Estadística descriptiva

Contenidos adaptados del libro “Probability and statistics for computer scientists, Second edition, M. Baron” (Capítulo 8)

Contenidos

1. Objetivos
2. Introducción
3. Cálculo de medidas estadísticas
4. Tabla de frecuencias (no está en el libro)
5. Gráficos estadísticos
6. Estadística descriptiva bidimensional
7. Resumen

4. Tabla de frecuencias

- Una **tabla de frecuencias** para una variable estadística cuantitativa o cualitativa representa:
  - el número de veces que se repite cada valor de la variable en una población o en una muestra (frecuencias absolutas),
  - el número de veces que se repite cada valor de la variable en una población o en una muestra dividido por el total de elementos de la población o muestra (frecuencias relativas).
- Una **tabla de frecuencias para datos agrupados** en intervalos para una variable estadística cuantitativa continua o discreta representa:
  - el número de valores que pertenecen a cada uno de los intervalos iguales en los que se haya divido el rango de la población o muestra
- Los intervalos se suelen denominar *bins* o “intervalos de clase”

4. Tabla de frecuencias
Columnas

- $\boldsymbol{Fila:}$ Número de fila de la tabla, desde la fila 1 hasta la fila k
- $\boldsymbol{x}$ **ó** $\boldsymbol{X}$: Valores diferentes $(v_1, v_2, \dots, v_k)$ que tiene la variable estadística en la población (x) o muestra (X)
- $\boldsymbol{f}$: Frecuencia absoluta: número de veces que aparece cada valor en la población o muestra
- $\boldsymbol{h}$: Frecuencia relativa
- $\boldsymbol{F}$: Frecuencia absoluta acumulada
- $\boldsymbol{H}$: Frecuencia relativa acumulada
- NOTAS:
  - El número de filas no es el tamaño de la población (N) o de la muestra (n), sino el número de valores diferentes de la variable en la población o muestra (k).
  - La suma de todas las filas de la columna f es el tamaño de la población (N) o muestra (n)
  - La suma de todas las filas de la columna h es 1

$\boldsymbol{Fila}$	$\boldsymbol{x}$ **ó** $\boldsymbol{X}$	$\boldsymbol{f}$	$\boldsymbol{h}$	$\boldsymbol{F}$	$\boldsymbol{H}$
1	v_1	f_1	h_1	F_1	H_1
2	v_2	f_2	h_2	F_2	H_2
…	…	…	…	…	…
i	v_i	f_i	$h_i = \frac{f_i}{N \text{ ó } n}$	$F_i = \sum_{j=1}^{i} f_j$	$H_i = \frac{F_i}{N \text{ ó } n}$
…	…	…	…	…	…
K	v_k	f_k	h_k	$F_k = N \text{ ó } n$	$H_k = 1$
		$\sum_{i=1}^{k} f_i = N \text{ ó } n$	$\sum_{i=1}^{k} h_i = 1$		

4. Tabla de frecuencias
Ejemplo 8.12 (tiempo de CPU)

- Para evaluar la efectividad de un procesador para un determinado tipo de tareas, registramos el tiempo de CPU para una muestra de 30 trabajos elegidos aleatoriamente (en segundos)
  - $S = (9, 15, 19, 22, 24, 25, 30, 34, 35, 35, 36, 36, 37, 38, 42, 43, 46, 48, 54, 55, 56, 56, 59, 62, 69, 70, 82, 82, 89, 139)$
  - $n = 30$
- Calcular la tabla de frecuencias (no está en el libro)

4. Tabla de frecuencias
Ejemplo 8.12 (solución)

- $S = (9, 15, 19, 22, 24, 25, 30, 34, 35, 35, 36, 36, 37, 38, 42, 43, 46, 48, 54, 55, 56, 56, 59, 62, 69, 70, 82, 82, 89, 139)$
- $n = 30$ (No es el número de filas → tiene 26 filas porque hay 26 valores diferentes en la muestra)
- X = Variable estadística "tiempo de CPU en segundos consumido por un trabajo de la muestra"

$\boldsymbol{Fila}$	$\boldsymbol{X}$	$\boldsymbol{f}$	$\boldsymbol{h}$	$\boldsymbol{F}$	$\boldsymbol{H}$
1	9s	1	$1/30 = 0.03$	1	$1/30 = 0.03$
…	…	…	…	…	…
9	35s	2	$2/30 = 0.07$	10	$10/30 = 0.33$
10	36s	2	$2/30 = 0.07$	12	$12/30 = 0.4$
…	…	…	…	…	…
24	82s	2	$2/30 = 0.07$	28	$28/30 = 0.93$
25	89s	1	$1/30 = 0.03$	29	$29/30 = 0.97$
26	139s	1	$1/30 = 0.03$	30	$30/30 = 1$
		$Total = 30$	$Total = 1$		

4. Tabla de frecuencias
Ejemplo con una variable cualitativa

- Variable cualitativa:
  - X = Tipo de sistema operativo del ordenador de una persona en una muestra de 10 personas: Windows, Mac o Linux
  - S = (Windows, Windows, Mac, Linux, Windows, Mac, Mac, Windows, Linux, Windows)
- Tabla de frecuencias

$\boldsymbol{Fila}$	$\boldsymbol{X}$	$\boldsymbol{f}$	$\boldsymbol{h}$	$\boldsymbol{F}$	$\boldsymbol{H}$
1	Linux	2	$2/10 = 0.2$	2	$2/10 = 0.2$
2	Mac	3	$3/10 = 0.3$	5	$5/10 = 0.5$
3	Windows	5	$5/10 = 0.5$	10	$10/10 = 1$
		$Total = 10$	$Total = 1$		

4. Tabla de frecuencias
Tabla de frecuencias para datos agrupados

- $\boldsymbol{Fila:}$ Número de fila de la tabla, desde la fila 1 hasta la fila k
- $\boldsymbol{x}$ **ó** $\boldsymbol{X}$: Intervalos iguales (clases) en los que se ha decidido dividir el rango de valores que tiene la variable estadística en la población (x) o muestra (X)
- $\boldsymbol{Marca}$ $(\boldsymbol{de\ clase})$: Representante del intervalo (punto medio)
- $\boldsymbol{f}$: Frecuencia absoluta: número de valores de la población o muestra que pertenecen a cada intervalo
- $\boldsymbol{h}$: Frecuencia relativa
- $\boldsymbol{F}$: Frecuencia absoluta acumulada
- $\boldsymbol{H}$: Frecuencia relativa acumulada
- NOTAS:
  - El número de filas no es el tamaño de la población (N) o de la muestra (n), sino el número de intervalos o clases (k) en los que se ha decidido dividir el rango de la población o muestra
  - La suma de todas las filas de la columna f es el tamaño de la población (N) o muestra (n)
  - La suma de todas las filas de la columna h es 1
  - Cada intervalo está abierto a la derecha excepto el último, para contener el valor más extremo de la población o muestra
  - Los intervalos suelen denominarse clases o contenedores (*bins*)
  - $a_i = b_{i-1}$

$\boldsymbol{Fila}$	$\boldsymbol{x}$ **ó** $\boldsymbol{X}$	$\boldsymbol{Marca}$	$\boldsymbol{f}$	$\boldsymbol{h}$	$\boldsymbol{F}$	$\boldsymbol{H}$
1	$[a_1, b_1)$	m_1	f_1	h_1	F_1	H_1
2	$[a_2, b_2)$	m_1	f_2	h_2	F_2	H_2
…	…	…	…	…	…	…
i	$[a_i, b_i)$ $a_i = b_{i-1}$	$m_i = \frac{a_i + b_i}{2}$	f_i	$h_i = \frac{f_i}{N \text{ ó } n}$	$F_i = \sum_{j=1}^{i} f_j$	$H_i = \frac{F_i}{N \text{ ó } n}$
…	…	…	…	…	…	…
k	$[a_k, b_k]$	m_k	f_k	h_k	F_k	H_k
			$\sum_{i=1}^{k} f_i = N$ ó n	$\sum_{i=1}^{k} h_i = 1$		

4. Tabla de frecuencias
Datos agrupados. Ejemplo 8.12 (tiempo de CPU)

- Para evaluar la efectividad de un procesador para un determinado tipo de tareas, registramos el tiempo de CPU para una muestra de 30 trabajos elegidos aleatoriamente (en segundos)
  - $S = (9, 15, 19, 22, 24, 25, 30, 34, 35, 35, 36, 36, 37, 38, 42, 43, 46, 48, 54, 55, 56, 56, 59, 62, 69, 70, 82, 82, 89, 139)$
  - $n = 30$
- Calcular la tabla de frecuencias con datos agrupados en 14 intervalos de 10 segundos (no está en el libro)

4. Tabla de frecuencias
Datos agrupados. Ejemplo 8.12 (solución)

- $S = (9, 15, 19, 22, 24, 25, 30, 34, 35, 35, 36, 36, 37, 38, 42, 43, 46, 48, 54, 55, 56, 56, 59, 62, 69, 70, 82, 82, 89, 139)$
- $n = 30$ (No es el número de filas → tiene 14 filas porque se ha dividido en 14 intervalos de 10s cada uno
- X = Variable estadística "tiempo de CPU en segundos consumido por un trabajo de la muestra"

$\boldsymbol{Fila}$	$\boldsymbol{X}$	$\boldsymbol{Marca}$	$\boldsymbol{f}$	$\boldsymbol{h}$	$\boldsymbol{F}$	$\boldsymbol{H}$
1	$[0, 10)$	5	1	$1/30 = 0.03$	1	$1/30 = 0.03$
2	$[10, 20)$	15	2	$2/30 = 0.07$	3	$3/30 = 0.1$
3	$[20, 30)$	25	3	$3/30 = 0.1$	6	$6/30 = 0.2$
3	$[30, 40)$	35	8	$8/30 = 0.27$	14	$14/30 = 0.47$
…	…	…	…	…	…	…
13	$[120, 130)$	125	0	$0/30 = 0$	29	$29/30 = 0.97$
14	$[130, 140]$	135	1	$1/30 = 0.03$	30	$30/30 = 1$
			$Total = 30$	$Total = 1$		

4. Tabla de frecuencias
Datos agrupados: Tamaño del intervalo

- Existen diferentes recomendaciones de expertos para decidir el tamaño óptimo de los intervalos
- Regla de Sturges (la más usada): Número de intervalos k para una muestra de n datos
  - $k = \lceil 1 + \log_2 n \rceil = \lceil 1 + 3{,}322 \log_{10} n \rceil$ *(Redondear por exceso)*
  - Ejemplos:
    - Para $n = 10$ → $k = 5$
    - Para $n = 30$ → $k = 6$
    - Para $n = 50$ → $k = 7$
    - Para $n = 100$ → $k = 8$
- Existen otras reglas, como las propuestas por Scott o Freedman-Diaconis

4. Tabla de frecuencias
Cálculo de medidas estadísticas

- Las medidas estadísticas pueden calcularse a partir de una tabla de frecuencias

Medida	Poblacional	Muestral
Media (aritmética)	$\mu = \frac{\sum_{i=1}^{k} f_i v_i}{N}$	$\bar{X} = \frac{\sum_{i=1}^{k} f_i v_i}{n}$
Varianza	$\sigma^2 = \frac{\sum_{i=1}^{k} f_i (v_i - \mu)^2}{N}$	$s^2 = \frac{\sum_{i=1}^{k} f_i (v_i - \bar{X})^2}{n-1}$

$\boldsymbol{Fila}$	$\boldsymbol{x}$ o $\boldsymbol{X}$	$\boldsymbol{f}$
1	v_1	f_1
2	v_2	f_2
…	…	…
i	v_i	f_i
…	…	…
k	v_k	f_k

4. Tabla de frecuencias
Cálculo de medidas estadísticas. Datos agrupados

- En el caso de tablas de frecuencias con datos agrupados hay que elegir un valor que represente a cada intervalo, llamado “marca de clase”
  - Suele ser el punto medio del intervalo:
    - $m_i = \frac{a_i + b_i}{2}$

Medida	Poblacional	Muestral
Media (aritmética)	$\mu = \frac{\sum_{i=1}^{k} f_i m_i}{N}$	$\bar{X} = \frac{\sum_{i=1}^{k} f_i m_i}{n}$
Varianza	$\sigma^2 = \frac{\sum_{i=1}^{k} f_i (m_i - \mu)^2}{N}$	$s^2 = \frac{\sum_{i=1}^{k} f_i (m_i - \bar{X})^2}{n-1}$

$\boldsymbol{Fila}$	$\boldsymbol{x}$ ó $\boldsymbol{X}$	$\boldsymbol{Marca}$	$\boldsymbol{f}$
1	$[a_1, b_1)$	m_1	f_1
2	$[a_2, b_2)$	m_2	f_2
…	…	…	…
i	$[a_i, b_i)$	m_i	f_i
…	…	…	…
k	$[a_k, b_k]$	m_k	f_k

