U-tad CENTRO UNIVERSITARIO DE TECNOLOGÍA Y ARTE DIGITAL

Estadística Descriptiva I

Mar Angulo Martínez
mar.angulo@u-tad.com
Curso 2023-2024

U-tad CENTRO UNIVERSITARIO DE TECNOLOGÍA Y ARTE DIGITAL

Temario

TEMA 1: Estadística descriptiva I. Distribuciones unidimensionales

1. Conceptos fundamentales.
2. Tablas estadísticas.
3. Gráficas estadísticas (visualización de datos).
4. Análisis de los datos.
    I. Medidas de posición.
    II. Medidas de variabilidad.
    III. Medidas de simetría.
    IV. Medidas de apuntamiento.

Conceptos fundamentales

Población: Conjunto de individuos definido por una o más propiedades que pretendemos observar y analizar.

Población finita. De tamaño N

vehículos matriculados en España en 2019
niños nacidos en Madrid en la última década
alumnos matriculados este curso en Universidades españolas

Población infinita. No podemos llegar a determinar su tamaño

resultados en el lanzamiento de un dado
tasa de crecimiento de un país

A veces las poblaciones finitas son de tamaño tan grande que se tratan como poblaciones infinitas (infinito numerable)

población mundial

Conceptos fundamentales

Muestra: Es un subconjunto formado por individuos de la población en el que se investiga/analiza alguna característica de una forma efectiva.

El tamaño de la muestra se denota n

- La condición esencial es que la muestra sea **representativa** de la población
- Una muestra perfecta sería la que "hiciera escala" con la población.

Muestreo: Procedimiento que permite extraer/seleccionar una muestra de tamaño n de una población infinita o de tamaño N

Conceptos fundamentales

Característica: es la cualidad o propiedad de los individuos que pretendemos analizar

Características cualitativas. No son susceptibles de medida.

- También se llaman variables categóricas o atributos
- Pueden ser ordinales o nominales. Y pueden ser booleanas (si/no).

fumadores/no fumadores, apto/no apto, presente/ausente
color del pelo (cualitativa nominal)
calificaciones: Ins/Sufic/Notable/Sobresal/MH (cualitativa ordinal)

Características cuantitativas. Son aquellas que se pueden medir numéricamente. Sus valores numéricos se denominan variables estadísticas

peso, estatura, IQ, nº viviendas en Madrid…

Conceptos fundamentales

Variables estadísticas
pueden ser:

En este curso, nos vamos a centrar en el estudio de variables estadísticas (discretas y continuas)

Variables discretas. Toman valores aislados. Concretamente no pueden tomar valores intermedios entre dos valores consecutivos previamente ordenados.

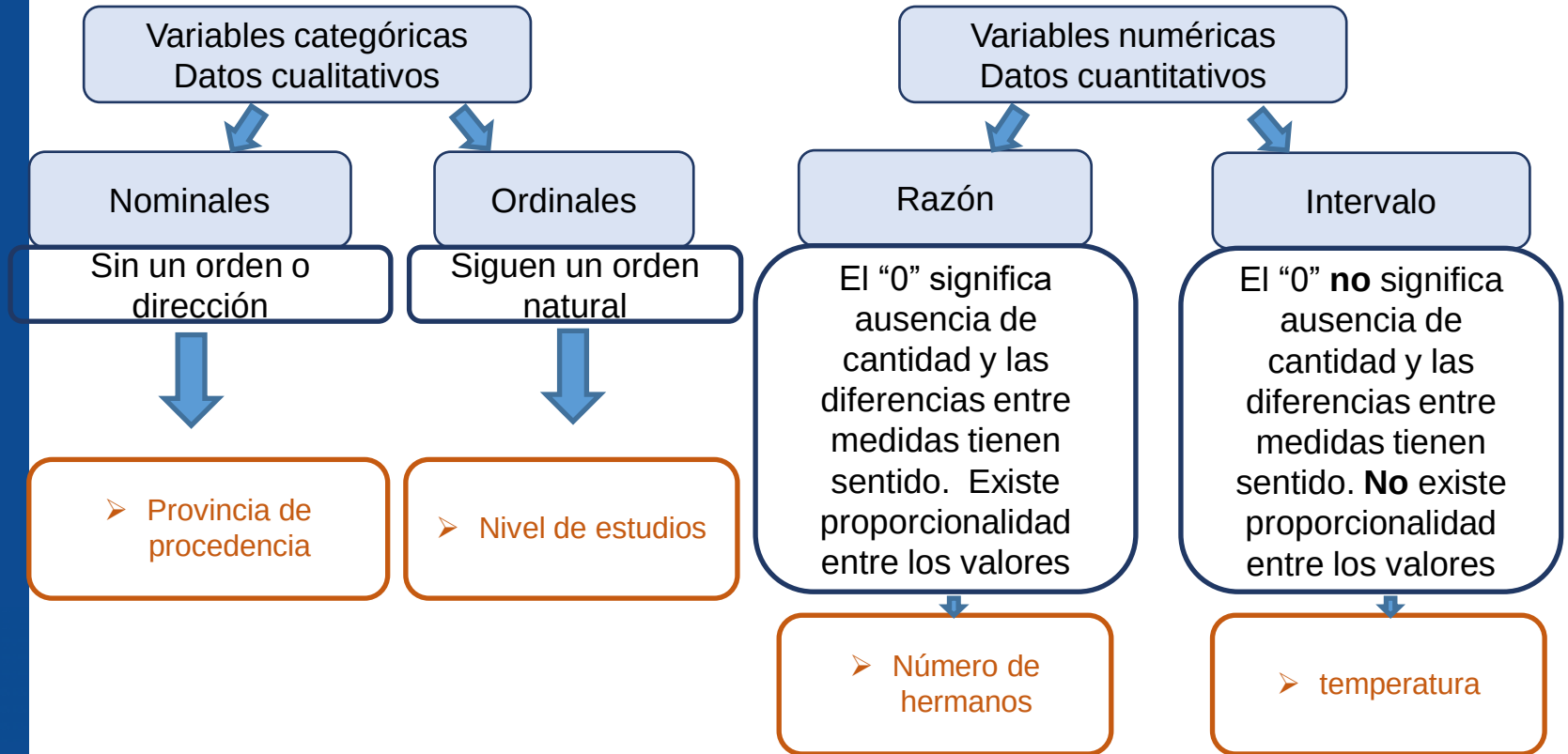
nº de encestes de un jugador en un torneo
nº de alumnos matriculados este curso en Universidades españolas

Variables continuas. Pueden tomar todos los valores de forma continua dentro del conjunto de los números reales

estatura/peso/nivel de glucosa en sangre…
tasa de crecimiento de un país

» ¿Cómo clasificarías la edad?
¿y los ingresos mensuales de un conjunto de familias?

