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

1. Objetivos

- Recordar conceptos de estadística descriptiva
- Diferenciar entre parámetro de una población y estadístico de una muestra
- Calcular medidas estadísticas de una población y de una muestra
- Elaborar tablas de frecuencias
- Dibujar e interpretar gráficos estadísticos

2. Introducción

- La estadística descriptiva es la rama de la Estadística que describe un conjunto de datos numéricamente y gráficamente
  - Para ello, se utilizan medidas estadísticas (media, mediana, moda, varianza, desviación estándar, rango, etc.),
  - y gráficos estadísticos (diagrama de barras, histograma, diagrama de sectores, diagrama de dispersión, diagrama de caja o boxplot, etc.)
- Las medidas estadísticas pueden referirse a:
  - Una población: se denominan “parámetros” de la población (media poblacional, varianza poblacional, ...)
  - Una muestra de una población: se denominan estadísticos de la muestra (media muestral, varianza muestral, …)

2. Introducción
Definiciones (población, muestra, muestreo)

- Población (*Population*)
  - Todas las unidades, individuos o elementos de interés sobre cuyas propiedades (peso, tamaño, etc.) se quiere realizar un estudio estadístico (descriptivo o inferencial).
  - Pueden ser personas, animales, planetas, cosas, ..
- Muestra (*Sample*)
  - Subconjunto de una población compuesto por una parte de los elementos de la población (denominadas observaciones).
- Muestreo (*Sampling*)
  - Acción de creación de una muestra y recolección de los valores de las propiedades de los elementos incluidos en la muestra.

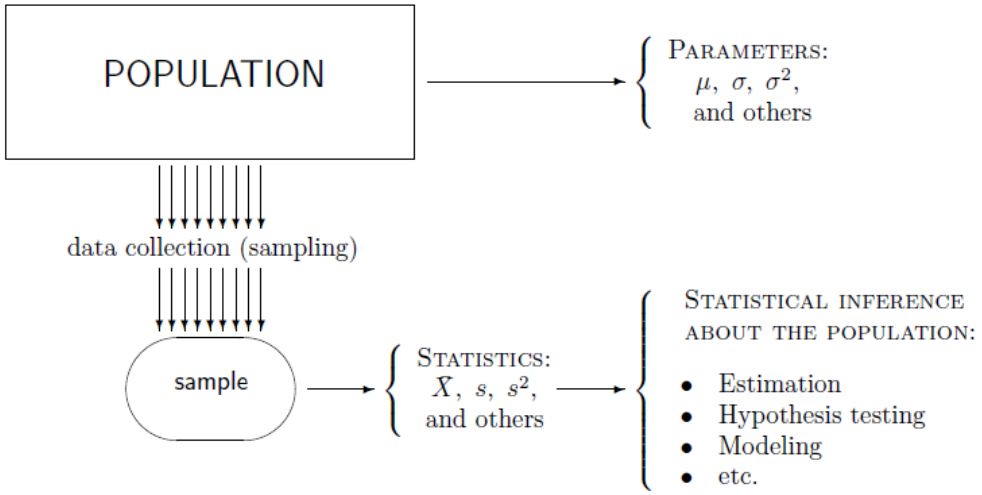


FIGURE 8.1: *Population parameters and sample statistics.*

2. Introducción
Definiciones (variable estadística)

- Una variable estadística es una propiedad de un individuo de una población o muestra
  - Por ejemplo, la edad de cada persona en una población o muestra de personas es una variable estadística
  - Se suele utilizar también el término “elemento” para referirse a “individuo”
  - Se suele utilizar también el término “carácter” para referirse a “propiedad”
  - El valor de la variable para un individuo concreto se denomina **observación**
    - Por ejemplo, en una muestra de 3 personas, se podrían tener como observaciones de la variable EDAD: 34, 18 y 45 años, como la edad de cada una de las tres personas
- Tipos de variables estadísticas
  - Cualitativas
  - Cuantitativas

2. Introducción
Definiciones (variable estadística cualitativa)

- También llamada variable categórica
- Es aquella en la que los valores posibles no son valores numéricos.
- Por ejemplo: estado de salud de una persona, eficiencia energética de un televisor, color de un coche, …
- Puede ser
  - Dicotómica: Cuando puede tener sólo dos valores. Ejemplo: estado de salud de una persona: (sana, enferma).
  - Ordinal: Cuando tiene sentido establecer un orden secuencial o jerarquía. Ejemplo: eficiencia energética de un televisor (A, B, C, D, E, F, G)
  - Nominal: Aquellas que no sugieren ningún orden o jerarquía. Ejemplo: color de un coche (blanco, rojo, azul, …)
- Otros ejemplos: mec.es, proyectodescartes.org