4. Tabla de frecuencias
Cálculo de medidas estadísticas. Ejemplo

- Ejemplo 8.12
- Media con datos sin agrupar
  - $\bar{X} = \frac{\sum_{i=1}^{26} f_i v_i}{30} = \frac{1 \cdot 9 + \cdots + 1 \cdot 139}{30} = 48.23\ seg$
- Varianza y desv. estándar sin agrupar
  - $s^2 = \frac{\sum_{i=1}^{26} f_i (v_i - 48.23)^2}{30-1} = 703.15\ seg^2$
  - $s = \sqrt{s^2} = \sqrt{703.15} = 26.52\ seg$
- Media con datos agrupados
  - $\bar{X} = \frac{\sum_{i=1}^{14} f_i m_i}{30} = \frac{1 \cdot 5 + \cdots + 1 \cdot 135}{30} = 48\ seg$
- Varianza y desv. con datos agrupados
  - $s^2 = \frac{\sum_{i=1}^{14} f_i (m_i - 48)^2}{30-1} = 697.59\ seg^2$
  - $s = \sqrt{s^2} = \sqrt{697.59} = 26.41\ seg$

$\boldsymbol{Fila}$	$\boldsymbol{X}$	$\boldsymbol{f}$
1	9	1
…	…	…
9	35	2
10	36	2
…	…	…
24	82	2
25	89	1
26	139	1
		$Total = 30$

$\boldsymbol{fila}$	$\boldsymbol{X}$	$\boldsymbol{Marca}$ $(\boldsymbol{m})$	$\boldsymbol{f}$
1	$[0, 10)$	5	1
2	$[10, 20)$	15	2
3	$[20, 30)$	25	3
3	$[30, 40)$	35	8
…	…		…
13	$[120, 130)$	125	0
14	$[130, 140]$	135	1
			$Total = 30$

4. Tabla de frecuencias
Ejercicios propuestos

- Ejercicios 8.1, 8.2, 8.8, 8.9 del libro
  - Calcular las tablas de frecuencias y la media y desviación estándar con datos agrupados aplicando la regla de Sturges en cada ejercicio y comparar los resultado con los datos sin agrupar
  - La respuesta de 8.2 es:
    - Datos sin agrupar: $\bar{X} = 17.95\ miles\ de\ usuarios, s = 3.16\ miles\ de\ usuarios$
    - Datos agrupados: $k = 7, \bar{X} = 17.86\ miles\ de\ usuarios, s = 3.09\ miles\ de\ usuarios$
- Otros ejercicios (resueltos): proyectodescartes.org

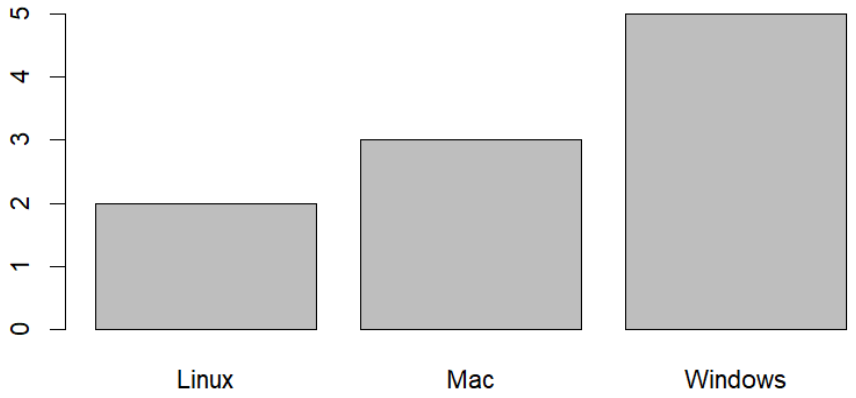
5. Gráficos estadísticos

- Diagrama de barras (no está en el libro)
- Diagrama de tarta (no está en el libro)
- Diagrama de Pareto (no está en el libro)
- Histograma
- Polígono de frecuencias (no está en el libro)
- Diagrama de tallo y hojas (*Stem-and-leaf plot*)
- Diagrama de caja (*Boxplot*)

5. Gráficos estadísticos
Diagrama de barras

- Se utiliza para:
  - variables cualitativas
  - variables cuantitativas que tienen pocos valores diferentes
- Es la representación gráfica de una tabla de frecuencias
- Se dibuja una barra vertical por cada posible valor de la variable
  - La altura de la barra representa la frecuencia del valor

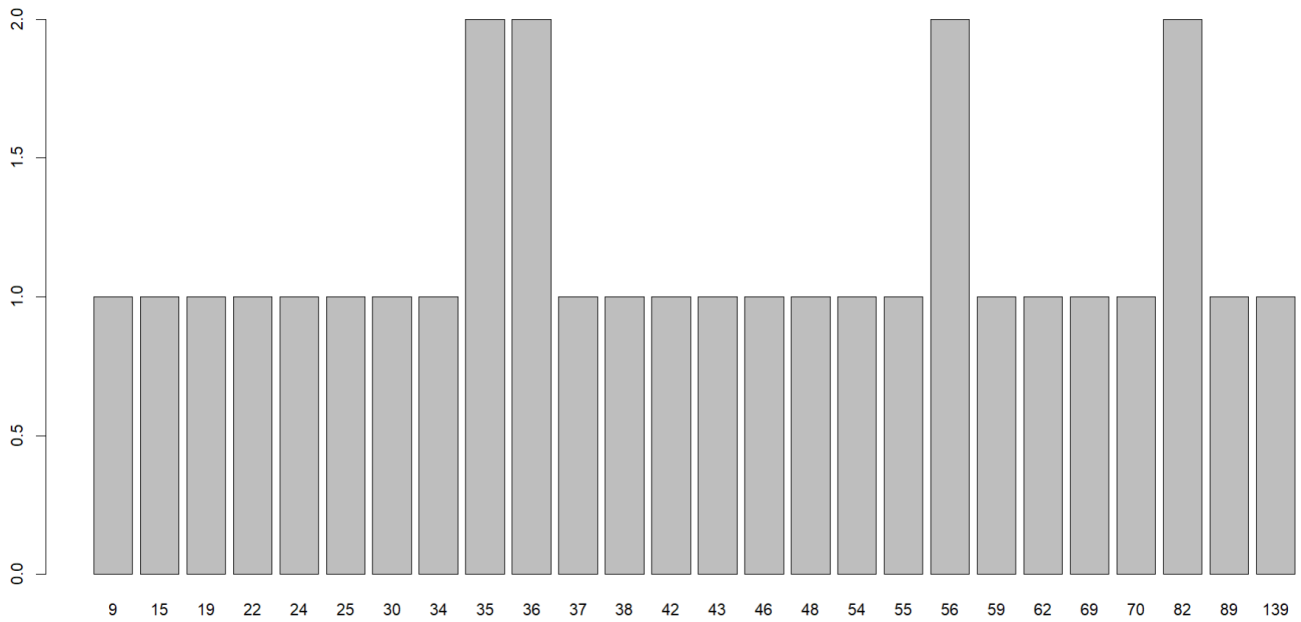
$Fila$	X	f	h
1	Linux	2	$2/10 = 0.2$
2	Mac	3	$3/10 = 0.3$
3	Windows	5	$5/10 = 0.5$
		10	1



5. Gráficos estadísticos
Diagrama de barras. Ejemplo 8.12

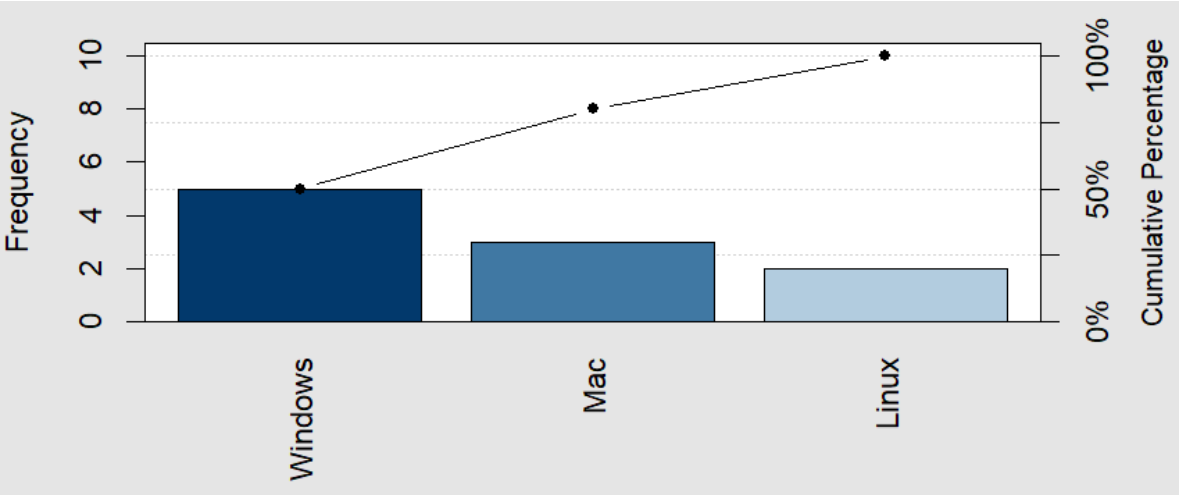
- Para evaluar la efectividad de un procesador para un determinado tipo de tareas, registramos el tiempo de CPU para una muestra de 30 trabajos elegidos aleatoriamente (en segundos)
- S=(9, 15, 19, 22, 24, 25, 30, 34, 35, 35, 36, 36, 37, 38, 42, 43, 46, 48, 54, 55, 56, 56, 59, 62, 69, 70, 82, 82, 89, 139)
- Dibujar el diagrama de barras de frecuencias absolutas

Fila	***X***	***f***
1	9	1
2	15	1
…	…	…
24	82	2
25	89	1
26	139	1
		30



5. Gráficos estadísticos
Diagrama de Pareto

- Es un diagrama de barras en orden descendente según frecuencias absolutas, en el que se superpone una línea con las frecuencias acumuladas
- Se suele utilizar en la gestión empresarial para clasificar gráficamente la información de mayor a menor relevancia, con el objetivo de reconocer los problemas más importantes

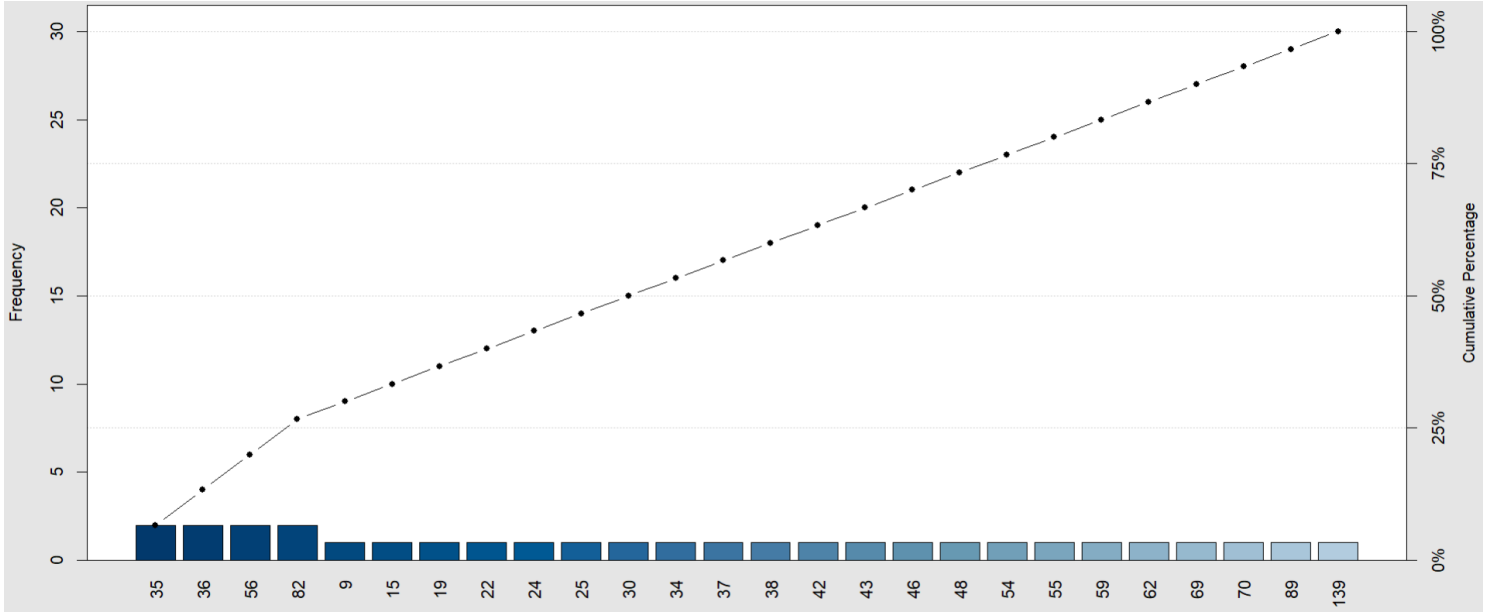


Fila	*X*	*f*	*F*
1	Linux	2	2
2	Mac	3	5
3	Windows	5	10

5. Gráficos estadísticos
Diagrama de Pareto. Ejemplo 8.12

- Para evaluar la efectividad de un procesador para un determinado tipo de tareas, registramos el tiempo de CPU para una muestra de 30 trabajos elegidos aleatoriamente (en segundos)
- S=(9, 15, 19, 22, 24, 25, 30, 34, 35, 35, 36, 36, 37, 38, 42, 43, 46, 48, 54, 55, 56, 56, 59, 62, 69, 70, 82, 82, 89, 139)
- Dibujar el diagrama de Pareto