Clasificación de variables/datos



	Tipo				Escala de medida			
	Categórica	Dummy	Numérica Discreta	Numérica Continua	Nominal	Ordinal	Intervalo	Razón
Gasto mensual (en euros) en actividades culturales de los universitarios de Madrid				X				X
Grado de satisfacción con un producto (mu bajo, bajo, medio…)	X					X		
Año de nacimiento			X				X	
Edad			X					X
Nivel en un juego (principiante, intermedio…)	X					X		
Existencia de parking en un hotel	X	X						
Salario de familias españolas				X				X

Conceptos fundamentales

El tratamiento de variables discretas es diferente al de variables continuas

Variable discreta: datos aislados

La variable toma los valores: $\{x_1, x_2, \dots, x_n\}$

Variable continua: datos agrupados en intervalos

La variable toma los valores dentro de intervalos: $\{I_1, I_2, \dots, I_k\}$, donde $k < n$

- Los intervalos se consideran cerrados por la izquierda y abiertos por la derecha: $I_i = [L_{i-1}, L_i)$
- El punto medio del intervalo se denomina **marca de clase**: $x_i = \frac{L_{i-1} + L_i}{2}$
- El ancho de cada intervalo se calcula por cada intervalo (y no tienen por qué ser siempre iguales): $c_i = |L_{i-1} - L_i|$

Conceptos fundamentales

Ejemplo: Calificaciones de un examen

Muestra:
{89, 80, 93, 64, 67, 72, 70, 66, 85, 89, 81, 97, 74, 82, 85, 63, 72, 81, 81, 95, 84, 81, 80, 72, 66, 60, 83, 85, 98, 84, 68, 90, 82, 69, 72 , 87}

Tamaño muestral:
$n = 36$

Una variable que toma un gran número de valores puede ser tratada como una variable continua. Podemos agrupar los valores en intervalos (número ideal de intervalos: entre 5 y 20).

Intervalo	Marca de clase	Número de calificaciones
[50, 60)	55	0
[60, 70)	65	8
[70, 80)	75	6
[80, 90)	85	17
[90, 100]	95	5

Tablas estadísticas (distribuciones de frecuencias)

Ejemplo 1:

X =“El número de encestes diarios de un jugador en un torneo en un mes”

Nº encestes x_i	nº días n_i	Frecuencia relativa f_i	Frecuencia absoluta acumulada N_i	Frecuencia relativa acumulada F_i
1	4	0.1333	4	0.1333
2	9	0.3	13	0.4333
3	8	0.2667	21	0.7
4	3	0.1	24	0.8
5	2	0.0667	26	0.8667
6	1	0.0333	27	0.9
7	0	0	27	0.9
8	3	0.1	**30**	**1**

$n = 30$ $\quad$ 1

¿Cómo se interpretan los datos del recuadro?

Tablas estadísticas (distribuciones de frecuencias)

Frecuencia absoluta (n_i): es el número de veces que se repite cada valor de la variable.

- *La suma de todas las frecuencias absolutas es n*

Frecuencia relativa (f_i): es el cociente entre la frecuencia absoluta y el tamaño de la muestra. Representa el tanto por uno que corresponde a cada uno de los valores de la muestra.

$$f_i = \frac{n_i}{n}$$

- *La suma de todas las frecuencias relativas es 1*

Tablas estadísticas (distribuciones de frecuencias)

Frecuencia absoluta acumulada (N_i): es la suma de las frecuencias absolutas correspondientes a x_i y a todos los valores que son menores que él

$$N_i = \sum_{j=1}^{i} n_j$$

- *La última frecuencia absoluta acumulada es n*

Tablas estadísticas (distribuciones de frecuencias)

Frecuencia relativa acumulada (F_i): es la suma de las frecuencias relativas correspondientes a x_i y a todos los valores que son menores que él

$$F_i = \frac{\sum_{j=1}^{i} n_j}{n} = \sum_{j=1}^{i} f_j$$

- *La última frecuencia relativa acumulada es 1*

Tablas estadísticas (distribuciones de frecuencias)

x_i	n_i	f_i	N_i	F_i
x_1	n_1	f_1	N_1	F_1
x_2	n_2	f_2	N_2	F_2
…	…	…	…	…
x_k	n_k	f_k	$N_k = n$	$F_k = 1$
	$\| \| n$	$\| \| 1$		

Caso discreto

Caso continuo

Intervalo	x_i	n_i	f_i	N_i	F_i
$I_1 = [L_0, L_1)$	x_1	n_1	f_1	N_1	F_1
$I_2 = [L_1, L_2)$	x_2	n_2	f_2	N_2	F_2
…	…	…	…	…	…
$I_k = [L_{k-1}, L_k)$	x_k	n_k	f_k	$N_k = n$	$F_k = 1$
		$\| \| n$	$\| \| 1$		

Tablas estadísticas (distribuciones de frecuencias)

Ejemplo 2:

X =“Valor de las compras en alimentación de un hotel durante el mes de agosto”

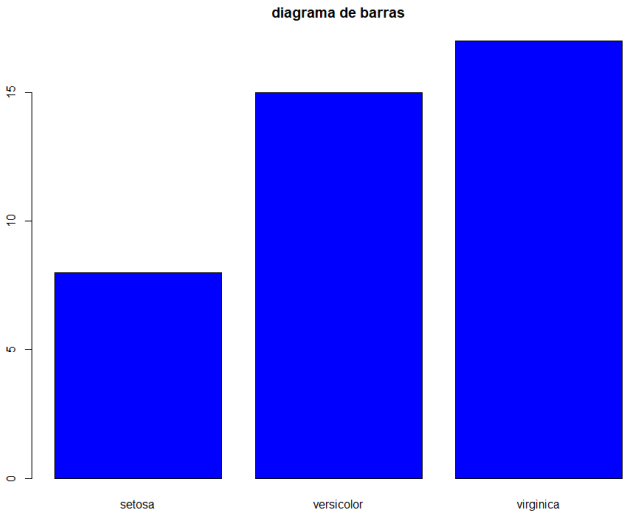
Gasto (€·10^3) I_i	nº días n_i	Frecuencia relativa f_i	Frecuencia absoluta acumulada N_i	Frecuencia relativa acumulada F_i
[0, 1.6)	8	0.2667	8	0.2667
[1.6, 3.2)	13	0.4333	21	0.7
[3.2, 4.8)	6	0.2	27	0.9
[4.8, 6.4)	1	0.0333	28	0.9333
[6.4, 8.0)	2	0.0667	**30**	**1**

$n = 30$ $\quad$ 1

¿Cómo se interpretan los datos del recuadro?