2. Introducción
Definiciones (variable estadística cuantitativa)

- También llamada variable numérica
- Es aquella en la que los valores posibles son números
- Por ejemplo: altura de una persona, número de puertas de un coche
- Puede ser
  - Continua: Cuando se mide dentro de un rango continuo infinito de valores numéricos y se registra con números reales. Por ejemplo: altura de una persona (1,643m, 1,8m, 0,99m, …)
  - Discreta o discontinua: Cuando sólo pueden tomar un número limitado de valores y se registra con números enteros. Ejemplo: número de puertas de un coche (2, 3, 4, 5, …)
- Otros ejemplos: mec.es, proyectodescartes.org

2. Introducción
Definiciones (parámetro, estadístico)

- Parámetro (de una población) (*Parameter*)
  - Una característica numérica sobre una variable estadística (propiedad de los individuos) de una población
  - Se utiliza para hacer afirmaciones sobre una variable estadística de la población
  - Por ejemplo, media poblacional (μ), mediana poblacional (M), varianza poblacional (σ^2), …
- Estadístico (de una muestra) (*Statistic*)
  - Una característica numérica sobre una variable estadística (propiedad de los individuos) de una muestra
  - Se utiliza como estimador del posible valor de un parámetro.
  - Por ejemplo, media muestral ($\bar{X}$), mediana muestral (m), varianza muestral (s^2), …
- Los parámetros y estadísticos son **medidas estadísticas**

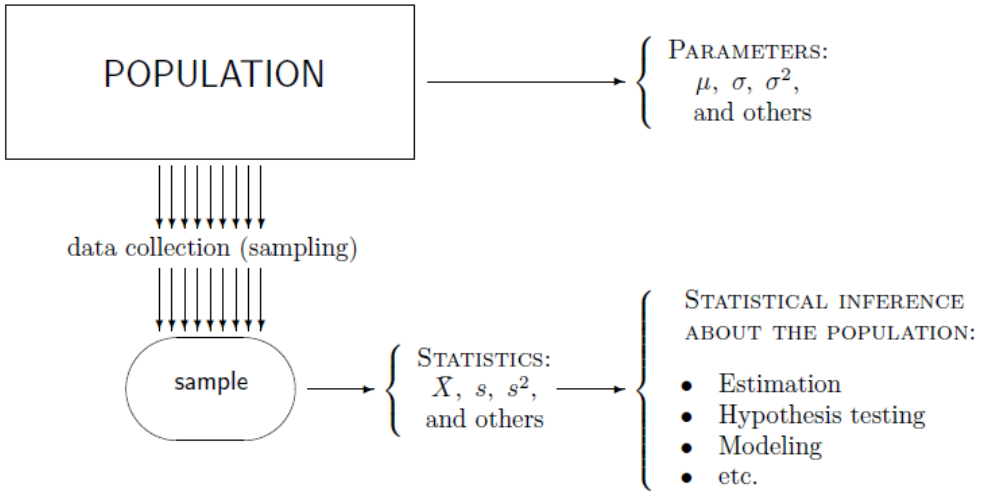


FIGURE 8.1: *Population parameters and sample statistics.*

3. Cálculo de medidas estadísticas

- Hay notaciones y fórmulas diferentes para las medidas sobre
  - una población (medidas poblacionales)
  - una muestra (medidas muestrales)
- Las medidas se pueden clasificar en
  - Tamaño
  - Medidas de centralización
  - Medidas de localización o posición
  - Medidas de dispersión o variabilidad
  - Medidas de forma (no está en el libro)
  - Proporción (no está en el libro)

3. Cálculo de medidas estadísticas
Tamaño (de una población o de una muestra)

- Tamaño de una población: N
  - Población (*Poblation*): $P = (x_1, x_2, .., x_N)$
  - Donde x_i es el valor de la propiedad (variable estadística) x para el elemento (individuo) i de la población
    - Ejemplo: si x representa el peso de una persona de una población, entonces x_i es el peso de la persona i del total de N personas de la población
- Tamaño de una muestra: n
  - Muestra (*Sample*): $S = (X_1, X_2, .., X_n)$ Siendo $n \leq N$
  - Donde X_i es el valor de la propiedad (variable estadística) X para el elemento (individuo) i de la muestra
    - Ejemplo: si X representa el peso de una persona de una muestra, entonces X es el peso de la persona i del total de n personas de la muestra