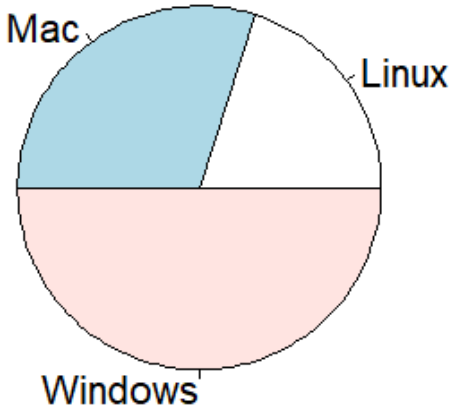
Fila	*X*	*f*	*F*
1	9	1	1
2	15	1	2
…	…	…	…
24	82	2	28
25	89	1	29
26	139	1	30



5. Gráficos estadísticos
Diagrama de tarta

- Se utiliza para:
  - variables cualitativas
  - variables cuantitativas que tienen pocos valores diferentes
- Es la representación gráfica de una tabla de frecuencias
- Se dibuja una porción (sector) de un círculo por cada posible valor de la variable
  - El área de la porción es proporcional a representa la frecuencia del valor

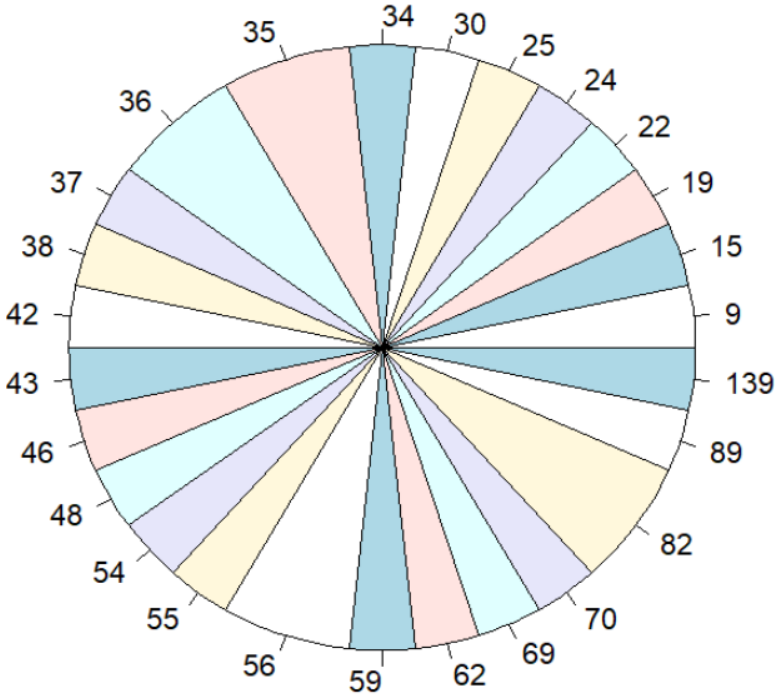
$Fila$	X	f	h
1	Linux	2	$2/10 = 0.2$
2	Mac	3	$3/10 = 0.3$
3	Windows	5	$5/10 = 0.5$
		10	1



5. Gráficos estadísticos
Diagrama de tarta. Ejemplo 8.12

- Para evaluar la efectividad de un procesador para un determinado tipo de tareas, registramos el tiempo de CPU para una muestra de 30 trabajos elegidos aleatoriamente (en segundos)
- S=(9, 15, 19, 22, 24, 25, 30, 34, 35, 35, 36, 36, 37, 38, 42, 43, 46, 48, 54, 55, 56, 56, 59, 62, 69, 70, 82, 82, 89, 139)
- Dibujar un diagrama de tarta

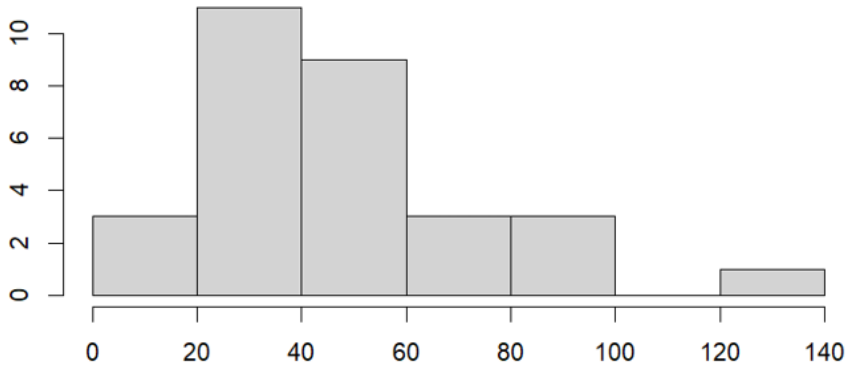
Fila	*X*	*f*
1	9	1
2	15	1
…	…	…
24	82	2
25	89	1
26	139	1
		30



5. Gráficos estadísticos
Histograma

- Se utiliza para variables cuantitativas
- Es la representación gráfica de una tabla de frecuencias para datos agrupados
- Las frecuencias se representan como barras verticales unidas
- En el eje horizontal se representan los intervalos como base de cada barra
- En el eje vertical se representan las frecuencias, como altura de cada barra
- Tipos
  - Histograma de frecuencias (absolutas)
  - Histograma de frecuencias relativas

$Fila$	X	f	h
1	$[0, 20)$	3	$3/30 = 0.10$
2	$[20, 40)$	11	$11/30 = 0.37$
3	$[40, 60)$	9	$9/30 = 0.30$
4	$[60, 80)$	3	$3/30 = 0.10$
5	$[80, 100)$	3	$3/30 = 0.10$
6	$[100, 120)$	0	$0/30 = 0$
7	$[120,140]$	1	$1/30 = 0.03$
		30	1



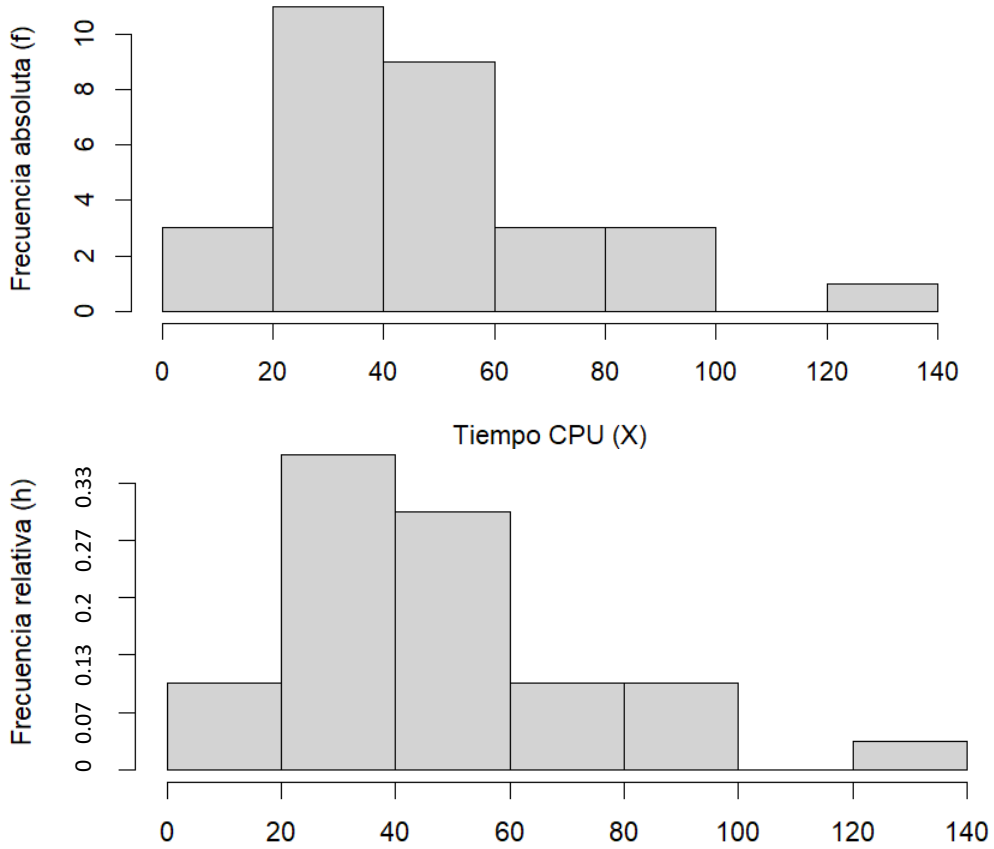
5. Gráficos estadísticos
Histograma. Ejemplo 8.12

- Para evaluar la efectividad de un procesador para un determinado tipo de tareas, registramos el tiempo de CPU para una muestra de 30 trabajos elegidos aleatoriamente (en segundos)
  - $S = (9, 15, 19, 22, 24, 25, 30, 34, 35, 35, 36, 36, 37, 38, 42, 43, 46, 48, 54, 55, 56, 56, 59, 62, 69, 70, 82, 82, 89, 139)$
- a) Dibujar los histogramas de frecuencias absolutas y relativas con 7 intervalos (no está en el libro)
- b) Dibujar los histogramas de frecuencias absolutas y relativas con 14 intervalos

5. Gráficos estadísticos
Histograma. Ejemplo 8.12 (solución)(a)

a) Con 7 intervalos ($k = 7$)

- S=(9, 15, 19, 22, 24, 25, 30, 34, 35, 35, 36, 36, 37, 38, 42, 43, 46, 48 54, 55, 56, 56, 59, 62, 69, 70, 82, 82, 89, 139)

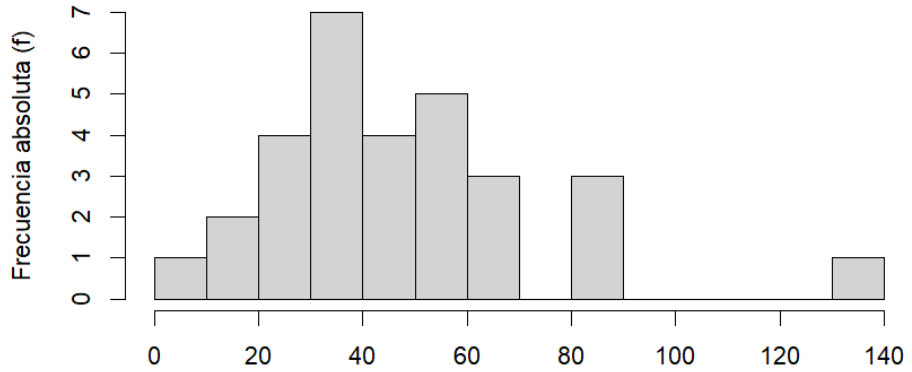


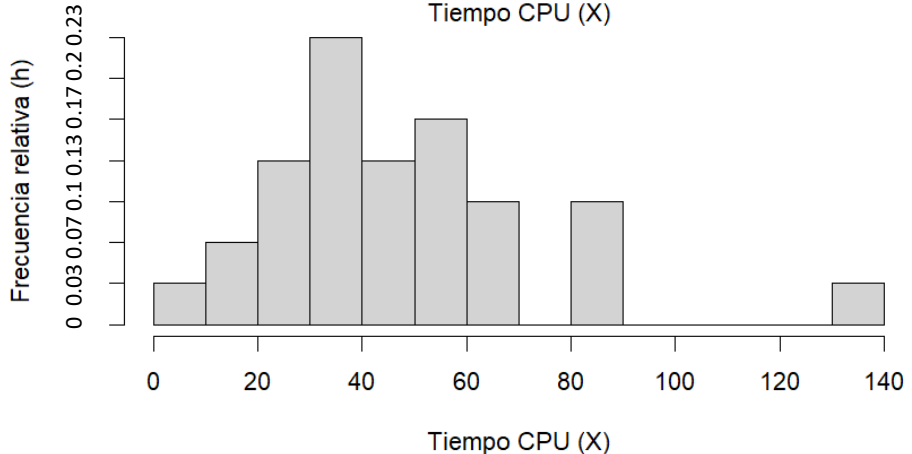
Fila	***X***	***f***	***h***
1	$[0, 20)$	3	$3/30 = 0.10$
2	$[20, 40)$	11	$11/30 = 0.37$
3	$[40, 60)$	9	$9/30 = 0.30$
4	$[60, 80)$	3	$3/30 = 0.10$
5	$[80, 100)$	3	$3/30 = 0.10$
6	$[100, 120)$	0	$0/30 = 0$
7	$[120,140]$	1	$1/30 = 0.03$
		30	1

5. Gráficos estadísticos
Histograma. Ejemplo 8.12 (solución)(b)

b) Con 14 intervalos ($k = 14$)