Gráficas estadísticas (visualización de datos)

Diagrama de barras: Para representar las frecuencias n_i

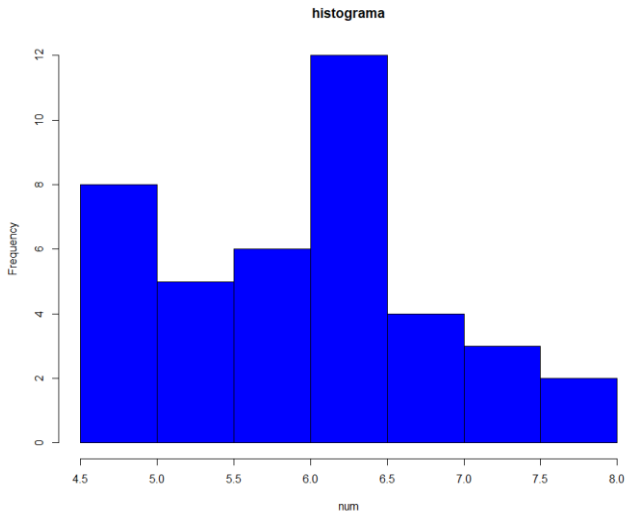


SOLO variables categóricas

```r
# Lectura del dataframe
datos = iris
head(datos)
dim(datos)

# Seleccion de numeros aleatorios
ids = sample(1:150, 40, replace=TRUE)

# Muestra aleatoria del dataframe
sample.datos = datos[ids,]
head(sample.datos)
dim(sample.datos)

# Variable categorica
cat = sample.datos$Species

# Frecuencias
table(cat)

# Diagrama de barras
barplot(table(cat), main = 'diagrama de barras', col = 'blue')
```

Gráficas estadísticas (visualización de datos)

Histograma: Para representar las frecuencias n_i



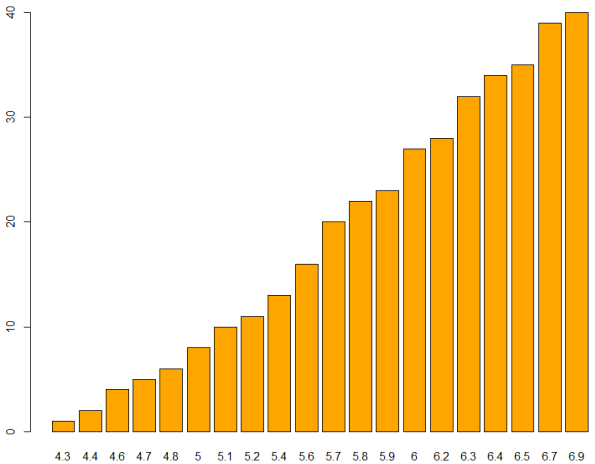
SOLO variables continuas

```r
# Variable numerica
num = sample.datos$Sepal.Length

# Frecuencias
table(num)

# Diagrama de barras
hist(num, main = 'histograma', col = 'blue')
hist(num, main = 'histograma', col = 'orange', breaks = 15)
```

Observamos que agrupa por intervalos

Gráficas estadísticas (visualización de datos)

Diagrama de frecuencias acumuladas: F_i, N_i

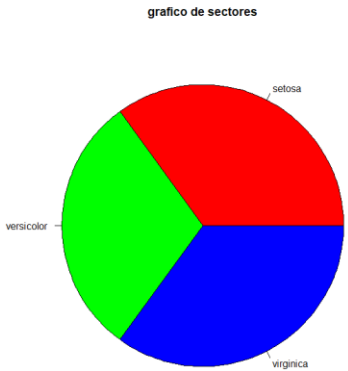


SOLO variables continuas

```r
# Variable numerica
num = sample.datos$Sepal.Length
N = length(num)

# Frecuencias
table(num)

# Frecuencias acumuladas (absolutas)
cumsum(table(num))

# Diagrama de frecuencias acumuladas (absolutas)
barplot(cumsum(table(num)), col = 'orange')

# Frecuencias acumuladas (relativas)
cumsum(table(num))/N

# Diagrama de frecuencias acumuladas (relativas)
barplot(cumsum(table(num))/N, col = 'red')
```

Gráficas estadísticas (visualización de datos)

Diagrama de sectores: f_i



SOLO variables categóricas

```r
# Variable categorica
cat = sample.datos$Species

# Frecuencias
table(cat)

# Diagrama de sectores
pie(table(cat), main = 'grafico de sectores', col = rainbow(3))
```

https://r-coder.com/grafico-sectores-r/

Análisis de datos

1. **Medidas de posición**
   I. Centrales
      - Media aritmética
      - Media geométrica
      - Media armónica
      - Mediana
      - Moda
   II. No centrales
      - Cuartiles
      - Deciles
      - Percentiles
2. **Medidas de dispersión**
   - Varianza
   - Desviación típica
   - Coeficiente de variación de Pearson
3. **Medidas de asimetría**
   - Coeficiente de asimetría de Pearson
   - Coeficiente de asimetría de Bowley
   - Coeficiente de asimetría de Fisher
4. **Medidas de kurtosis o apuntamiento**
   - Coeficiente de apuntamiento de Fisher

Medidas de posición central

La **Media aritmética** de una serie de observaciones $\{x_1, x_2, \dots, x_n\}$ se define como el promedio de los valores de la muestra

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

A partir de la muestra

Si los datos están agrupados, y $\{x_1, x_2, \dots, x_k\}$ son las marcas de clase (puntos medios de los intervalos), la media aritmética es

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i \cdot n_i}{n} = \sum_{i=1}^{k} x_i \cdot f_i$$

A partir de la tabla de frecuencias

Ventaja: en su cálculo intervienen todos los valores
Inconveniente: se ve muy afectada por los valores extremos

Medidas de posición central

La **Media geométrica** de una serie de observaciones $\{x_1, x_2, \dots, x_n\}$, se define como la raíz n-ésima del producto de los valores de la variable elevados a sus correspondientes frecuencias

$$G = \sqrt[n]{\prod_{i=1}^{n} x_i} = \sqrt[n]{\prod_{i=1}^{k} x_i{}^{n_i}}$$

- *Se utiliza para promediar variables que presentan variaciones acumulativas como porcentajes, tasas, índices…*
- *No es aplicable cuando algún valor es nulo o negativo.*

Ejemplo: El nivel de ahorro de una persona durante los dos primeros años es del 10% y el tercer año es del 20%, el cuarto año sólo logra un 15% y los tres siguientes, un 17%. Calcular la tasa de ahorro promedio en estos años.