Medida	Poblacional	Muestral
Tamaño	N	n

3. Cálculo de medidas estadísticas
Medidas de centralización

- Las medidas de centralización o de posición de tendencia central indican un valor alrededor del cual se distribuyen las observaciones
  - Media: Es el promedio de los valores de las observaciones
    - Aritmética: cociente entre la suma de todos los valores y el número de valores
    - Otras medias: cuadrática, geométrica, ponderada, …
  - Mediana: Es un número que es superado por, como máximo, la mitad de las observaciones y es precedido por, como máximo, la mitad de las observaciones
  - Moda (no está en el libro): Es el valor (o valores) que más se repite. Puede haber varias modas

3. Cálculo de medidas estadísticas
Medidas de centralización (cálculo)

Medida	Poblacional	Muestral
Media (aritmética)	$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$
Mediana	Se ordenan los N elementos de menor a mayor: $Población\ ordenada = (x_1, x_2, .., x_N)$ Si $N/2$ no es un número entero: $M = x_{\lfloor N/2 \rfloor + 1}$ Si $N/2$ es un número entero: $M = \frac{x_{N/2} + x_{N/2+1}}{2}$	Se ordenan los n elementos de menor a mayor: $Muestra\ ordenada = (X_1, X_2, .., X_n)$ Si n/2 no es un número entero: $\widehat{M} = X_{\lfloor n/2 \rfloor + 1}$ Si n/2 es un número entero: $\widehat{M} = \frac{X_{n/2} + X_{n/2+1}}{2}$
Moda	Valor x_i que más se repite	Valor X_i que más se repite

NOTA: El operador $\lfloor número \rfloor$ representa la función suelo, es decir, la parte entera del número que hay en su interior.

3. Cálculo de medidas estadísticas
Medidas de centralización (interpretación)

- Si la mediana es igual a la media
  - Hay un número similar de valores inferiores y superiores a la media
- Si la mediana es menor que la media
  - La mayoría de los valores están por debajo de la media
- Si la mediana es mayor que la media
  - La mayoría de los valores están por encima de la media

3. Cálculo de medidas estadísticas
Medidas de centralización. 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)
  - $S = (70,\ 36,\ 43,\ 69,\ 82, 48,\ 34, 62, 35, 15, 59, 139, 46, 37, 42,$ $30, 55, 56, 36, 82, 38, 89, 54, 25, 35, 24, 22, 9, 56, 19)$
  - $n = 30$
- Calcular las siguientes medidas de centralización para la muestra
  a) Media (aritmética)
  b) Mediana
  c) Moda (no está en el libro)

3. Cálculo de medidas estadísticas
Medidas de centralización. Ejemplo 8.12 (solución)

- Primero ordenamos los $n = 30$ valores de menor a mayor
- $S = (9, 15, 19, 22, 24, 25, 30, 34, 35, 35, 36, 36, 37, 38, 42, 43, 46, 48, 54, 55, 56, 56, 59, 62, 69, 70, 82, 82, 89, 139)$
- a) Media (aritmética)
  - $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{\sum_{i=1}^{30} X_i}{30} = \frac{9+15+\cdots+89+139}{30} = \frac{1447}{30} = 48.23\ seg$
- b) Mediana
  - $\widehat{M} = \frac{X_{n/2}+X_{(n+2)/2}}{2} = \frac{X_{15}+X_{16}}{2} = \frac{42+43}{2} = 42.5\ seg$
  - Como $\widehat{M} < \bar{X}$, la distribución de valores es asimétrica hacia la derecha (right-skewed), es decir hay más valores por debajo de la media (18 de 30).
- c) Moda
  - Hay cuatro modas: 35, 36, 56 y 82, porque estos valores se repiten 2 veces

