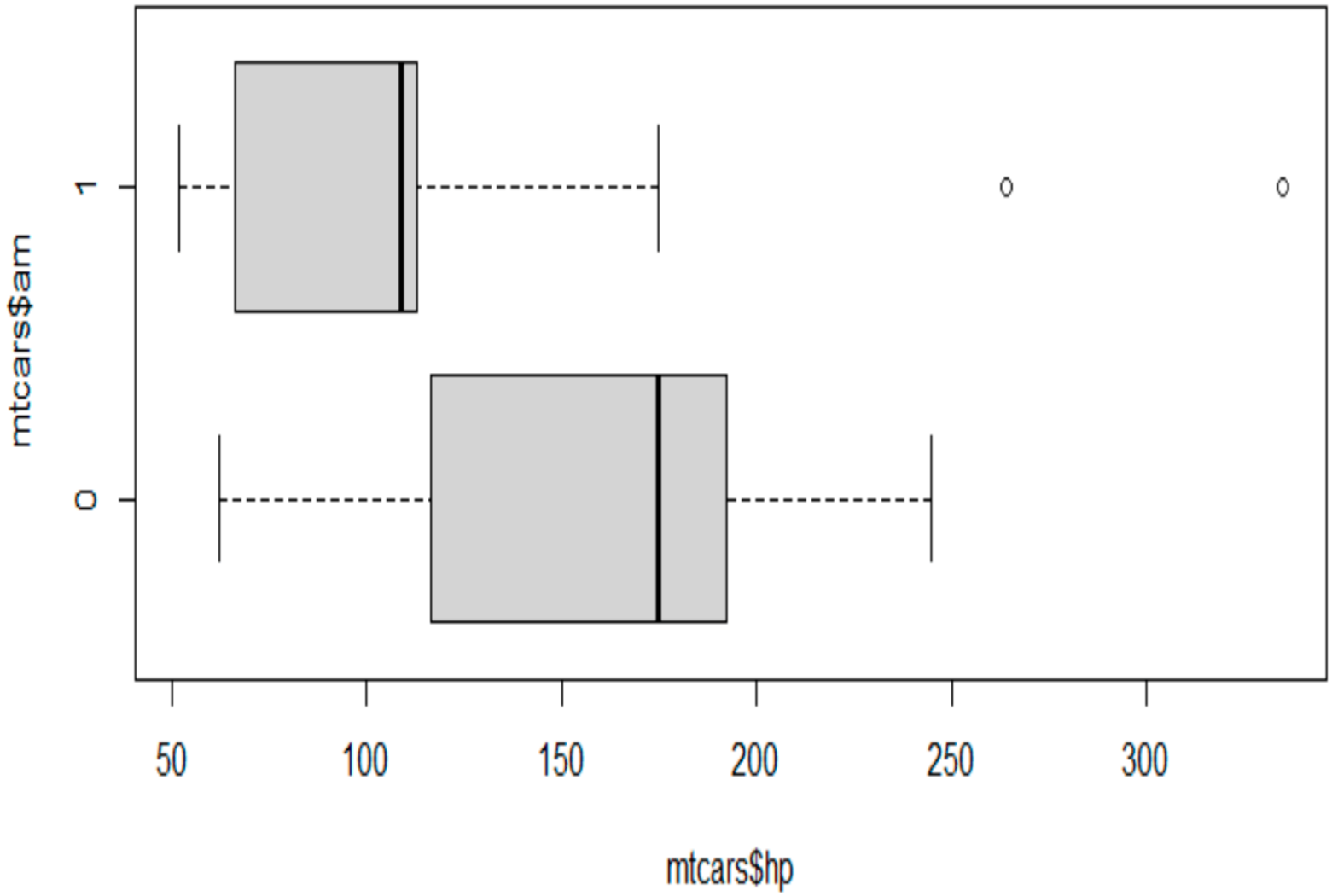
```
boxplot(mtcars$hp)
boxplot(mtcars$hp,horizontal = TRUE)
```