$G = \sqrt[7]{10^2 . 20^1 15^1 . 17^3}$ = 14.687% de ahorro promedio

Medidas de posición central

La **Media armónica** de una serie de observaciones $\{x_1, x_2, \dots, x_n\}$, se utiliza para promediar rendimientos, velocidades, productividades… variables que se expresan en términos relativos

$$H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{n}{\sum_{i=1}^{k} \frac{n_i}{x_i}}$$

Ejemplo: Cuatro fincas han producido 100, 120, 150 y 200 Qm con unos rendimientos de trigo de 10, 15, 12 y 18 Qm/Ha. Obtener el rendimiento medio.

$$H = \frac{570}{\frac{100}{10} + \frac{120}{15} + \frac{150}{12} + \frac{200}{18}} = 13.698 \text{ Qm/Ha}$$

Medidas de posición central

Ejemplo 1: X ="El número de encestes diarios de un jugador en un torneo en un mes" $n = 30$

x_i	n_i	f_i	N_i	F_i
1	4	0.1333	4	0.1333
2	9	0.3	13	0.4333
3	8	0.2667	21	0.7
4	3	0.1	24	0.8
5	2	0.0667	26	0.8667
6	1	0.0333	27	0.9
7	0	0	27	0.9
8	3	0.1	**30**	**1**

$$\bar{x} = \frac{4 \cdot 1 + 9 \cdot 2 + 8 \cdot 3 + 3 \cdot 4 + 2 \cdot 5 + 1 \cdot 6 + 7 \cdot 0 + 3 \cdot 8}{n} = \frac{98}{30} = 3.27$$

$$G = \sqrt[30]{1^4 . 2^9 . 3^8 . 4^3 . 5^2 . 6^1 . 7^0 . 8^3} = 2.758$$

$$H = \frac{30}{\frac{4}{1} + \frac{9}{2} + \frac{8}{3} + \frac{3}{4} + \frac{2}{5} + \frac{1}{6} + \frac{0}{7} + \frac{3}{8}} = 2.334$$

Las tres medias se expresan en las mismas unidades que la variable; en este ejemplo sería el número medio de encestes diarios del jugador en el torneo

Medidas de posición central

Ejemplo 2: X =“Valor de las compras en alimentación de un hotel durante el mes de agosto” $n = 30$

I_i	n_i	x_i	$n_i x_i$
[0, 1.6)	8	0.8	6.4
[1.6, 3.2)	13	2.4	31.2
[3.2, 4.8)	6	4	24
[4.8, 6.4)	1	5.6	5.6
[6.4, 8.0)	2	7.2	14.4
			$\|\|$
			81.6

$$\bar{x} = \frac{81.6}{30} = 2.72$$

- *Podemos usar la tabla como referencia para realizar los cálculos más cómodamente*

Medidas de posición central

La **Mediana** de una serie de observaciones $\{x_1, x_2, \dots, x_n\}$, representa el valor de la variable tal que, si los valores están ordenados de menor a mayor, la mitad de los valores son menores o iguales y el resto mayores. Se denota Me.

Cálculo de la mediana:

1) Se calcula $n/2$

a) Variable discreta (Valores aislados)

2) Si $N_{i-1} < n/2 < N_i \rightarrow Me = x_i$

2) Si $N_{i-1} = n/2 < N_i \rightarrow Me = \frac{x_{i-1}+x_i}{2}$

b) Variable continua (Valores agrupados en intervalos)

2) Si $N_{i-1} \leq n/2 < N_i \rightarrow Me$ está en el intervalo $[L_{i-1}, L_i)$

3) $Me = L_{i-1} + \frac{n/2-N_{i-1}}{n_i} c_i$

- *Ventaja:*
  Se ve menos afectada que la media por los valores extremos
- *Inconveniente:*
  En su cálculo no intervienen todos los valores de la variable

Si a los valores se les suma (multiplica) una constante, la media y la mediana quedan aumentadas (multiplicadas) en la misma cantidad

Medidas de posición central

La **Moda** de una serie de observaciones $\{x_1, x_2, \dots, x_n\}$, representa el valor más frecuente de la variable. Es el valor al que corresponde la mayor frecuencia absoluta. Se denota Md.

Cálculo de la moda:

a) Variable discreta (Valores aislados)

1) Es el valor con mayor frecuencia n_i

b) Variable continua (Valores agrupados en intervalos)

1) Se observa el mayor n_i (o el mayor $k_i = n_i/c_i$) → Md está en el intervalo $[L_{i-1}, L_i)$

2) Si tiene amplitudes iguales (mismos c_i): $Md = L_{i-1} + \frac{n_i - n_{i-1}}{(n_i - n_{i-1}) + (n_i - n_{i+1})} c_i$

2) Si tiene amplitudes distintas (distintos c_i): $Md = L_{i-1} + \frac{k_i - k_{i-1}}{(k_i - k_{i-1}) + (k_i - k_{i+1})} c_i$,

donde $k_i = \frac{n_i}{c_i}$ es la densidad de frecuencia.

Medidas de posición central

Ejemplo 2: X =“Valor de las compras en alimentación de un hotel durante el mes de agosto”

I_i	n_i	N_i
[0, 1.6)	8	8
[1.6, 3.2)	13	21
[3.2, 4.8)	6	27
[4.8, 6.4)	1	28
[6.4, 8.0)	2	30

$n = 30$

- *Moda:*

Dado que los intervalos son de amplitud constante, la moda está en el intervalo de mayor n_i.
Por tanto, Md está en [1.6, 3.2):

$$Md = L_{i-1} + \frac{n_i - n_{i-1}}{(n_i - n_{i-1}) + (n_i - n_{i+1})}\, c_i = 1.6 + \frac{13 - 8}{(13 - 8) + (13 - 6)} 1.6 = 2.26$$

Medidas de posición central

Ejemplo 3: Supongamos una variable con la siguiente distribución no uniforme de frecuencias

I_i	n_i	N_i
[2-4)	9	9
[4-6)	15	24
[6-8)	5	29
[8-12)	9	38
[12-20)	8	46

$n = 46$

- *Moda:*

Dado que los intervalos NO son de amplitud constante, la moda está en el intervalo de mayor k_i:

I_i	n_i	N_i	c_i	k_i
[2-4)	9	9	2	4.5
[4-6)	15	24	2	7.5
[6-8)	5	29	2	2.5
[8-12)	9	38	4	2.25
[12-20)	8	46	8	1

→ Por tanto, Md está en [4, 6)

$$Md = L_{i-1} + \frac{k_i - k_{i-1}}{(k_i - k_{i-1}) + (k_i - k_{i+1})} c_i = 4 + \frac{7.5 - 4.5}{(7.5 - 4.5) + (7.5 - 2.5)} \cdot 2 = 4.75$$

Medidas de posición NO central

Las medidas de posición NO central,

- Se expresan también en las mismas unidades que la variable
- No intervienen en su cálculo todos los valores de la variable

Medidas de posición NO central

Cuartiles: Q_1 , Q_2, Q_3

- **Primer Cuartil Q_1** es el valor de la variable tal que, si los valores están ordenados de menor a mayor, la cuarta parte de los valores ($n/4$) son menores o iguales que él y el resto mayores.
- **Tercer Cuartil Q_3** es el valor de la variable tal que, si los valores están ordenados de menor a mayor, tres cuartas partes de los valores ($3n/4$) son menores o iguales que él y el resto mayores.
- **Segundo Cuartil Q_2** = Me

- *Son por tanto los valores que separan en cuartas partes el total de los datos*
- *Cálculo de k-Cuartil: igual que la mediana cambiando $n/2$ por $kn/4$*