- S=(9, 15, 19, 22, 24, 25, 30, 34, 35, 35, 36, 36, 37, 38, 42, 43, 46, 48 54, 55, 56, 56, 59, 62, 69, 70, 82, 82, 89, 139)

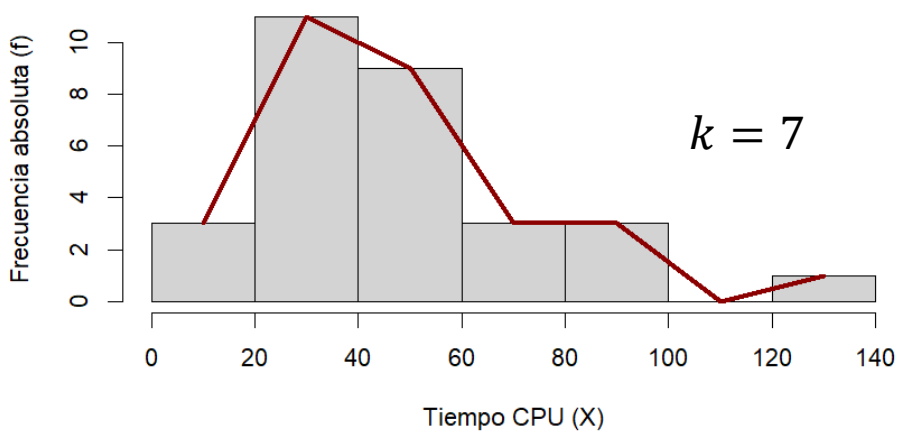


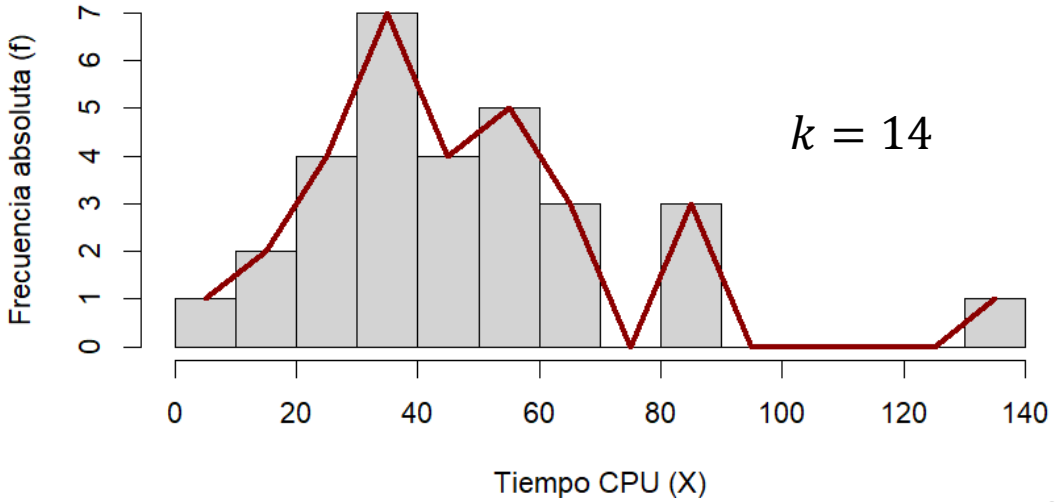


Fila	*X*	*f*	*h*
1	$[0, 10)$	1	$1/30 = 0.03$
2	$[10, 20)$	2	$2/30 = 0.07$
3	$[20, 30)$	3	$3/30 = 0.10$
4	$[30, 40)$	8	$8/30 = 0.27$
5	$[40, 50)$	4	$4/30 = 0.13$
6	$[50, 60)$	5	$5/30 = 0.17$
7	$[60,70]$	2	$2/30 = 0.07$
8	$[70, 80)$	1	$3/30 = 0.03$
9	$[80, 90)$	3	$3/30 = 0.10$
10	$[90, 100)$	0	$0/30 = 0$
11	$[100, 110)$	0	$0/30 = 0$
12	$[110, 120)$	0	$0/30 = 0$
13	$[120, 130)$	0	$0/30 = 0$
14	$[130,140]$	1	$1/30 = 0.03$
		30	1

5. Gráficos estadísticos
Polígono de frecuencias

- Un polígono de frecuencias se dibuja con líneas que unen los puntos medios de la parte superior de las barras de un histograma
- Ejemplo 8.12
  - S=(9, 15, 19, 22, 24, 25, 30, 34, 35, 35, 36, 36, 37, 38, 42, 43, 46, 48 54, 55, 56, 56, 59, 62, 69, 70, 82, 82, 89, 139)





5. Gráficos estadísticos
Diagrama de tallo y hojas (*Stem-and-leaf plot*)

- Se utiliza para variables estadísticas cuantitativas
- Se agrupan los valores por intervalos en filas
  - Ejemplo: primera fila $[0,9]$, segunda fila $[10,19]$, tercera fila $[20,29]$, …
- Se dibuja una barra vertical
  - La parte izquierda es el "tallo"
  - La parte derecha son las "hojas"
- En cada línea horizontal (fila) se escribe:
  - a la derecha de la barra el primer dígito de cada uno de los valores que pertenece al intervalo
  - a la izquierda, el resto de los dígitos de cada valor
- Ejemplo 8.12
  - S=(9, 15, 19, 22, 24, 25, 30, 34, 35, 35, 36, 36, 37, 38, 42, 43, 46, 48, 54, 55, 56, 56, 59, 62, 69, 70, 82, 82, 89, 139)
  - Se podría dibujar un diagrama con 14 intervalos (filas), del $[0,9]$ al $[130,139]$
  - En la tercera fila se escribirían los valores 22, 24 y 25, que pertenecen al intervalo$[20,29]$. A la derecha de la barra vertical el primer dígito de cada uno esos valores de menor a mayor: 2, 4 y 5, y a la izquierda el dígito restante: 2.

```
 0 | 9
 1 | 59
 2 | 245
 3 | 04556678
 4 | 2368
 5 | 45669
 6 | 29
 7 | 0
 8 | 229
 9 |
10 |
11 |
12 |
13 | 9
```

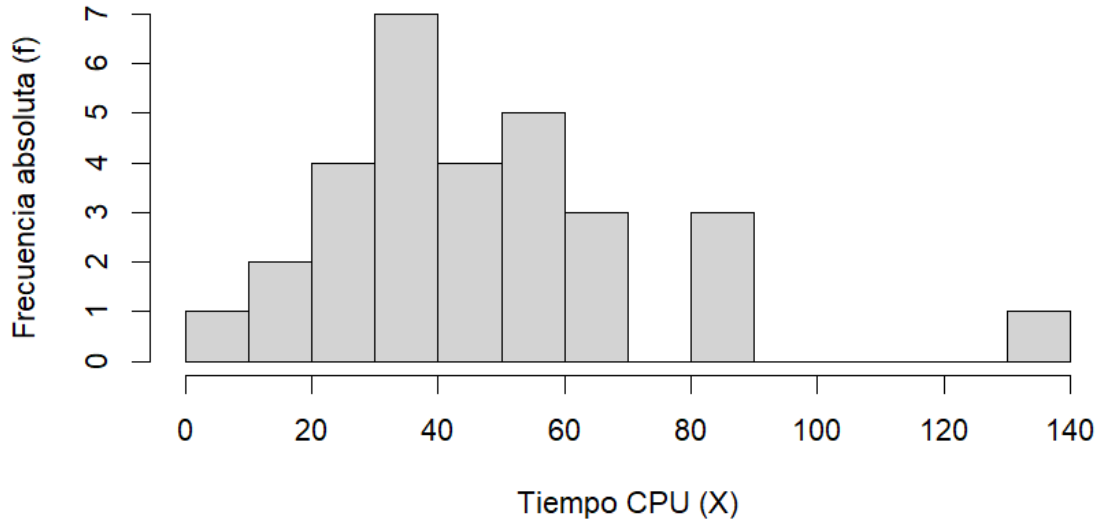
5. Gráficos estadísticos
Diagrama de tallo y hojas. Escala

- Se pueden representar números con decimales multiplicando los dígitos por una escala
- Por defecto la escala es 1, para números enteros
- Para número con decimales se puede usar 0.1, 0.01, 0.001, etc.
- Cada número es: $(10 \cdot tallo + hoja) \cdot escala$
- Ejemplo 8.19:
  - *S=(0.003, 0.004, 0.010, 0.016, 0.019, 0.029, 0.038, 0.046, 0.066, 0.067, 0.071, 0.078)*
  - El primero sería $(10 \cdot 0 + 3) \cdot 0.001 = 0.003$
  - El último sería $(10 \cdot 7 + 8) \cdot 0.001 = 0.078$

0	|	34
1	|	069
2	|	9
3	|	8
4	|	6
5	|	
6	|	67
7	|	18

5. Gráficos estadísticos
Diagrama de tallo y hojas vs Histograma

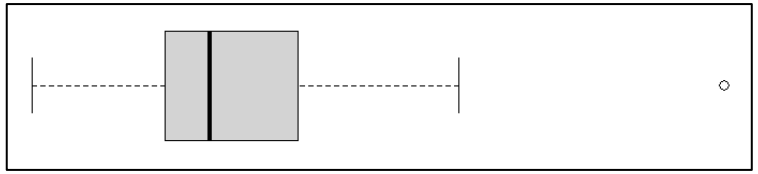
- Si se gira un diagrama de tallo y hojas, se asemeja a un histograma
- Pero el diagrama de tallo y hojas tiene más información → se pueden ver todos los valores, no sólo las frecuencias absolutas




```
  9
  59
  245
  0455667 8
  2368
  45669
  29
  0
  229
  9
0 1 2 3 4 5 6 7 8 9 10 11 12 13
```

5. Gráficos estadísticos
Diagrama de caja (*Boxplot*)

- Es un diagrama para las siguientes medidas de una variable estadística cuantitativa:
  - Mínimo, máximo, primer cuartil, tercer cuartil, mediana, rango intercuartílico
  - También se representan los valores atípicos
  - En algunos casos, también se representa la media



- Se dibuja:
  - una caja entre el primer y tercer cuartil
  - una línea continua dentro de la caja que representa la mediana
  - unos "bigotes" (*whiskers*) que llegan hasta el primer valor y último valor (derecha) que no son datos atípicos, aplicando la regla $1.5(IQR)$
  - unos puntos que representan los valores atípicos
  - En algunos casos, una línea discontinua o un punto (o cruz) dentro de la caja que representa la media
- Se puede representar en vertical u horizontal

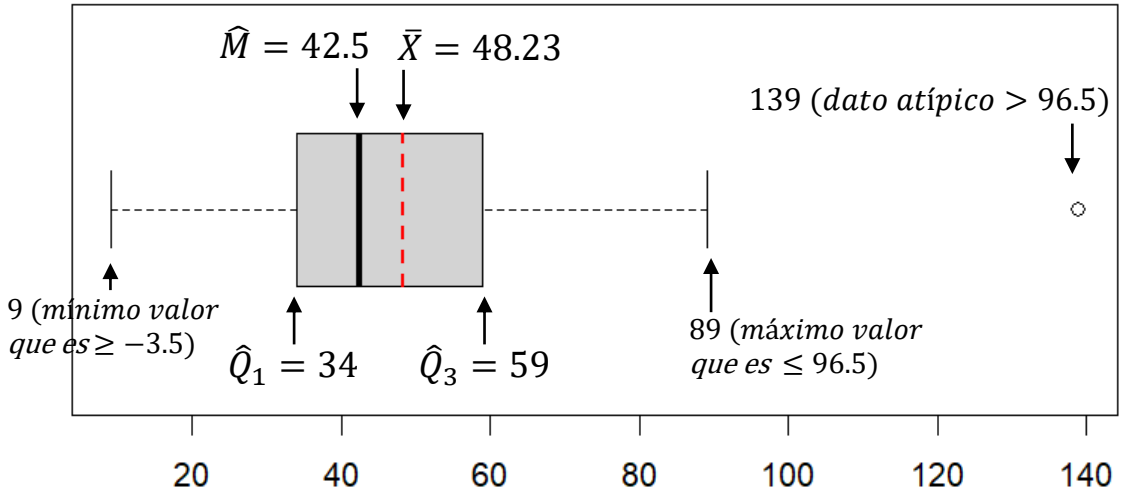
5. Gráficos estadísticos
Diagrama de caja. Ejemplo 8.12

- Para evaluar la efectividad de un procesador para un determinado tipo de tareas, registramos el tiempo de CPU para una muestra de 30 trabajos elegidos aleatoriamente (en segundos)
  - $S = (9, 15, 19, 22, 24, 25, 30, 34, 35, 35, 36, 36, 37, 38, 42, 43, 46, 48, 54, 55, 56, 56, 59, 62, 69, 70, 82, 82, 89, 139)$
- a) Dibujar el diagrama de caja en horizontal (incluyendo la media)
- b) Dibujar el diagrama de caja en vertical (incluyendo la media)