3. Cálculo de medidas estadísticas
Medidas de localización (o posición)

- Las medidas de localización o de posición de tendencia no central, permiten conocer otros puntos característicos de la distribución de los datos que no son los valores centrales.
- Son valores de la distribución ordenada de menor a mayor que la dividen en partes iguales, es decir en intervalos que comprenden el mismo número de datos
  - Cuantiles: genérico
    - Un cuantil p es cualquier número que supera como maximo al 100*p% de los datos y es superado como máximo por el 100*(1-p)% de los datos
  - Percentiles: cien intervalos
    - Un percentil γ es cualquier número que supera como maximo al $\gamma\%$ de los datos, y es superado como maximo por el $(100-\gamma)\%$ de los datos
  - Cuartiles: cuatro intervalos
    - Primer cuartil: Es cualquier número que supera como máximo a una cuarta parte (1/4 o 25%) de los datos, y es superado como máximo por tres cuartas partes (3/4 o 75%) de las observaciones.
    - Segundo cuartil: Es igual a la mediana
    - Tercer cuartil: Es cualquier número que supera como máximo a las tres cuartas partes (3/4 o 75%) de los datos, y es superado como máximo por una cuarta parte (1/4 o 25%) de los datos.

3. Cálculo de medidas estadísticas
Medidas de localización. Ejemplos

- Ejemplo de cuantil 0.2: con $p = 0.2$ y $S = (1, 2, \mathbf{3}, 4, 5, 6, 7, 8, 9, 10, 11)$
  - El 100*0.2% de 11 datos es 2.2
  - El 100*(1-0.2)% de 11 datos es 8.8
  - El cuantil 0.2 es “3” porque se cumple que
    - supera como maximo a 2.2 datos, ya que sólo supera a los 2 que tiene a su izquierda
    - y es superado como máximo por 8.8 datos, ya que sólo es superado por los 8 que tiene a su derecha
- Ejemplo de cuantil 0.2: con $p = 0.2$ y $S = (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)$
  - El 100*0.2% de 10 datos es 2
  - El 100*(1-0.2)% de 10 datos es 8
  - Un cuantil 0.2 podría ser “2.5” porque
    - supera como maximo a 2 datos, ya que el valor 2.5 sólo supera al 1 y al 2
    - y es superado como máximo por 8 datos, ya que 2.5 es superado por 3, 4, 5, 6, 7, 8, 9 y 10
  - NOTA: En este caso, podría ser también 2.1, 2.2, etc., pero se suele usar el valor medio de los dos valores a ambos lados del cuantil, en este ejemplo el valor medio de 2 y 3, que es 2.5

3. Cálculo de medidas estadísticas
Medidas de localización (cuantiles y percentiles)

Medida	Poblacional	Muestral
Cuantil p	$Población\ ordenada = (x_1, x_2, .., x_N)$ Si pN no es un número entero: $q_p = x_{\lfloor pN \rfloor + 1}$ Si pN es un número entero: $q_p = \frac{x_{pN} + x_{pN+1}}{2}$	$Muestra\ ordenada = (X_1, X_2, .., X_n)$ Si pn no es un número entero: $\hat{q}_p = X_{\lfloor pn \rfloor + 1}$ Si pn es un número entero: $\hat{q}_p = \frac{X_{pn} + X_{pn+1}}{2}$
Percentil γ	Si $\gamma N/100$ no es un número entero: $\pi_\gamma = x_{\lfloor \gamma N/100 \rfloor + 1}$ Si $\gamma N/100$ es un número entero: $\pi_\gamma = \frac{x_{\gamma N/100} + x_{\gamma N/100+1}}{2}$	Si $\gamma n/100$ no es un número entero: $\hat{\pi}_\gamma = X_{\lfloor \gamma n/100 \rfloor + 1}$ Si $\gamma n/100$ es un número entero: $\hat{\pi}_\gamma = \frac{X_{\gamma n/100} + X_{\gamma n/100+1}}{2}$

NOTA: Existen otras fórmulas alternativas para el cálculo de los cuantiles cuando pN o pn es un número entero.