Medidas de posición NO central

Percentiles: $P_1, P_2, \dots, P_{99}$

- **11-Percentil $\boldsymbol{P_{11}}$** es el valor de la variable tal que, si los valores están ordenados de menor a mayor, un 11% de los valores ($11n/100$) son menores o iguales que él y el resto mayores.
- **76-Percentil $\boldsymbol{P_{76}}$** es el valor de la variable tal que, si los valores están ordenados de menor a mayor, un 76% de los valores ($76n/100$) son menores o iguales que él y el resto mayores.
- **50-Percentil $\boldsymbol{P_{50}}$** = Me

- *Son por tanto los valores que separan en centésimas partes el total de los datos*
- *Cálculo de k-Percentil: igual que la mediana cambiando $n/2$ por $kn/100$*

Medidas de posición NO central

Deciles: D_1 , $D_{2,\ldots}$ D_9

- **3-Decil D_3** es el valor de la variable tal que, si los valores están ordenados de menor a mayor, 3 décimas partes de los valores $(3n/10)$ son menores o iguales que él y el resto mayores.
- **9-Decil D_9** es el valor de la variable tal que, si los valores están ordenados de menor a mayor, 9 décimas partes de los valores $(9n/10)$ son menores o iguales que él y el resto mayores.
- **5- Decil D_5** = Me

- *Son por tanto los valores que separan en décimas partes el total de los datos*
- *Cálculo de k-Decil: igual que la mediana cambiando $n/2$ por $kn/10$*

Medidas de posición NO central

Ejemplo 1: X ="El número de encestes de un jugador en un torneo en un mes"

x_i	n_i	N_i
1	4	4
2	9	13
3	8	21
4	3	24
5	2	26
6	1	27
7	0	27
8	3	**30**

$$n = 30$$

- *Mediana:*

$$n/2 = 15 \Longrightarrow 13 < 15 < 21 \Longrightarrow Me = 3$$

- *Moda:*

$$Md = 2$$

- *Cuartiles:*

$$n/4 = 7{,}5 \Longrightarrow 4 < 7{,}5 < 13 \Longrightarrow Q_1 = 2$$
$$3n/4 = 22{,}5 \Longrightarrow 21 < 22{,}5 < 24 \Longrightarrow Q_3 = 4$$

- *Percentil 20 y 95:*

$$20n/100 = 6 \Longrightarrow 4 < 6 < 13 \Longrightarrow P_{20} = 2$$
$$95n/100 = 28{,}5 \Longrightarrow P_{95} = 8$$

- *Deciles 3 y 7:*

$$3n/10 = 9 \Longrightarrow D_3 = 2$$
$$7n/10 = 21 \Longrightarrow D_7 = \frac{3+4}{2} = 3{,}5$$

Medidas de posición NO central

Ejemplo 2: X =“Valor de las compras en alimentación de un hotel durante el mes de agosto” $n = 30$

I_i	n_i
[0, 1.6)	8
[1.6, 3.2)	13
[3.2, 4.8)	6
[4.8, 6.4)	1
[6.4, 8.0)	2

- *Mediana:*
- *Moda:*
- *Cuartiles:*
- *Percentil 20 y 95:*
- *Deciles 3 y 9:*

Medidas de posición NO central

Ejemplo 2: X =“Valor de las compras en alimentación de un hotel durante el mes de agosto” $n = 30$

I_i	n_i	Ni
[0, 1.6)	8	8
[1.6, 3.2)	13	21
[3.2, 4.8)	6	27
[4.8, 6.4)	1	28
[6.4, 8.0)	2	30

- *Mediana:*

Calculamos $n/2 = 15$ → Me está en [1.6, 3.2)

$$Me = L_{i-1} + \frac{n/2 - N_{i-1}}{n_i}|L_i - L_{i-1}| = 1.6 + \frac{15 - 8}{13}|3.2 - 1.6| = 2.46$$

- *Cuartiles:*

Calculamos $n/4 = 7.5$ y $3n/4 = 22.5$ → Q_1 y Q_3 están, respectivamente, en [0, 1.6) y [3.2, 4.8)

$$Q_1 = L_{i-1} + \frac{n/4 - N_{i-1}}{n_i}|L_i - L_{i-1}| = 0 + \frac{7.5 - 0}{8}|1.6 - 0| = 1.5$$

$$Q_3 = L_{i-1} + \frac{3n/4 - N_{i-1}}{n_i}|L_i - L_{i-1}| = 3.2 + \frac{22.5 - 21}{6}|4.8 - 3.2| = 3.6$$

Medidas de posición NO central

Y vamos a interpretarlos

La mitad de los días de agosto el gasto del hotel en alimentación fue como mucho de 2.460 euros

La cuarta parte del mes el gasto no pasó de 1.500 euros y 3.600 euros es el gasto a partir del cual un día está entre la cuarta parte (=25%) de los días en que más se gastó.

Medidas de posición NO central

Ejemplo 2: X =“Valor de las compras en alimentación de un hotel durante el mes de agosto” $n = 30$

- *Percentiles 20 y 95:*

Calculamos $20n/100 = 6$ y $95n/100 = 28{,}5$ → P_{20} y P_{95} están, respectivamente, en [0, 1.6) y [6.4, 8.0)

$$P_{20} = L_{i-1} + \frac{20n/100 - N_{i-1}}{n_i}|L_i - L_{i-1}| = 0 + \frac{6-0}{8}|1.6 - 0| = 1.2$$

$$P_{95} = L_{i-1} + \frac{95n/100 - N_{i-1}}{n_i}|L_i - L_{i-1}| = 6.4 + \frac{28{,}5-28}{2}|8.0 - 6.4| = 6.8$$

I_i	n_i	Ni
[0, 1.6)	8	8
[1.6, 3.2)	13	21
[3.2, 4.8)	6	27
[4.8, 6.4)	1	28
[6.4, 8.0)	2	30

A partir de 6.800 euros, un día está entre el 5% de los de mayor gasto del hotel en alimentación

El 20% de los días el gasto es como máximo de 1.200 euros

Medidas de dispersión

Recorrido o Rango de una serie de observaciones $\{x_1, x_2, \dots, x_n\}$, es la diferencia entre el valor mayor y el menor de las mismas

$$Recorrido = \max\{x_1, x_2, \dots, x_n\} - \min\{x_1, x_2, \dots, x_n\}$$

Rango intercuartílico es la diferencia entre el tercer y primer cuartil

$$R.I. = Q_3 - Q_1$$

- *Las medidas de dispersión miden el grado de "esparcimiento" de los datos, si están concentrados o dispersos en torno a una medida de posición central.*