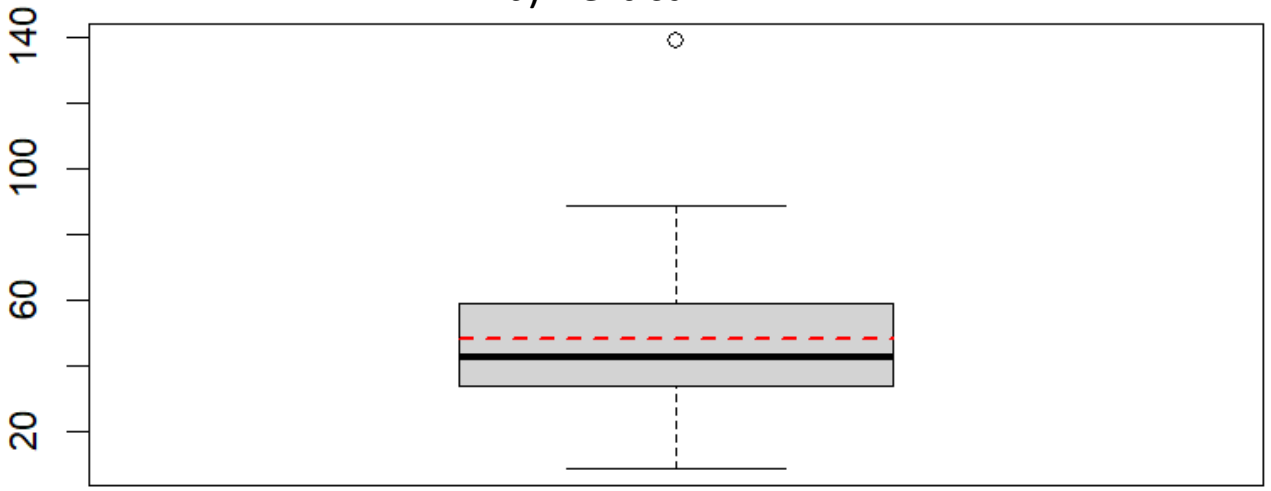
5. Gráficos estadísticos
Diagrama de caja. Ejemplo 8.12 (solución)

- *S=(9, 15, 19, 22, 24, 25, 30, 34, 35, 35, 36, 36, 37, 38, 42, 43, 46, 48, 54, 55, 56, 56, 59, 62, 69, 70, 82, 82, 89, 139)*
- $min = 9, max = 139, \bar{X} = 48.23s, \hat{M} = 42.5s, \hat{Q}_1 = 34s, \hat{Q}_3 = 59s, \widehat{IQR} = 25s, \hat{Q}_1 - 1.5 \cdot \widehat{IQR} = -3.5s, \hat{Q}_3 + 1.5 \cdot \widehat{IQR} = 96.5s$

a) Horizontal

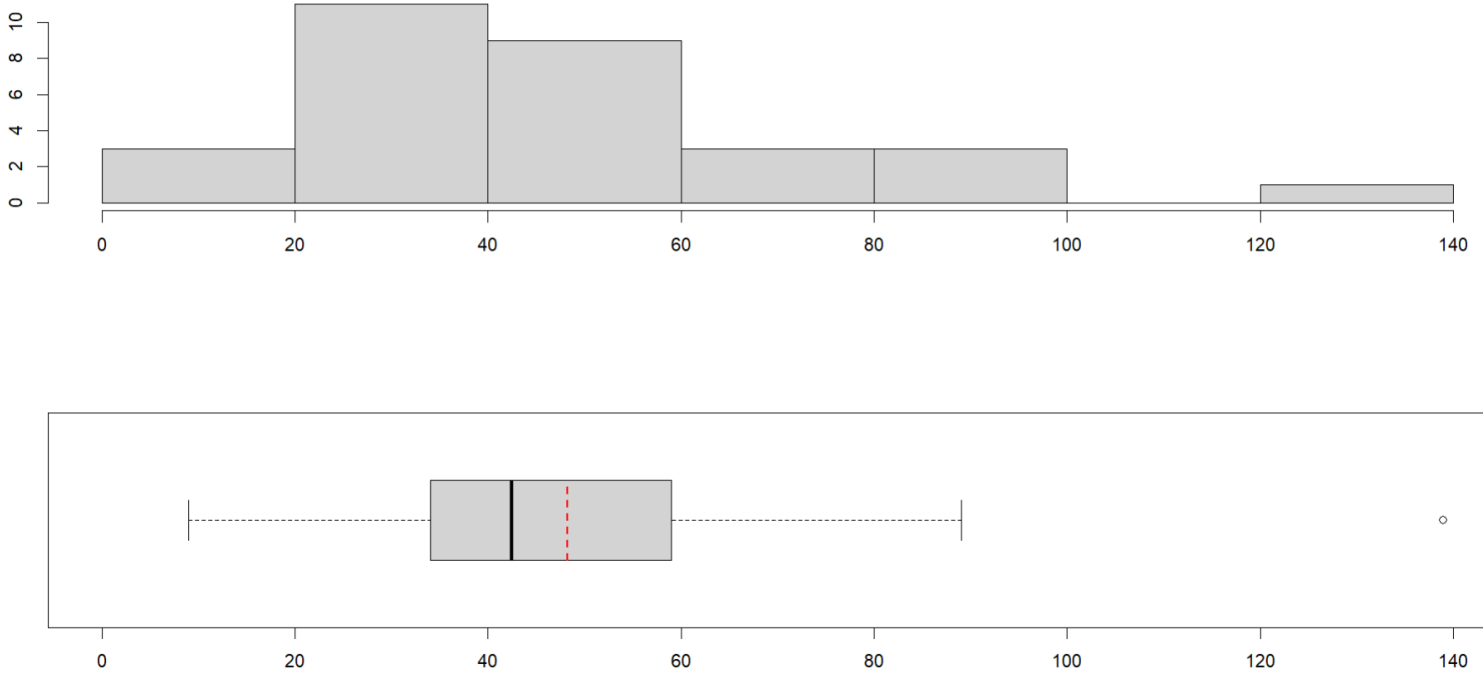


b) Vertical



5. Gráficos estadísticos
Diagrama de caja vs Histograma

- El diagrama de caja y el histograma son complementarios
- Si la mediana está a la izquierda de la media (es menor) en el diagrama de caja, se confirma que hay una asimetría a la derecha en el histograma
- Las columnas del histograma se concentran sobre todo en la parte que abarcan los bigotes en el diagrama de caja
- Si hay columnas alejadas en el histograma, se confirma que hay valores atípicos en el diagrama de caja



5. Gráficos estadísticos
Diagrama de caja vs Histograma. Ejemplo

- ¿Qué diagrama de caja corresponde a cada histograma?

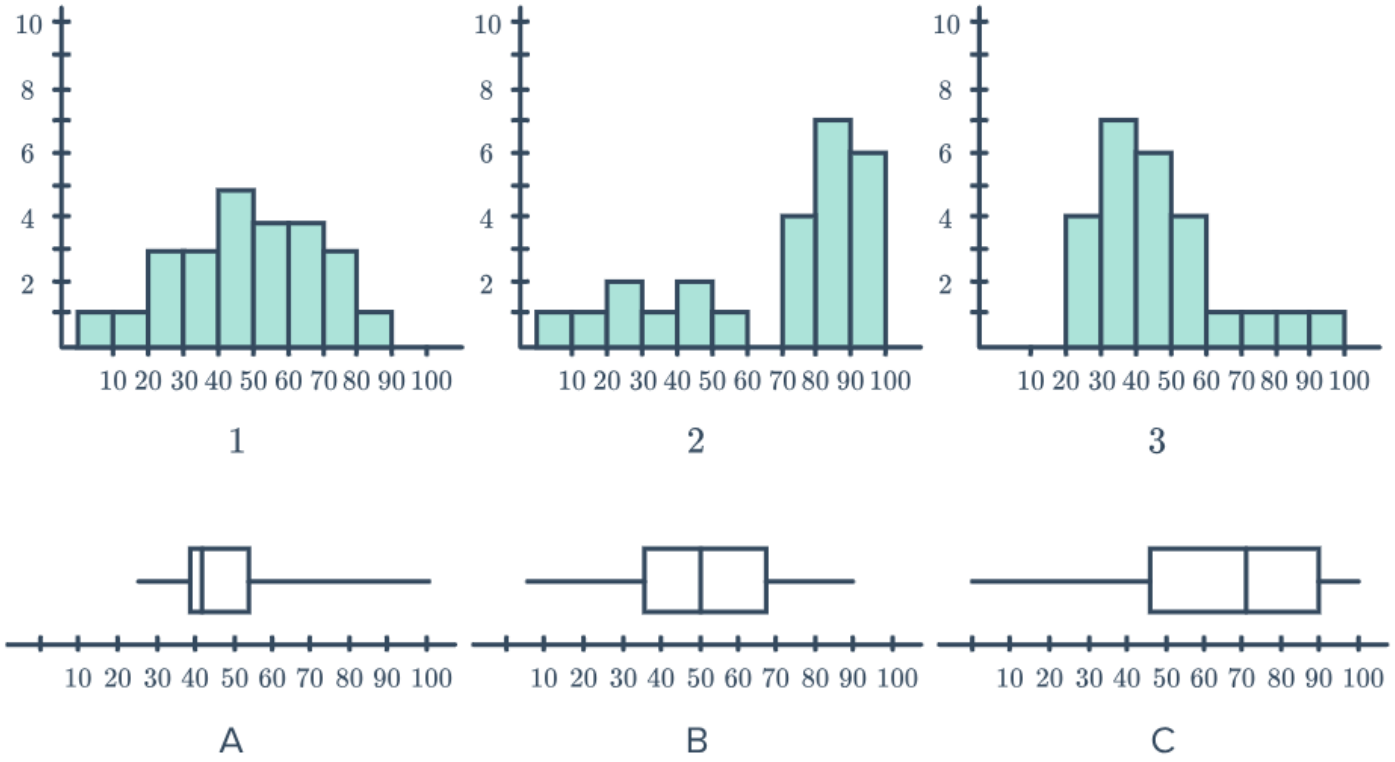
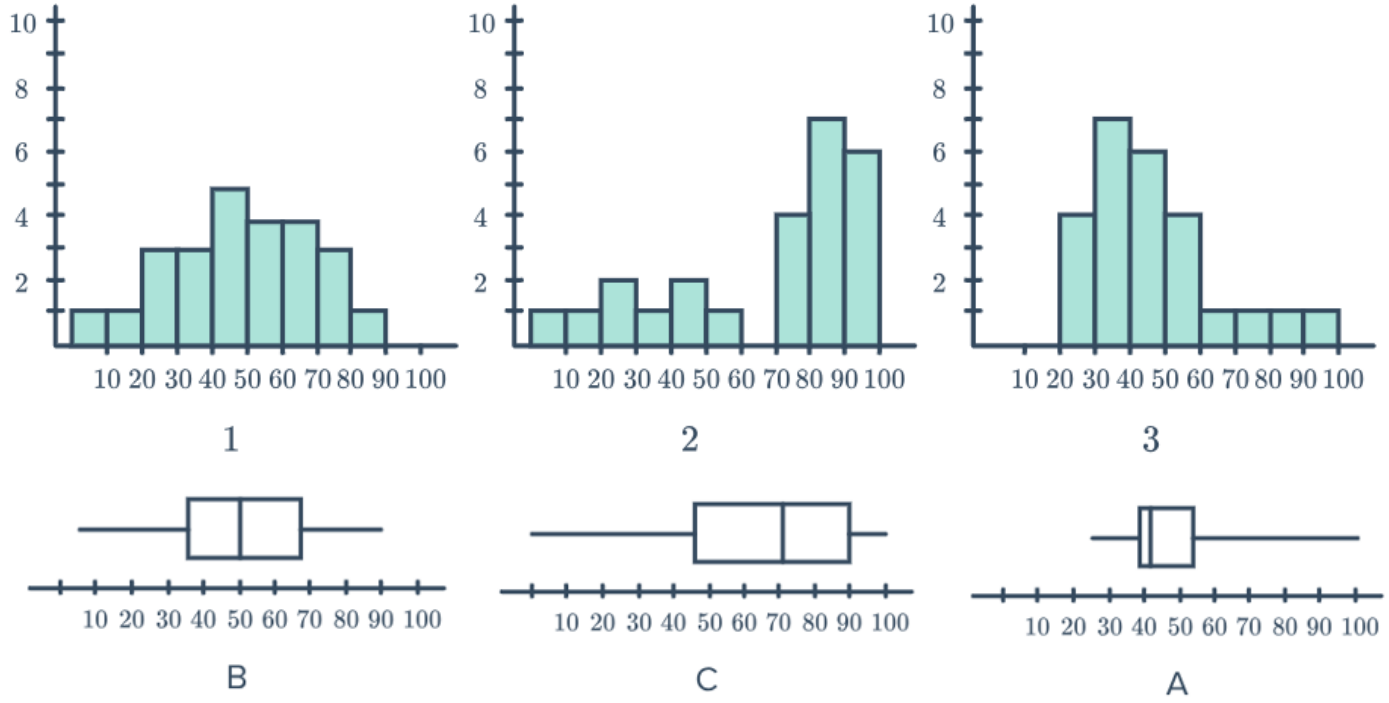