3. Cálculo de medidas estadísticas
Medidas de localización (cuartiles)

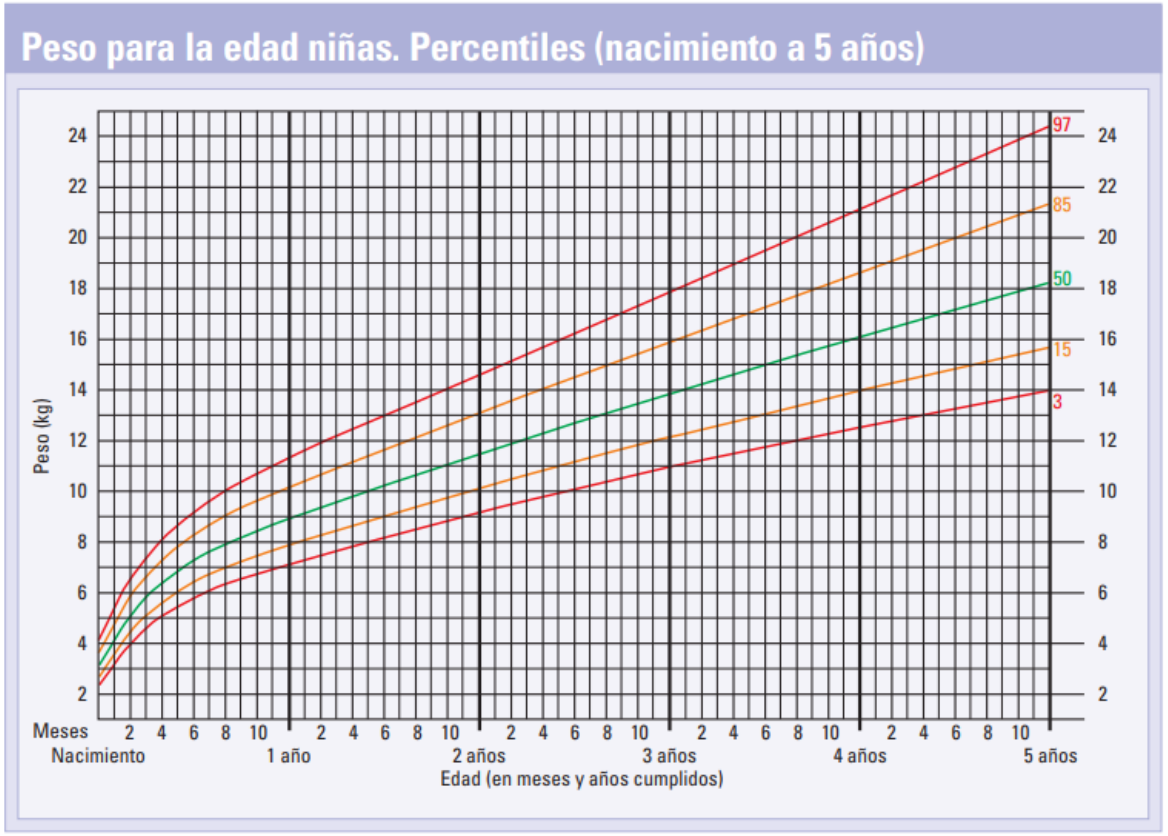
Medida	Poblacional	Muestral
Primer cuartil	$Población\ ordenada = (x_1, x_2, .., x_N)$ Si $N/4$ no es un número entero: $Q_1 = x_{\lfloor N/4 \rfloor + 1}$ Si $N/4$ es un número entero: $Q_1 = \frac{x_{N/4} + x_{N/4+1}}{2}$	$Muestra\ ordenada = (X_1, X_2, .., X_n)$ Si $n/4$ no es un número entero: $\hat{Q}_1 = X_{\lfloor N/4 \rfloor + 1}$ Si $n/4$ es un número entero: $\hat{Q}_1 = \frac{X_{n/4} + X_{n/4+1}}{2}$
Segundo cuartil	$Q_2 = M$	$\hat{Q}_2 = \hat{M}$
Tercer cuartil	Si $3N/4$ no es un número entero: $Q_3 = x_{\lfloor 3N/4 \rfloor + 1}$ Si $3N/4$ es un número entero: $Q_3 = \frac{x_{3N/4} + x_{3N/4+1}}{2}$	Si $3n/4$ no es un número entero: $\hat{Q}_3 = X_{\lfloor 3N/4 \rfloor + 1}$ Si $3n/4$ es un número entero: $\hat{Q}_3 = \frac{X_{3n/4} + X_{3n/4+1}}{2}$

3. Cálculo de medidas estadísticas
Medidas de localización (equivalencias)

- El cuantil p es igual al percentil p*100
  - $q_p = \pi_{p100}$
- El primer cuartil es igual al cuantil 0.25 y percentil 25
  - $Q_1 = q_{0.25} = \pi_{25}$
- El segundo cuartil es igual a la mediana y al cuantil 0.5 y percentil 50
  - $Q_2 = M = q_{0.5} = \pi_{50}$
- El tercer cuartil es igual al cuantil 0.75 y percentil 75
  - $Q_3 = q_{0.75} = \pi_{75}$