Medidas de dispersión

Varianza (muestral) de una serie de observaciones $\{x_1, x_2, \dots, x_n\}$

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = \frac{\sum_{i=1}^{n} {x_i}^2}{n} - \bar{x}^2 = \frac{\sum_{i=1}^{k} n_i \cdot (x_i - \bar{x})^2}{n} = \frac{\sum_{i=1}^{k} n_i \cdot {x_i}^2}{n} - \bar{x}^2$$

- *Inconveniente: No se expresa en las mismas unidades que la variable.*
- *Si los datos se multiplican por una constante $\delta \in \mathbb{R}$, la varianza queda multiplicada por δ^2.*
- *La varianza es invariante respecto a un cambio de localización.*
- *La varianza es siempre positiva.*

Medidas de dispersión

Desviación típica (muestral) de una serie de observaciones $\{x_1, x_2, \dots, x_n\}$ es la raíz cuadrada positiva de la varianza:

$$s = +\sqrt{s^2}$$

- *Es siempre positiva y tiene las mismas unidades que los datos observados.*
- *Si los datos se multiplican por una constante $\delta \in \mathbb{R}$, la varianza queda multiplicada por δ.*
- *La desviación típica es invariante respecto a un cambio de localización.*
- *La desviación típica es siempre positiva.*

Medidas de dispersión

El **Coeficiente de Variación de Pearson (C.V.)** es una medida de dispersión relativa que da la proporción existente entre la desviación típica y la media,

$$C.V. = \frac{s}{\bar{x}}$$

- *Es invariante a los cambios de escala*
- *No es invariante a los cambios de origen*
- *La media no puede ser nula para que tenga sentido*

Medidas de dispersión

Ejemplo 1: X ="El número de encestes diarios de un jugador en un torneo en un mes" $n = 30$

x_i	n_i	N_i	$n_i \cdot x_i$	$n_i \cdot {x_i}^2$
1	4	4	4	4
2	9	13	18	36
3	8	21	24	72
4	3	24	12	48
5	2	26	10	50
6	1	27	6	36
7	0	27	0	0
8	3	30	24	192
			=	=
			98	438

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i \cdot n_i}{n} = \frac{98}{30} = 3.27$$

$$s^2 = \frac{\sum_{i=1}^{k} n_i \cdot {x_i}^2}{n} - \bar{x}^2 = \frac{438}{30} - 3.27^2 = 3.9$$

$$s = +\sqrt{s^2} = +\sqrt{3.9} = 1.9$$

$$C.V. = \frac{s}{\bar{x}} = \frac{1.9}{3.27} = 0.6$$

Medidas de dispersión

Ejemplo 2: X =“Valor de las compras en alimentación de un hotel durante el mes de agosto” $n = 30$

I_i	n_i	x_i	$n_i \cdot x_i$	$n_i \cdot {x_i}^2$
[0, 1.6)	8	0.8	6,4	5.12
[1.6, 3.2)	13	2.4	31,2	74.8
[3.2, 4.8)	6	4	24	96
[4.8, 6.4)	1	5.6	5,6	31.36
[6.4, 8.0)	2	7.2	14,4	103.8
			$\sum$	$\sum$
			81.6	311.16

$$\bar{x} = \frac{81.6}{30} = 2.72$$

$$s^2 = \frac{311.16}{30} - 2.72^2 = 2.97$$

$$s = +\sqrt{s^2} = +\sqrt{2.97} = 1.72$$

$$C.V. = \frac{s}{\bar{x}} = \frac{1.72}{2.72} = 0.62$$

Transformaciones lineales

Ejemplo : X =“salarios de seis trabajadores de una empresa”

Salarios (dólares)
1000
2500
1650
1550
1268
1900

Medidas (dólares)	Valores
Media	1.644,67 $
Varianza	272.650,67 $dólares^2$
Desviación típica	522,16 dólares
Coeficiente de variación	31,37%

Y=X+250
1250
2750
1900
1800
1518
2150

T=1,07X
1070
2675
1765,5
1658,5
1356,76
2033

Cambio de origen Y=X+250	Valores
Media	1.894,67 $
Varianza	272.650,67 $dólares^2$
Desviación típica	522,16 dólares
Coeficiente de variación	27,56%

Cambio de escala Y=1,07X	Valores
Media	1.759,79 $=1,07x1644,67
Varianza	312.157,75 $dólares^2$=$(1,07)^2$x 272.650,67
Desviación típica	558,71 dólares=522,16x1,07
Coeficiente de variación	31,37%

U-tad
CENTRO UNIVERSITARIO DE TECNOLOGÍA Y ARTE DIGITAL

Medidas de dispersión

Datos atípicos (*outliers*):

- Son las observaciones que, numéricamente, distan del resto de los datos de una manera importante.
- En muchos casos, los *outliers* son datos erróneos insertados en las bases de datos.
- La inclusión de estos datos en cualquier estudio estadístico produciría una distorsión importante.
- El tratamiento de los datos atípicos es pues, decisivo, pero subjetivo.

Valores atípicos leves:

$$q < Q_1 - 1.5\ RI \text{ y } q > Q_3 + 1.5\ RI$$

Valores atípicos extremos:

$$q < Q_1 - 3\ RI \text{ y } q > Q_3 + 3\ RI$$

Elementos de un Box plot

•Ordenar los datos y obtener el valor mínimo, el máximo, los cuartiles Q_1, Q_2 y Q_3 y el rango intercuartílico RI
•Calcular Lím. Inferior = Q1 – 1.5 RI y Lím. Sup = Q3 + 1.5 RI

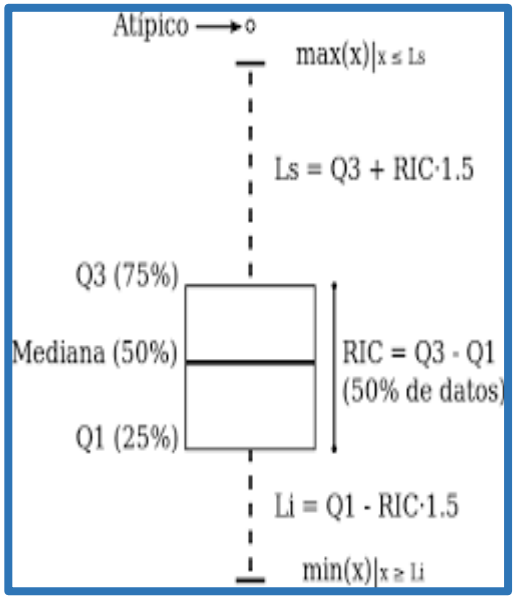
- el 50% de los valores se encuentra dentro de la caja
- El rango intercuartílico RI corresponde a la longitud de la caja

- Trazar líneas desde los extremos de la caja (bigotes) hasta los datos mínimo y máximo que son respectivamente >=Linf o <=Lsup
- Si hay valores atípicos: trazar líneas hasta Linf y/o Lsup y anotar los valores atípicos con círculos y los valores extremos con asteriscos.

- Si los datos están distribuidos con normalidad, aproximadamente el 95% de los datos se encuentre entre los bigotes.