Imagen: mathspace.co

5. Gráficos estadísticos
Diagrama de caja vs Histograma. Ejemplo (sol.)

- El histograma 1 y el boxplot B son casi simétricos
- El histograma 2 es asimétrico a la izquierda y el boxplot C tiene más largo el lado Izquierdo de la caja
- El histograma 3 es asimétrico a la derecha y el boxplot A tiene más largo el lado derecho de la caja



5. Gráficos estadísticos
Diagrama de caja. Varias cajas

- Los diagramas de caja se pueden utilizar para comparar diferentes poblaciones o muestras
- Para ello se dibujan juntos sus diagramas de caja
- Ejemplo: Tráfico diario de Internet de un servidor durante las 52 semanas de un año
  - Siete muestras (días) de 52 valores (semanas)
  - El mayor volumen de tráfico se tiene los viernes
  - Los viernes también tiene la mayor variabilidad
  - El menor volumen de ocurre los fines de semana
  - Cada día hay algunos valores atípicos, excepto los sábados

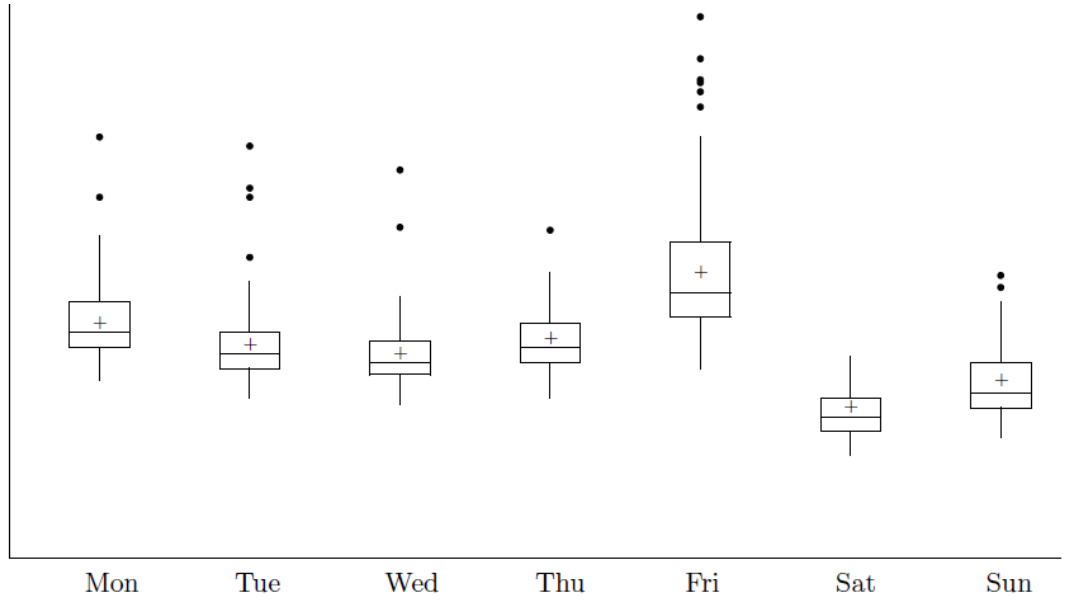


FIGURE 8.10: *Parallel boxplots of internet traffic.*

5. Gráficos estadísticos
Ejercicios propuestos

- Ejercicios 8.1, 8.2, 8.8, 8.9 del libro
  - Dibujar todos los diagramas posibles con los datos de cada ejercicio
- Otros ejercicios (resueltos): proyectodescartes.org

6. Estadística descriptiva bidimensional

- La estadística descriptiva bidimensional describe simultáneamente dos variables estadísticas que representan dos propiedades diferentes de cada individuo de una población o muestra
  - Las dos variables pueden ser cuantitativas y/o cualitativas
- El conjunto de los dos valores de las variables para un individuo se denomina variable estadística bidimensional
- Caso de una población: $P = \big((x_1, y_1), (x_2, y_2), .., (x_N, y_N)\big)$
  - (x_i, y_i) es el valor de la variable bidimensional (x, y) para el individuo i de la población
  - Ejemplo: si x representa el peso e y representa la altura de una persona de una población, entonces (x_i, y_i) es el peso y altura de la persona i del total de N personas de la población
- Caso de una muestra: $S = \big((X_1, Y_1), (X_2, Y_2), .., (X_n, Y_n)\big)$
  - (X_i, Y_i) es el valor de la variable bidimensional (X, Y) para el individuo i de la muestra
  - Ejemplo: si X representa el peso e Y representa la altura de una persona de una muestra, entonces (X_i, Y_i) es el peso y altura de la persona i del total de n personas de la muestra

6. Estadística descriptiva bidimensional Medidas estadísticas. Cálculo

Medida	Poblacional	Muestral
Media (aritmética) marginal de x o X	$\mu_x = \frac{\sum_{i=1}^{N} x_i}{N}$	$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$
Media (aritmética) marginal de y o Y	$\mu_y = \frac{\sum_{i=1}^{N} y_i}{N}$	$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$
Varianza marginal de x o X	$\sigma_x{}^2 = \frac{\sum_{i=1}^{N}(x_i - \mu_x)^2}{N}$	$s_X{}^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$
Varianza marginal de y o Y	$\sigma_y{}^2 = \frac{\sum_{i=1}^{N}(y_i - \mu_y)^2}{N}$	$s_Y{}^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}$
Covarianza	$\sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N}$	$s_{xy} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$
Coeficiente de correlación (de Pearson)	$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$	$r = \frac{s_{xy}}{s_x s_y}$

6. Estadística descriptiva bidimensional
Medidas estadísticas. Coeficiente de correlación (I)

- El coeficiente de correlación (de Pearson) es una medida de dependencia lineal entre dos variables cuantitativas que representan dos propiedades de los elementos o individuos de una población (peso, altura, edad, etc.)
- Cuando se aplica a toda la población, se denomina coeficiente de correlación poblacional (ρ).
- Cuando se aplica a una muestra de la población, se denomina coeficiente de correlación muestral (r).

6. Estadística descriptiva bidimensional Medidas estadísticas. Coeficiente de correlación

- Dada una muestra de pares de valores:
  - $(X_1, Y_1), (X_2, Y_2), \ldots (X_n, Y_n)$ de dos variables X e Y
  - $r =$ coeficiente de correlación muestral $= \frac{s_{XY}}{s_X s_Y}$
- Valores posibles
  - $-1 \leq r \leq 1$
- Valores de r cercanos a:
  - 1 indican una fuerte correlación lineal positiva
  - – 1 muestran una fuerte correlación lineal negativa
  - 0 muestran una correlación débil o ninguna correlación
- $|r| = 1$ es posible sólo cuando todos los valores de X e Y se encuentran en una línea recta
- NOTA: Es el coeficiente de correlación de Pearson, existen otros coeficientes de correlación, como el de Kendall o el de Spearman

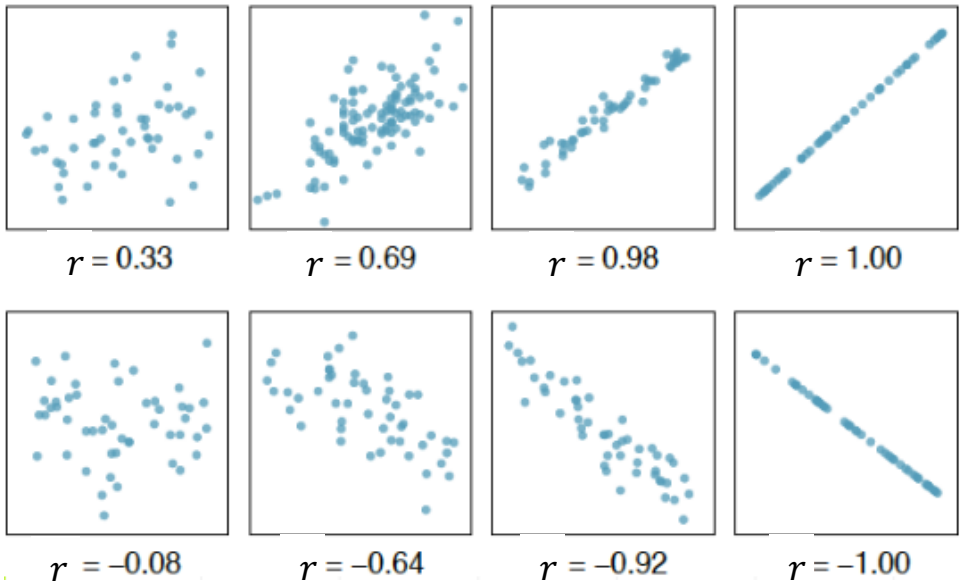


Imagen: RPubs

6. Estadística descriptiva bidimensional
Medidas estadísticas. Coef. de correlación. Ejemplo 8.22

- Según la Base Internacional de Datos de la Oficina del Censo de los Estados Unidos, la población mundial entre 1950 y 2010 creció según esta tabla, en la que:
  - X representa un año
  - Y representa la población mundial en ese año (millones de personas)

X	1950	1955	1960	1965	1970	1975	1980	1985	1990	1995	2000	2005	2010
Y	2558	2782	3043	3350	3712	4089	4451	4855	5287	5700	6090	6474	6864

- Calcular el coeficiente de correlación (de Pearson)

6. Estadística descriptiva bidimensional
Medidas estadísticas. Coef. de correlación. Ejemplo 8.22 (sol.)

X	1950	1955	1960	1965	1970	1975	1980	1985	1990	1995	2000	2005	2010
Y	2558	2782	3043	3350	3712	4089	4451	4855	5287	5700	6090	6474	6864

- $\bar{X} = \frac{\sum_{i=1}^{13} X_i}{13} = \frac{1950+\cdots+2010}{13} = 1980$
- $\bar{Y} = \frac{\sum_{i=1}^{13} Y_i}{13} = \frac{2558+\cdots+6864}{13} = 4558.1$
- $s_X^2 = \frac{\sum_{i=1}^{13}(X_i-1980)^2}{13-1} = \frac{(1950-1980)^2+\cdots+(2010-1980)^2}{12} = 379.17 \rightarrow s_X = \sqrt{379.17} = 19.47$
- $s_Y^2 = \frac{\sum_{i=1}^{13}(Y_i-4558.1)^2}{13-1} = \frac{(2558-4558.1)^2+\cdots+(6864-4558.1)^2}{12} = 2092943.41 \rightarrow s_Y = \sqrt{2092943.41} = 1446.7$
- $s_{XY} = \frac{\sum_{i=1}^{13}(X_i-1980)(Y_i-4558.1)}{13-1} = \frac{(1950-1980)(2558-4558.1)+\cdots+(2010-1980)(6864-4558.1)}{12} = 28104.17$
- $r = \frac{s_{XY}}{s_X s_Y} = \frac{28104.17}{(19.47)(1446.7)} = \mathbf{0.998}$ (Valor muy próximo a 1, hay una gran correlación entre las variables)

6. Estadística descriptiva bidimensional
Tabla de contingencia

- Es la tabla de frecuencias para una variable estadística bidimensional
- También se denomina tabla de frecuencias de doble entrada
- Frecuencia absoluta **marginal** del valor v_i de la variable X:
  - $f_{v_i} = \sum_{j=1}^{l} f_{i,j}$
  - Siendo l el número de valores diferentes de Y
- Frecuencia absoluta **marginal** del valor w_j de la variable Y:
  - $f_{w_j} = \sum_{i=1}^{k} f_{i,j}$
  - Siendo k el número de valores diferentes de X
- Frecuencias relativas **marginales**
  - $h_{v_i} = \frac{f_{v_i}}{N \text{ ó } n}$
  - $h_{w_j} = \frac{f_{w_j}}{N \text{ ó } n}$

	$\boldsymbol{w_1}$	$\boldsymbol{w_2}$	…	$\boldsymbol{w_j}$	…	$\boldsymbol{w_l}$	$\boldsymbol{f_X}$	$\boldsymbol{h_X}$
$\boldsymbol{v_1}$	$f_{1,1}$	$f_{1,2}$	…	$f_{1,j}$	…	$f_{1,l}$	f_{v_1}	h_{v_1}
$\boldsymbol{v_2}$	$f_{2,1}$	$f_{2,2}$	…	$f_{2,j}$	…	$f_{2,l}$	f_{v_2}	h_{v_2}
…	…	…	…	…	…	…	…	…
$\boldsymbol{v_i}$	$f_{i,1}$	$f_{i,2}$	…	$f_{i,j}$	…	$f_{i,l}$	f_{v_i}	h_{v_i}
…	…	…	…	…	…	…	…	…
$\boldsymbol{v_k}$	$f_{k,1}$	$f_{k,2}$	…	$f_{k,j}$	…	$f_{k,l}$	f_{v_k}	h_v
$\boldsymbol{f_Y}$	f_{w_1}	f_{w_2}	…	f_{w_j}	…	f_{w_l}	N ó n	
$\boldsymbol{h_Y}$	h_w	h_w	…	h_{w_j}	…	h_{w_l}		1