3. Cálculo de medidas estadísticas
Medidas de localización (interpretación)

- Las medidas de localización permiten comparar la evolución de un individuo dentro de una población o muestra
- Se usan percentiles para conocer cuál es el patrón normal (estándar) de crecimiento de los niños y niñas para detectar a tiempo la aparición de algún problema
- La Organización Mundial de la Salud (OMS) y otros organismos publican tablas de crecimiento para ir observando el percentil al que corresponde el niño o niña en cada momento, de un conjunto de cinco percentiles importantes según la OMS: 3, 15, 50, 85 y 97.
- Si el niño no se mantiene en un mismo percentil o cercano, habría que investigar qué puede ocurrir
- Por ejemplo,
  - si una niña con 1 año pesa 10kg estaría en el percentil 85 (comparado con las de su edad, un 85% pesan menos que ella),
  - y si con 2 años sigue pesando 10kg, estaría ahora en el percentil 15 (sólo un 15% de las niñas de su edad pesan menos que ella), y habría que analizar las causas.



Patrones de crecimiento infantil de la OMS.

Fuente: AEPAP

3. Cálculo de medidas estadísticas
Medidas de localización. Ejemplo 8.14

- Para evaluar la efectividad de un procesador para un determinado tipo de tareas, registramos el tiempo de CPU para una muestra de n = 30 trabajos elegidos aleatoriamente (en segundos)
- $S = (9, 15, 19, 22, 24, 25, 30, 34, 35, 35, 36, 36, 37, 38, 42, 43, 46, 48, 54, 55, 56, 56, 59, 62, 69, 70, 82, 82, 89, 139)$
- Calcular las siguientes medidas de localización para la muestra
  a) Primer cuartil
  b) Segundo cuartil
  c) Tercer cuartil
  d) Cuantil 0.1 (no está en el libro)
  e) Percentil 60 (no está en el libro)

3. Cálculo de medidas estadísticas
Medidas de localización. Ejemplo 8.14 (solución)(I)

- La muestra de 30 valores ya está ordenada de menor a mayor
  - $S = (9, 15, 19, 22, 24, 25, 30, 34, 35, 35, 36, 36, 37, 38, 42, 43, 46, 48, 54, 55, 56, 56, 59, 62, 69, 70, 82, 82, 89, 139)$
- La mediana se calculó en el ejemplo 8.12 → $\widehat{M} = 42.5\ seg$
- a) Primer cuartil ($n/4 = 7.5$, no es un número entero)
  - $\hat{Q}_1 = X_{\lfloor N/4 \rfloor + 1} = X_{\lfloor 7.5 \rfloor + 1} = X_{7+1} = X_8 = 34\ seg$
  - *La cuarta parte (25%) de los trabajos consumen un tiempo de CPU inferior a 34s*
- b) Segundo cuartil
  - $\hat{Q}_2 = \widehat{M} = 42.5\ seg$
  - *La mitad de los trabajos consumen un tiempo de CPU inferior a 42.5s*
- c) Tercer cuartil ($3n/4 = 22.5$, no es un número entero)
  - $\hat{Q}_3 = X_{\lfloor 3N/4 \rfloor + 1} = X_{\lfloor 22.5 \rfloor + 1} = X_{22+1} = X_{23} = 59\ seg$
  - *Las tres cuartas partes (75%) de los trabajos consumen un tiempo de CPU inferior a 59s*

3. Cálculo de medidas estadísticas
Medidas de localización. Ejemplo 8.14 (solución)(II)

- La muestra de 30 valores ya está ordenada de menor a mayor
  - $S = (9, 15, 19, 22, 24, 25, 30, 34, 35, 35, 36, 36, 37, 38, 42, 43, 46, 48, 54, 55, 56, 56, 59, 62, 69, 70, 82, 82, 89, 139)$
- d) Cuantil 0.1 (p$n = 0.1 \cdot 30 = 3$, es un número entero)
  - $\hat{q}_{0.1} = \frac{X_{pn}+X_{pn+1}}{2} = \frac{X_3+X_{3+1}}{2} = \frac{X_3+X_4}{2} = \frac{19+22}{2} = 20.5\ seg$
  - Es igual que el percentil 10: $\hat{q}_{0.1} = \hat{\pi}_{10} = 20.5\ seg$
  - *El 10% de los trabajos consumen un tiempo de CPU inferior a 20.5s*
- e) Percentil 60 ($\gamma n/100$=$60 \cdot \frac{30}{100} = 18$, es un número entero)
  - $\hat{\pi}_{60} = \frac{X_{\gamma n/100}+X_{\gamma n/100+1}}{2} = \frac{X_{18}+X_{19}}{2} = \frac{48+54}{2} = 51\ seg$
  - Es igual que el cuantil 0.6: $\hat{\pi}_{60} = \hat{q}_{0.6} = 51\ seg$
  - *El 60% de los trabajos consumen un tiempo de CPU inferior a 51s*