Diagrama de cajas y bigotes (*boxplot*):

- ✓ Proporcionan una visión general de la simetría de la distribución de los datos;
- ✓ Son útiles para ver la presencia de valores atípicos (outliers)
- ✓ Se puede ver también la presencia o no de valores extremos



Medidas de dispersión

Diagrama de cajas y bigotes (*boxplot*):

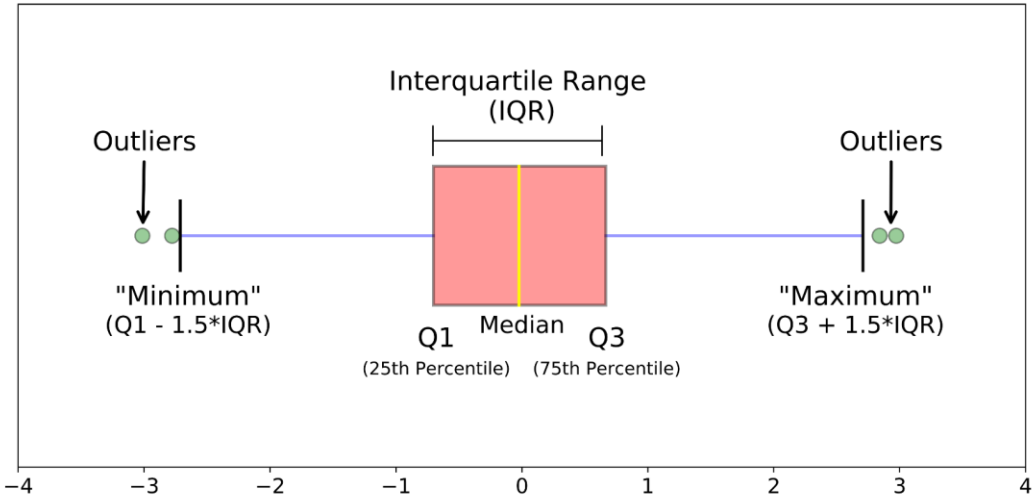
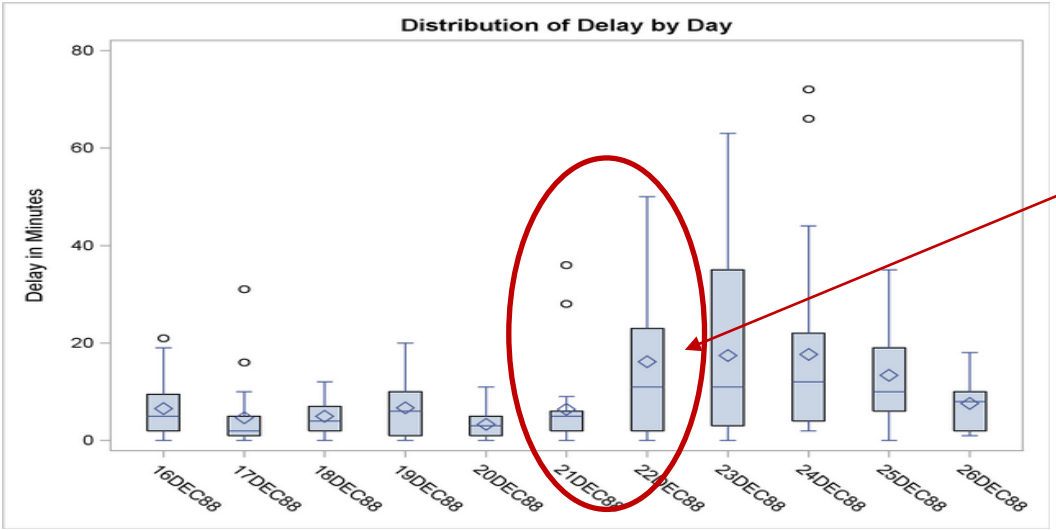


Diagrama de cajas y bigotes (*boxplot*):



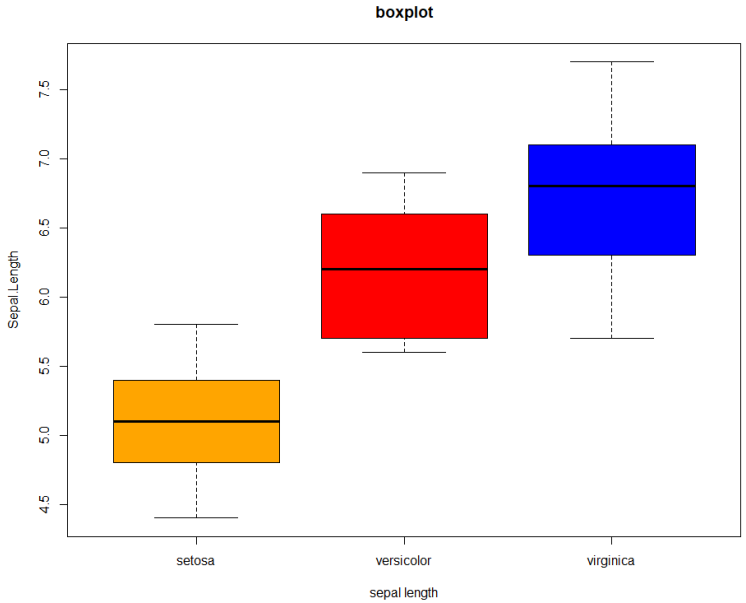
- Retrasos mínimos similares
- Q1 similares
- Pero… salvo outliers, el retraso máximo del día 21 fue menor que el que se produjo en más de la mitad de los vuelos del día 22

¿Cómo interpretar el box plot?

- ¿Qué día dirías que hubo menos retrasos? excluyendo valores atípicos, el 21 diciembre no hubo ningún vuelo con retraso superior a 20 minutos. Similar el día 17 diciembre
- El 23 de diciembre (y en menor medida el día 22) hubo una gran dispersión sobre todo en los retrasos mayores, hasta más de una hora. ¿Entre qué valores está el 50% central?