6. Estadística descriptiva bidimensional Tabla de contingencia. Ejemplo

- X= Tipo de sistema operativo del móvil de una persona: Android o iPhone
- Y = Tipo de sistema operativo del ordenador de una persona: Windows, Mac o Linux
- Muestra de 10 personas
  - *S = ((Android, Windows), (Android, Windows), (iPhone, Mac), (Android, Linux), (iPhone, Windows), (iPhone, Mac), (iPhone, Mac), (Android, Windows), (Android, Linux), (Android, Windows))*

X/Y	$Windows$	Mac	$Linux$	f_X	h_X
$Android$	4	0	2	6	0.6
$iPhone$	1	3	0	4	0.4
f_Y	5	3	2	10	
h_Y	0.5	0.3	0.2		1

6. Estadística descriptiva bidimensional
Tabla de contingencia. Datos agrupados

- Cuando una de las variables es cuantitativa continua o discreta con muchos valores diferentes, se suele agrupar en intervalos (clases)

	$\boldsymbol{w_1}$	$\boldsymbol{w_2}$	…	$\boldsymbol{w_j}$	…	$\boldsymbol{w_l}$	$\boldsymbol{f_X}$	$\boldsymbol{h_X}$
$\boldsymbol{c_1 = [a_1, b_1)}$	$f_{1,1}$	$f_{1,2}$	…	$f_{1,j}$	…	$f_{1,l}$	f_{c_1}	h_{c_1}
$\boldsymbol{c_2 = [a_2, b_2)}$	$f_{2,1}$	$f_{2,2}$	…	$f_{2,j}$	…	$f_{2,l}$	f_{c_2}	h_{c_2}
…	…	…	…	…	…	…	…	…
$\boldsymbol{c_i = [a_i, b_i)}$	$f_{i,1}$	$f_{i,2}$	…	$f_{i,j}$	…	$f_{i,l}$	f_{c_i}	h_{c_i}
…	…	…	…	…	…	…	…	…
$\boldsymbol{c_k = [a_k, b_k]}$	$f_{k,1}$	$f_{k,2}$	…	$f_{k,j}$	…	$f_{k,l}$	f_{c_k}	h_{c_k}
$\boldsymbol{f_Y}$	f_{w_1}	f_{w_2}	…	f_{w_j}	…	f_{w_l}	N ó n	
$\boldsymbol{h_Y}$	h_{Y_1}	h_{Y_2}	…	h_{Y_j}	…	h_{Y_l}		1

6. Estadística descriptiva bidimensional
Tabla de contingencia. Ejemplo datos agrupados

- X= Edad de una persona, entre 18 y 32 años
- Y = Tipo de sistema operativo del ordenador de una persona: Windows, Mac o Linux
- Muestra de 10 personas
  - *S = ((32, Windows), (20, Windows), (24, Mac), (28, Linux), (23, Windows), (20, Mac), (30, Mac), (18, Windows), (25, Linux), (29, Windows))*

X/Y	$Windows$	Mac	$Linux$	f_X	h_X
$[18, 23)$	2	1	0	3	0.3
$[23, 28)$	1	1	1	3	0.3
$[28, 32]$	2	1	1	4	0.4
f_Y	5	3	2	10	
h_Y	0.5	0.3	0.2		1

6. Estadística descriptiva bidimensional
Tabla de contingencia. Medidas marginales

- Las medidas estadísticas marginales de las dos variables (si son cuantitativas) pueden calcularse a partir de una tabla de frecuencias

Medida	Poblacional	Muestral
Medias	$\mu_x = \frac{\sum_{i=1}^{k} f_{v_i} v_i}{N}$	$\bar{X} = \frac{\sum_{i=1}^{k} f_{v_i} v_i}{n}$
	$\mu_y = \frac{\sum_{i=1}^{l} f_{w_i} w_i}{N}$	$\bar{Y} = \frac{\sum_{i=1}^{l} f_{w_i} w_i}{n}$
Varianzas	${\sigma_x}^2 = \frac{\sum_{i=1}^{k} f_{v_i} (v_i - \mu_x)^2}{N}$	${s_X}^2 = \frac{\sum_{i=1}^{k} f_{v_i} (v_i - \bar{X})^2}{n-1}$
	${\sigma_y}^2 = \frac{\sum_{i=1}^{k} f_{w_i} \left(w_i - \mu_y\right)^2}{N}$	${s_Y}^2 = \frac{\sum_{i=1}^{k} f_{w_i} (w_i - \bar{Y})^2}{n-1}$

	$\boldsymbol{w_1}$	$\boldsymbol{w_2}$	…	$\boldsymbol{w_j}$	…	$\boldsymbol{w_l}$	$\boldsymbol{f_X}$	$\boldsymbol{h_X}$
$\boldsymbol{v_1}$	$f_{1,1}$	$f_{1,2}$	…	$f_{1,j}$	…	$f_{1,l}$	f_{v_1}	h_{v_1}
$\boldsymbol{v_2}$	$f_{2,1}$	$f_{2,2}$	…	$f_{2,j}$	…	$f_{2,l}$	f_{v_2}	h_{v_2}
…	…	…	…	…	…	…	…	…
$\boldsymbol{v_i}$	$f_{i,1}$	$f_{i,2}$	…	$f_{i,j}$	…	$f_{i,l}$	f_{v_i}	h_{v_i}
…	…	…	…	…	…	…	…	…
$\boldsymbol{v_k}$	$f_{k,1}$	$f_{k,2}$	…	$f_{k,j}$	…	$f_{k,l}$	f_{v_k}	h_v
$\boldsymbol{f_Y}$	f_{w_1}	f_{w_2}	…	f_{w_j}	…	f_{w_l}	N ó n	
$\boldsymbol{h_Y}$	h_w	h_w	…	h_{w_j}	…	h_{w_l}		1

6. Estadística descriptiva bidimensional
Tabla de contingencia. Medidas marginales. Ejemplo

- X= Edad de una persona
- Y = Sistema operativo del ordenador de la persona
- n=10
- Como valores de X se utilizan las marcas de cada clase: $20.5, 25.5, 30$
- Edad media de las personas de la muestra
  - $\bar{X} = \frac{(3)(20.5)+(3)(25.5)+(4)(30)}{10} = 25.8\ años$

X/Y	$Windows$	Mac	$Linux$	f_X	h_X
$[18, 23)$ **20.5**	2	1	0	3	0.3
$[23, 28)$ **25.5**	1	1	1	3	0.3
$[28, 32]$ **30**	2	1	1	4	0.4
f_Y	5	3	2	10	
h_Y	0.5	0.3	0.2		1

6. Estadística descriptiva bidimensional
Tabla de contingencia. Medidas condicionadas

- Se pueden calcular medidas estadísticas de una variable (si es cuantitativa) cuando la otra tiene un determinado valor

Medida	Poblacional	Muestral
Media de x (ó X) cuando y (ó Y) = w_j	$\mu_x \mid w_j = \frac{\sum_{i=1}^{k} f_{i,j} v_i}{f_{w_j}}$	$\bar{X} \mid w_j = \frac{\sum_{i=1}^{k} f_{i,j} v_i}{f_{w_j}}$
Media de y (ó Y) cuando x (ó X) = v_i	$\mu_y \mid v_i = \frac{\sum_{j=1}^{l} f_{i,j} w_j}{f_{v_i}}$	$\bar{Y} \mid v_i = \frac{\sum_{j=1}^{l} f_{i,j} w_j}{f_{v_i}}$

	$\boldsymbol{w_1}$	$\boldsymbol{w_2}$	…	$\boldsymbol{w_j}$	…	$\boldsymbol{w_l}$	$\boldsymbol{f_X}$	$\boldsymbol{h_X}$
$\boldsymbol{v_1}$	$f_{1,1}$	$f_{1,2}$	…	$f_{1,j}$	…	$f_{1,l}$	f_{v_1}	h_{v_1}
$\boldsymbol{v_2}$	$f_{2,1}$	$f_{2,2}$	…	$f_{2,j}$	…	$f_{2,l}$	f_{v_2}	h_{v_2}
…	…	…	…	…	…	…	…	…
$\boldsymbol{v_i}$	$f_{i,1}$	$f_{i,2}$	…	$f_{i,j}$	…	$f_{i,l}$	f_{v_i}	h_{v_i}
…	…	…	…	…	…	…	…	…
$\boldsymbol{v_k}$	$f_{k,1}$	$f_{k,2}$	…	$f_{k,j}$	…	$f_{k,l}$	f_{v_k}	h_v
$\boldsymbol{f_Y}$	f_{w_1}	f_{w_2}	…	f_{w_j}	…	f_{w_l}	N ó n	
$\boldsymbol{h_Y}$	h_w	h_w	…	h_{w_j}	…	h_{w_l}		1

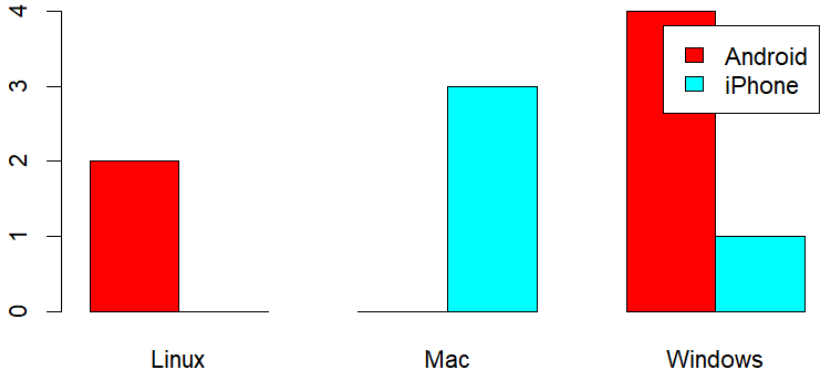
6. Estadística descriptiva bidimensional
Tabla de contingencia. Medidas condicionadas. Ejemplo

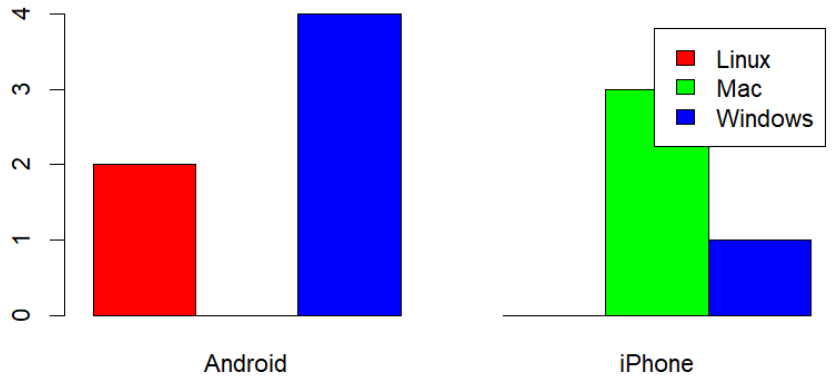
- X= Edad de una persona
- Y = Sistema operativo del ordenador de la persona
- n=10
- Como valores de X se utilizan las marcas de cada clase: $20.5, 25.5, 30$
- Media de la edad de las personas según el sistema operativo del ordenador que usan
  - $\bar{X}|Windows = \frac{(2)(20.5)+(1)(25.5)+(2)(30)}{5} = 25.3\ años$
  - $\bar{X}|Mac = \frac{(1)(20.5)+(1)(25.5)+(1)(30)}{3} = 25.33\ años$
  - $\bar{X}|Linux = \frac{(0)(20.5)+(1)(25.5)+(1)(30)}{2} = 27.75\ años$

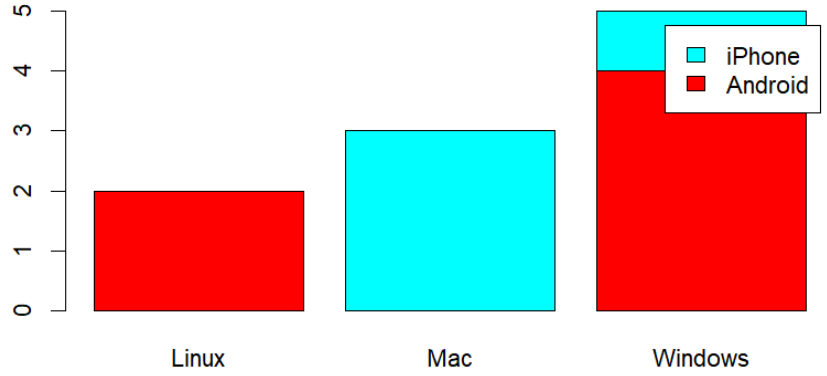
X/Y	$Windows$	Mac	$Linux$	f_X	h_X
$[18, 23)$ **20.5**	2	1	0	3	0.3
$[23, 28)$ **25.5**	1	1	1	3	0.3
$[28, 32]$ **30**	2	1	1	4	0.4
f_Y	5	3	2	10	
h_Y	0.5	0.3	0.2		1