0
50
100
150
200
250
300

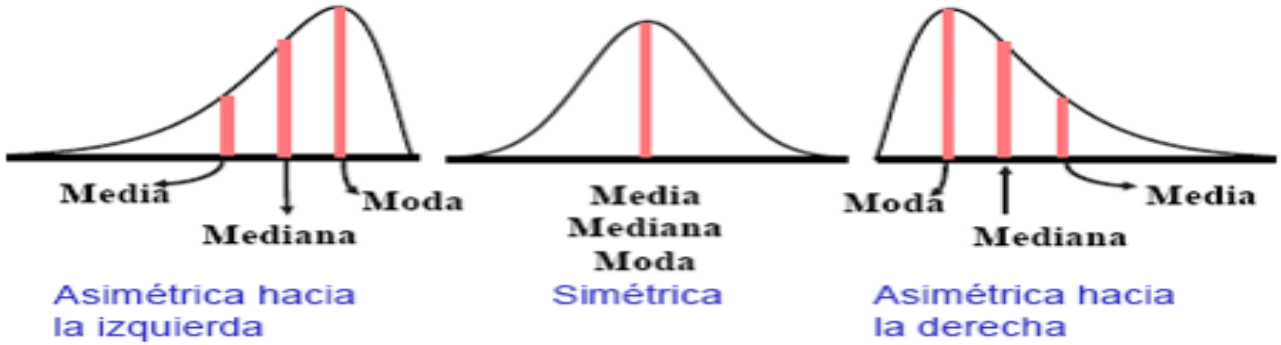
La línea gruesa es la mediana.
La parte inferior de la caja es el primer cuartil (1st Qu.), esto es, el valor que separa al primer 25% de los datos (luego de ordenarlos de menor a mayor).
La mediana (corresponde al segundo cuartil) permite dividir al conjunto de datos en dos partes de modo tal que cada parte tendrá el 50% de los datos (luego de ordenar todo el conjunto de datos de menor a mayor).
El tercer cuartil (3st Qu.) separa al 75% de los datos una vez ordenados de menor a mayor.
Ahora, la distancia que hay entre el primer y el tercer cuartil se conoce como rango intercuartílico (IQR) y permite tener una idea de la dispersión (acumulación) de los valores alrededor de la mediana

Código en R

boxplot(mtcars$hp $\sim$ $mtcars$am,horizontal = TRUE)