Gráficas estadísticas (visualización de datos)

Boxplots



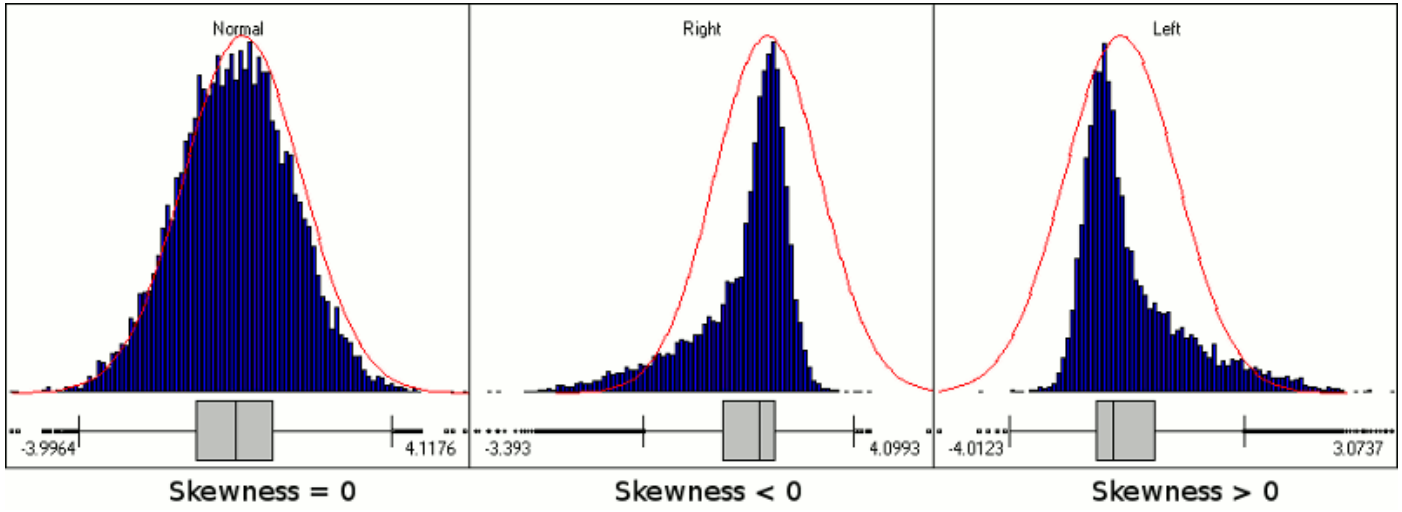
SOLO variables continuas

```r
# Boxplot
boxplot(num,
        main="boxplot",
        col = 'orange',
        xlab = "sepal length")

boxplot(Sepal.Length~Species,
        data=sample.datos,
        main="boxplot",
        col = c('orange','red', 'blue'),
        xlab = "sepal length")

q1    <- quantile(iris$sepal_length, 0.25)
q2    <- quantile(iris$sepal_length, 0.50) # Mediana
q3    <- quantile(iris$sepal_length, 0.75)
RI    <- q3-q1
left  <- q1-1.5*RI
right <- q3+1.5*RI
```

Medidas de asimetría

Una distribución de datos es **Asimétrica a la derecha (positiva)** si las frecuencias descienden más lentamente por la derecha.

Una distribución de datos es **Asimétrica a la izquierda (negativa)** si las frecuencias descienden más lentamente por la izquierda



Medidas de asimetría

Coeficiente de asimetría de Fisher:

$$A_F = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n \cdot s^3} = \frac{\sum_{i=1}^{k} n_i \cdot (x_i - \bar{x})^3}{n \cdot s^3}$$

Si $A_F > 0$ → Asimetría positiva.

Si $A_F = 0$ → Simétrica.

Si $A_F < 0$ → Asimetría negativa.

- *Si la distribución es simétrica, entonces coinciden media, mediana y moda.*

Medidas de asimetría

Coeficiente de asimetría de Bowley:

$$A_B = \frac{(Q_3 - Me) - (Me - Q_1)}{Q_3 - Q_1}$$

Si $A_B > 0$ → Asimétrica por la derecha.

Si $A_B = 0$ → Simétrica.

Si $A_B < 0$ → Asimétrica por la izquierda.

Medidas de asimetría

Coeficiente de asimetría de Pearson:

$$A_P = \frac{\bar{x} - Md}{s}$$

Si $A_P > 0$ → Asimétrica por la derecha.

Si $A_P = 0$ → Simétrica.

Si $A_P < 0$ → Asimétrica por la izquierda.

Medida de asimetría y apuntamiento

Ejemplo 2: X ="Valor de las compras en alimentación de un hotel durante el mes de agosto" $n = 30$

I_i	n_i	x_i	$(x_i - \bar{x})^3 . n_i$	$(x_i - \bar{x})^4 . n_i$
[0, 1.6)	8	0,8	-56.623	
[1.6, 3.2)	13	2,4	-0.426	
[3.2, 4.8)	6	4	12.582	
[4.8, 6.4)	1	5,6	23.88	
[6.4, 8.0)	2	7,2	179.83	

159.24

$$A_F = \frac{\sum_{i=1}^{k} n_i \cdot (x_i - \bar{x})^3}{n \cdot s^3} = \frac{159{,}24}{30 \cdot 1.72^3} > 0$$

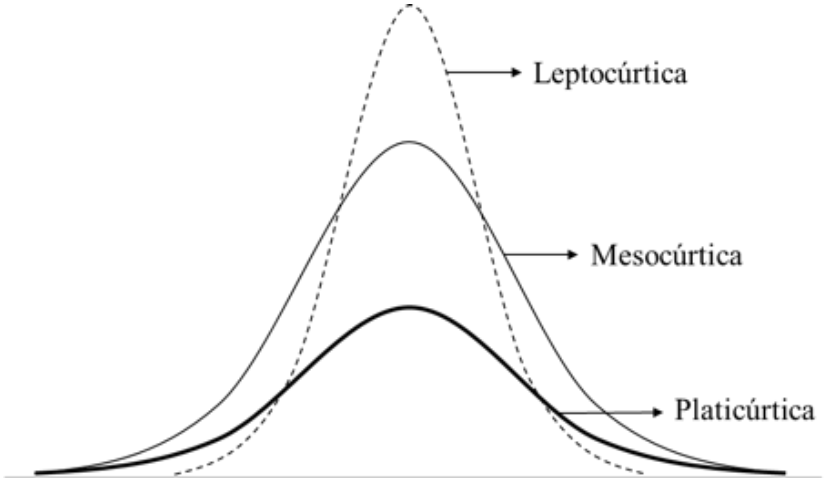
$$A_B = \frac{(Q_3 - Me) - (Me - Q_1)}{Q_3 - Q_1} = \frac{(3.6 - 2.46) - (2.46 - 1.5)}{3.6 - 1.5} = 0.086 > 0$$

$$A_P = \frac{\bar{x} - Md}{s} = \frac{2.72}{1.69} = 1.093 > 0$$

Asimétrica por la derecha

Medidas de apuntamiento

El **apuntamiento** (**Kurtosis**) de una distribución de datos se compara con el de una distribución normal,



Medidas de apuntamiento

Coeficiente de apuntamiento de Fisher:

$$g_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{n \cdot s^4} = \frac{\sum_{i=1}^{k} n_i \cdot (x_i - \bar{x})^4}{n \cdot s^4}$$

Si $g_2 > 3$ → Más apuntada de la normal: leptocúrtica.

Si $g_2 = 3$ → Igual a la campana de Gauss (distribución normal): mesocúrtica.

Si $g_2 < 3$ → Menos apuntada de la normal: platicúrtica.

- *Si la distribución es simétrica, entonces coinciden media, mediana y moda.*