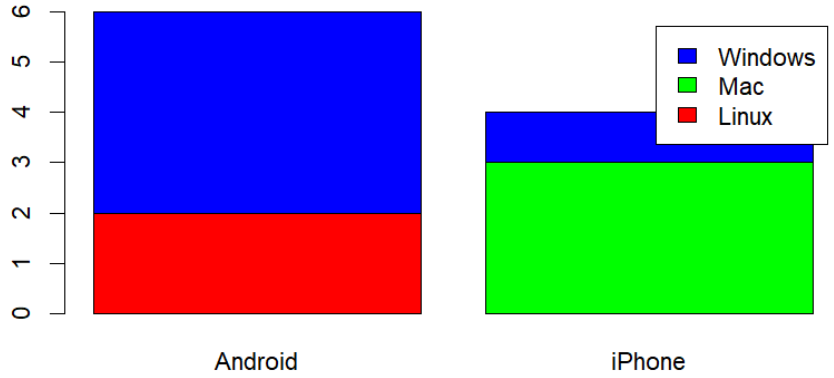
6. Estadística descriptiva bidimensional Gráficos. Diagrama de barras

- En el eje vertical se representan las frecuencias de una variable
- En el eje horizontal, se representan los valores de la otra variable
  - Se pueden agrupar las barras para un valor en vertical o en horizontal









6. Estadística descriptiva bidimensional Gráficos. Diagrama de dispersión

- En inglés *Scatter plot*
- Se representan los puntos correspondientes a las parejas de valores de la variable bidimensional
- El eje horizontal representa los valores de la variable x (si es una población) o X (si es una muestra)
- El eje vertical representa los valores de la variable y (si es una población) o Y (si es una muestra)
- La forma del diagrama depende del valor del coeficiente de correlación ρ (si es una población) o r (si es una muestra)

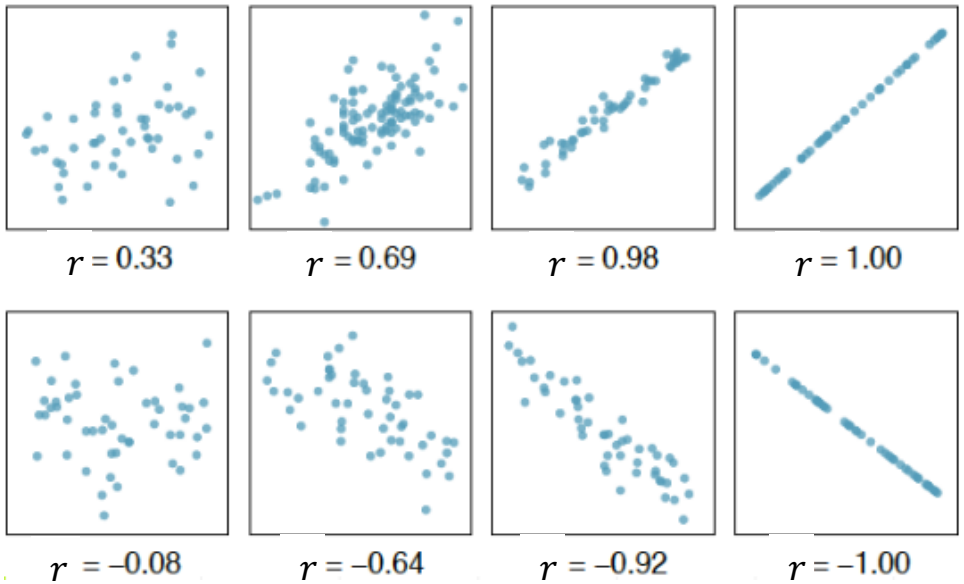


Imagen: RPubs

6. Estadística descriptiva bidimensional
Gráficos. Diagrama de dispersión (Ejemplo 8.20)

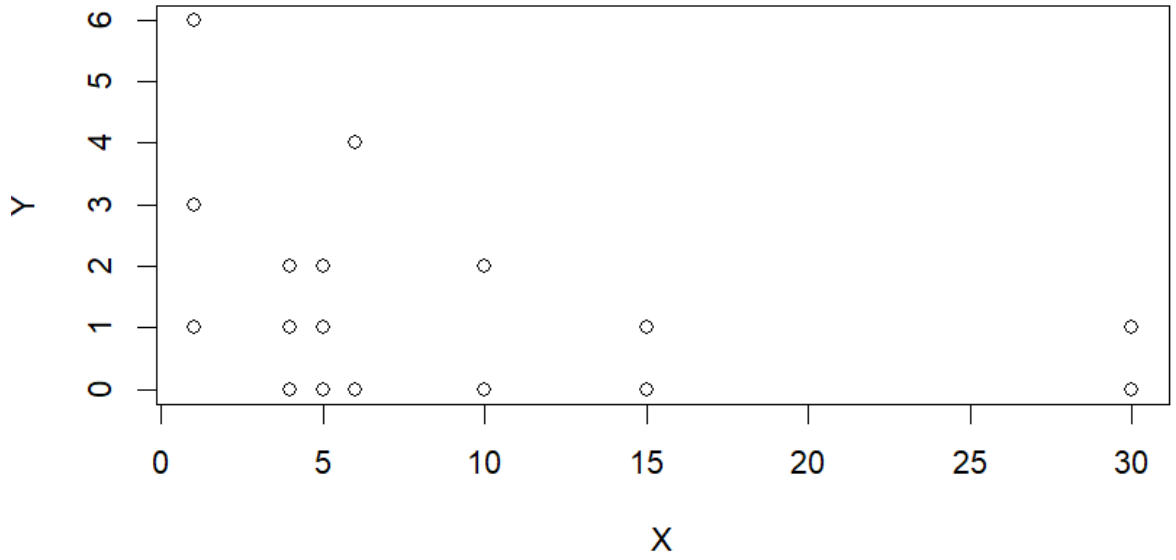
- Un administrador informático registra el número de veces que un software antivirus se ejecuta en cada ordenador de una empresa durante un mes (variable X) y el número de virus detectados (variable Y).
- Los datos de una muestra 30 ordenadores están en la siguiente tabla

X	30	30	30	30	30	30	30	30	30	30	30	15	15	15	10	10	10	6	6	5	5	5	4	4	4	4	4	1	1	1
Y	0	0	1	0	0	0	1	1	0	0	0	0	1	1	0	0	2	0	4	1	2	0	2	1	0	1	0	6	3	1

- ¿Existe una relación entre la frecuencia de ejecución del software antivirus y el número de virus en los ordenadores de la empresa?
- Calcular el coeficiente de correlación

6. Estadística descriptiva bidimensional
Gráficos. Diag. dispersión (Ejemplo 8.20) (solución)

- Diagrama de dispersión
  - X: Número de veces que se ejecuta el antivirus en cada ordenador en un mes
  - Y: Número de virus detectados en un mes
- Respuesta:
  - El diagrama muestra claramente que el número de virus se reduce, en general, cuando el software antivirus se emplea con más frecuencia
  - Esta relación, sin embargo, no es segura, porque no se detectó virus en algunos ordenadores "afortunados", aunque el software antivirus se ejecutó solo una vez a la semana en ellos
  - Coeficiente de correlación: $r = -0.4533$
    - Demuestra que, en general, si aumenta X disminuye Y, pero no hay una clara correlación entre las variables, al no ser un valor elevado



6. Estadística descriptiva bidimensional
Gráficos. Diagrama de dispersión (Ejemplo 8.22)

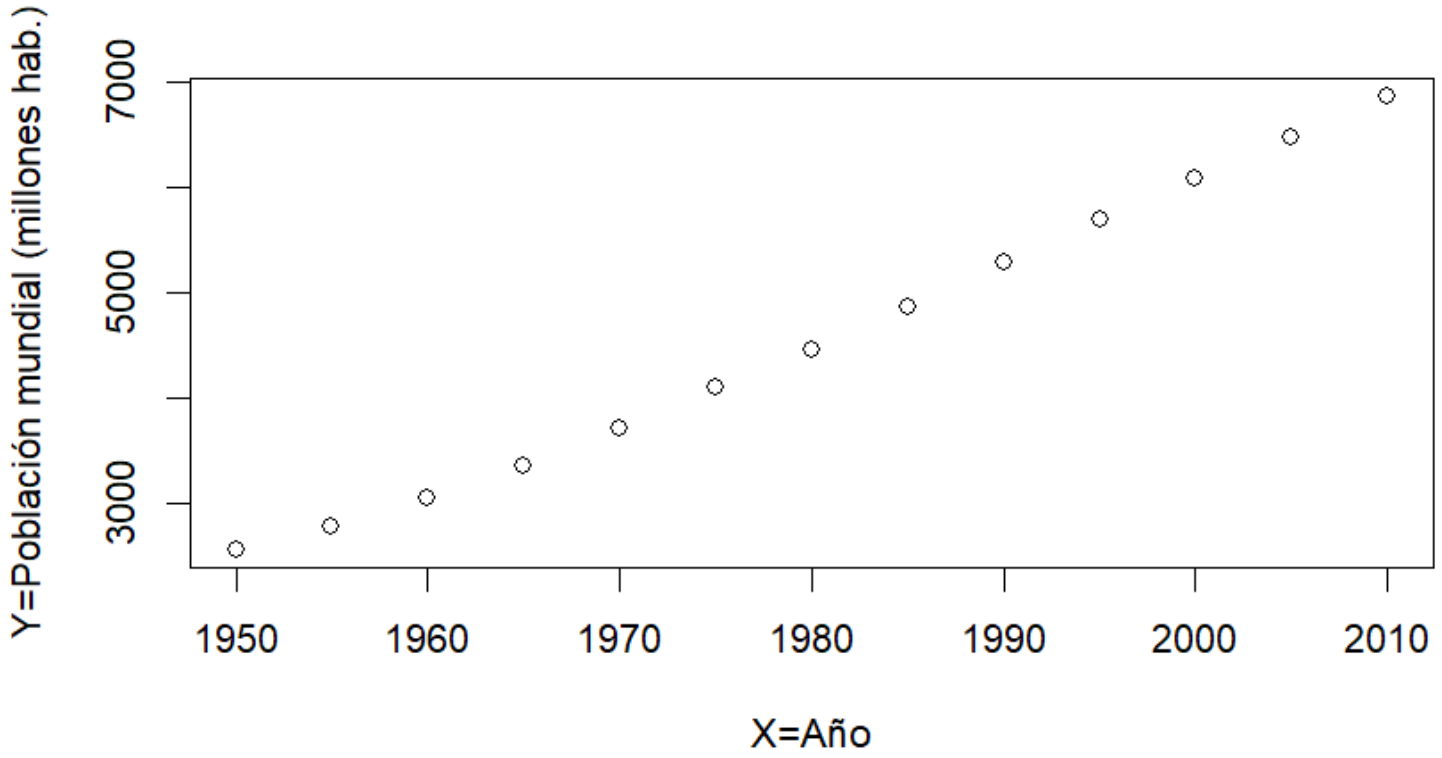
- Según la Base Internacional de Datos de la Oficina del Censo de los Estados Unidos, la población mundial entre 1950 y 2010 creció según esta tabla, en la que:
  - X representa un año
  - Y representa la población mundial en ese año (millones de personas)

X	1950	1955	1960	1965	1970	1975	1980	1985	1990	1995	2000	2005	2010
Y	2558	2782	3043	3350	3712	4089	4451	4855	5287	5700	6090	6474	6864

- Representar el diagrama de dispersión y calcular el coeficiente de correlación
- NOTA: Cuando la variable X representa tiempo, se trata de un *time plot* que muestra la evolución de la variable Y a lo largo del tiempo

6. Estadística descriptiva bidimensional
Gráficos. Diag. dispersión (Ejemplo 8.22)(solución)

X	1950	1955	1960	1965	1970	1975	1980	1985	1990	1995	2000	2005	2010
Y	2558	2782	3043	3350	3712	4089	4451	4855	5287	5700	6090	6474	6864



Coeficiente de correlación:

$$r = 0.9976$$

Es muy próximo a 1, los puntos están prácticamente alineados, existe una fuerte correlación lineal entre las dos variables

6. Estadística descriptiva bidimensional Ejercicios propuestos

- Ejercicios 8.5, 8.6, 8.7 del libro
  - Además de lo que se pide, calcular el coeficiente de correlación y una tabla de contingencia

7. Resumen

- La **Estadística descriptiva** es la rama de la Estadística que describe un conjunto de datos numéricamente (con medidas estadísticas) y gráficamente (con gráficos estadísticos)
- Una **variable estadística** es una propiedad de un elemento (individuo) de una población o muestra, y puede ser cualitativa o cuantitativa
- Una **medida estadística** es una característica numérica sobre una variable estadística de una población (parámetro) o de una muestra (estadístico)
  - Las medidas estadísticas pueden ser de tamaño, centralización, localización o posición, dispersión o variabilidad, forma, o proporción
- Una **tabla de frecuencias** para una variable estadística cuantitativa o cualitativa representa el número de veces que se repite cada valor de la variable en una población o en una muestra
- Los principales **gráficos estadísticos** son el diagrama de barras, diagrama de tarta, diagrama de Pareto, histograma, polígono de frecuencias, diagrama de tallo y hojas (*stem-and-leaf plot*), y diagrama de caja (*Boxplot*)
- La **estadística descriptiva bidimensional** describe simultáneamente dos variables estadísticas que representan dos propiedades diferentes de cada individuo de una población o muestra