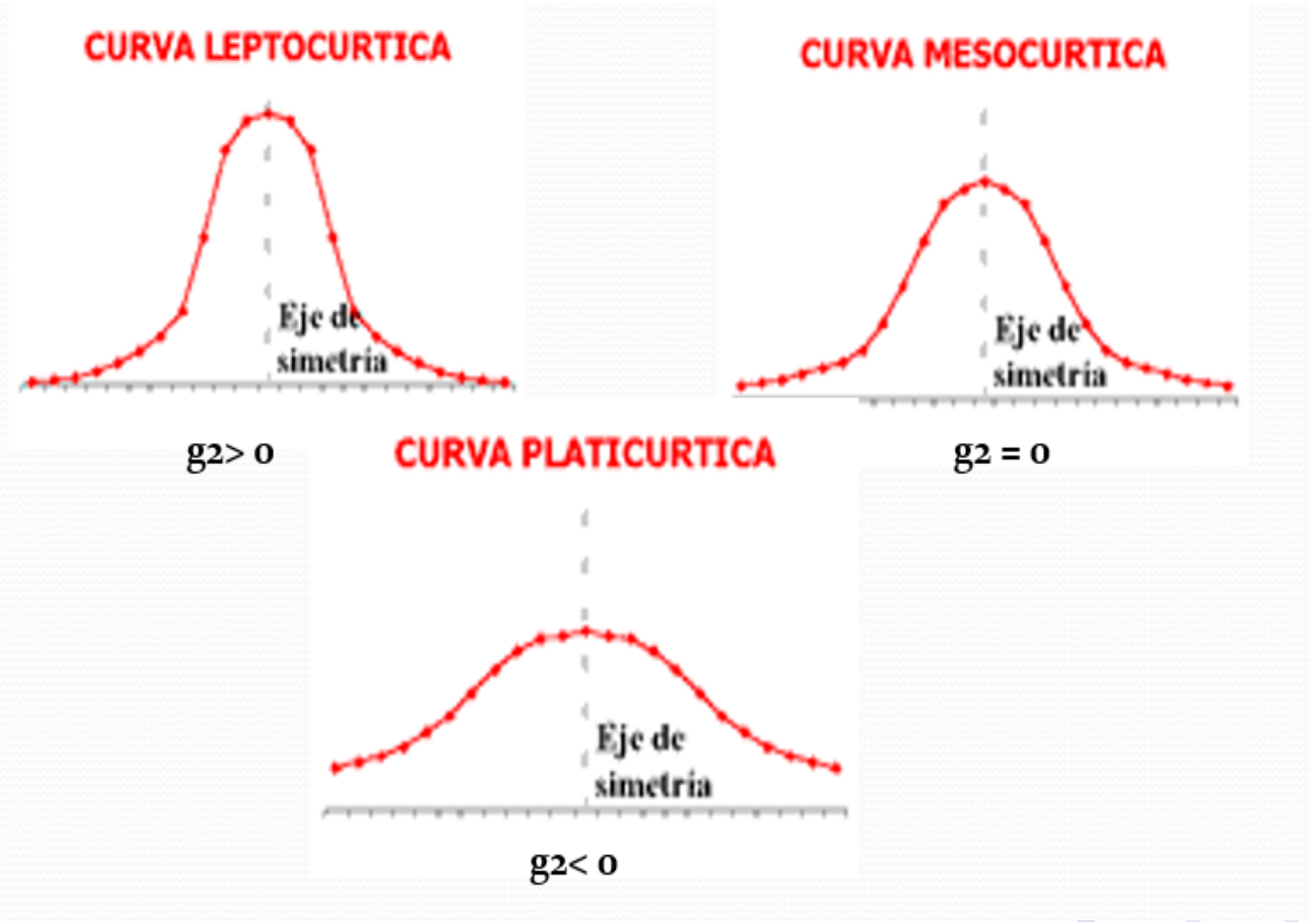
Asimetría



- $(g1 = 0)$: Se acepta que la distribución es Simétrica, es decir, existe aproximadamente la misma cantidad de valores a los dos lados de la media.
- $(g1 > 0)$: La curva es asimétricamente positiva por lo que los valores se tienden a reunir más en la parte izquierda que en la derecha de la media.
- $(g1 < 0)$: La curva es asimétricamente negativa por lo que los valores se tienden a reunir más en la parte derecha de la media.

CURTOSIS

Esta medida determina el grado de concentración que presentan los valores en la región central de la distribución.



Funciones en R de Asimetría y Curtosis

Para obtener estas dos medidas se utiliza el paquete **moments**
Funcion que instala paquete: install.packages("moments")
library(moments)

Asimetría

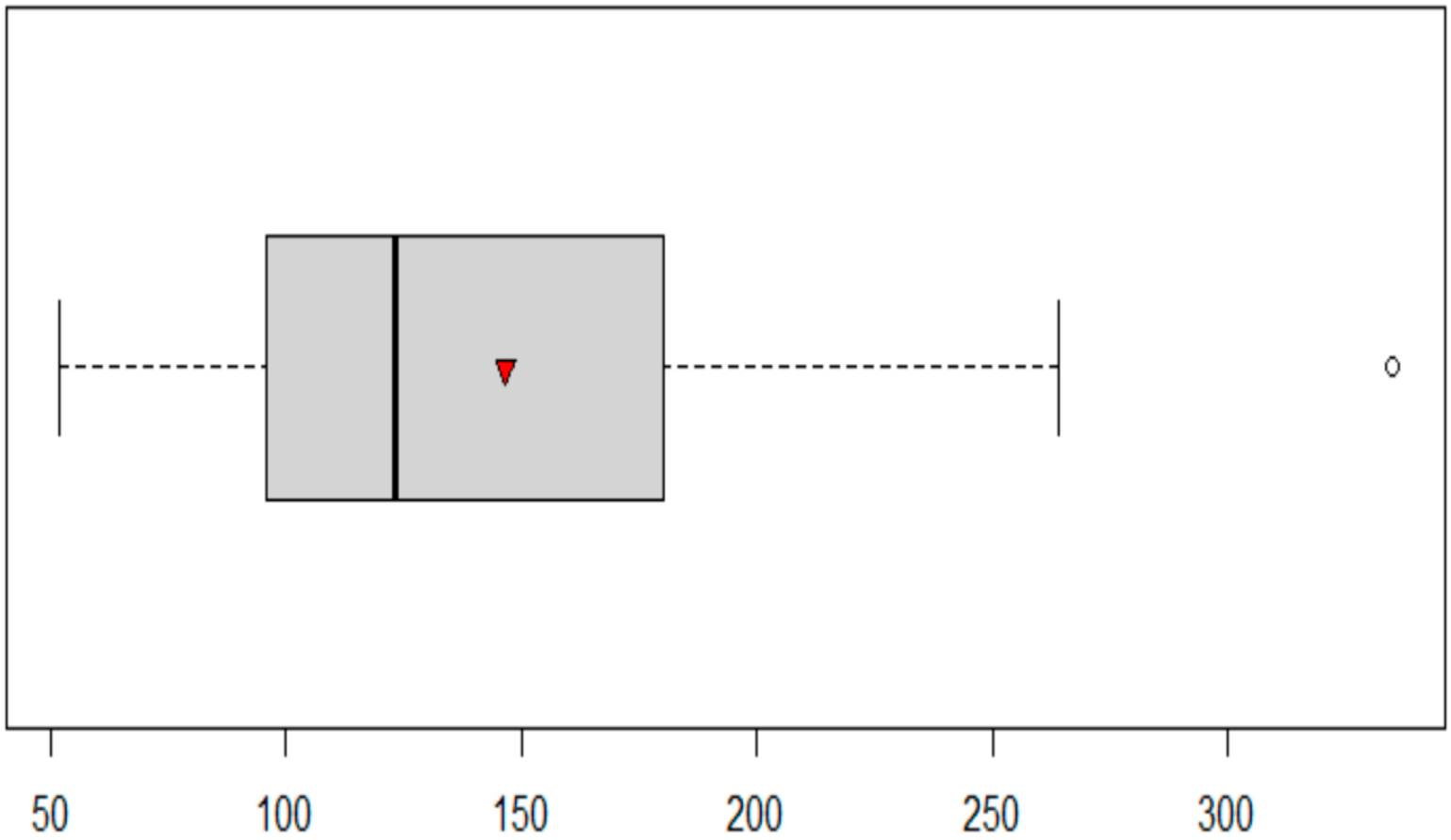
```
> skewness(mtcars$hp)
[1] 0.7614356
```

Curtosis

```
> kurtosis(mtcars$hp)
[1] 3.052233
```

Código en R

```
boxplot(mtcars$hp,horizontal = TRUE)
points(mean(mtcars$hp), 1, pch=25, bg="red")
```



Función tapply

Realiza una operación (parámetro 3) respecto a un vector (parámetro 1) agrupada por los factores que se indiquen como argumento (parámetro 2).

tapply(parámetro 1, parámetro 2, parámetro 3)

Código en R

```
> tapply(mtcars$hp, mtcars$am, mean)
0 1
160.2632 126.8462
```

En este ejemplo se calcula el promedio(mean) de potencia de los autos agrupados por el tipo de transmisión(0=automatica,1=manual)

Reto Parámetros Estadísticos

Utiliza el archivo "cosecha_jal.xlsx" el cual contiene la superficie cosechada de 7 diferentes cultivos en un periodo comprendido de 2004 a 2009 en el estado Jalisco.
Esta información fue obtenida del Banco de información sociodemográfica y económica (INEGI).

- Calcular el promedio, mediana, varianza y desviación estándar de la superficie cosecha en el estado de Jalisco.
- Calcular los cuartiles de la superficie cosechada en el estado de Jalisco y construye su diagrama de cajas.
- Calcular el promedio y desviación estándar de la superficie cosechada para cada uno de los cultivos.

- Calcular el coeficiente de asimetría y curtosis para la superficie cosechada del estado de Jalisco y explicar los resultados obtenidos.
- Construye un grafica donde se comparen los diagramas de cajas de las superficies cosechadas por tipos de cultivos.