3. Cálculo de medidas estadísticas
Medidas de dispersión (o variabilidad)

- Las medidas de dispersión o variabilidad reflejan la heterogeneidad de las observaciones y dan una idea sobre la representatividad de las medidas de centralización, de tal forma que a mayor dispersión menor representatividad
- Rango: La diferencia entre el valor más grande y el más pequeño
  - Tiene las mismas unidades que la variable estadística
- Varianza: Mide la variabilidad entre las observaciones
  - Es un valor positivo con las unidades las de la variable estadística al cuadrado
- Desviación estándar (o típica): Raíz cuadrada de la varianza
  - Tiene las mismas unidades que la variable estadística
- Coeficiente de variación: Cociente entre la desviación estándar y la media
  - Es un valor sin unidades, por tanto, no depende de cambios de escala
- Rango intercuartílico (*Interquartile range*): diferencia entre el tercer cuartil y el primer cuartil
  - Evita el efecto de posibles datos atípicos (*outliers*), que afectan a las otras medidas

3. Cálculo de medidas estadísticas
Medidas de dispersión (cálculo)

Medida	Poblacional	Muestral
Valor mínimo, valor máximo, rango	Si la población está ordenada de menor a mayor: $\min x = x_1$, $\max x = x_N$ $rango\ x = x_N - x_1$	Si la muestra está ordenada de menor a mayor: $\min X = X_1$, $\max X = X_n$ $rango\ X = X_n - X_1$
Varianza	$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$	También llamada cuasivarianza muestral $s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$
Desviación estándar	$\sigma = \sqrt{\sigma^2}$	$s = \sqrt{s^2}$
Coeficiente de variación	$cv = \frac{\sigma}{\mu}$	$CV = \frac{s}{\bar{X}}$
Rango intercuartílico	$IQR = Q_3 - Q_1$ Los datos atípicos están fuera del intervalo: $[Q_1 - 1.5 \cdot IQR,\ Q_3 + 1.5 \cdot IQR]$	$\widehat{IQR} = \hat{Q}_3 - \hat{Q}_1$ Los datos atípicos están fuera del intervalo: $\left[\hat{Q}_1 - 1.5 \cdot \widehat{IQR}, \hat{Q}_3 + 1.5 \cdot \widehat{IQR}\right]$

3. Cálculo de medidas estadísticas Medidas de dispersión (interpretación)(I)

- Si la varianza, desviación estándar y coeficiente de variación son pequeños:
  - Hay poca dispersión
  - Es un conjunto de datos homogéneo
  - La media es representativa del conjunto de datos
  - Ejemplo: En la fabricación de productos, valores pequeños suponen mejor calidad del producto
- Si la varianza, desviación estándar y coeficiente de variación son grandes:
  - Hay mucha dispersión
  - Es un conjunto de datos heterogéneo
  - La media no es representativa del conjunto de datos
  - Ejemplo: En economía, valores grandes suponen un mayor riesgo al invertir el dinero
- ¿Cuál es el límite entre valores pequeños y grandes?
  - No hay consenso entre los expertos
  - Algunos expertos afirman que en el caso del coeficiente de variación sería 0.3 (30%)
  - Pero otros expertos proponen valores diferentes, dependiendo del ámbito de estudio

3. Cálculo de medidas estadísticas
Medidas de dispersión (interpretación)(II)

- El rango intercuartílico permite comprobar si existen datos atípicos (*outliers*).
- Regla $\mathbf{1.5 \cdot IQR}$: Los datos fuera del siguiente intervalo son sospechosos de ser atípicos y habría que plantearse si conviene eliminarlos o no
  - $[Q_1 - 1.5 \cdot IQR,\ Q_3 + 1.5 \cdot IQR]$

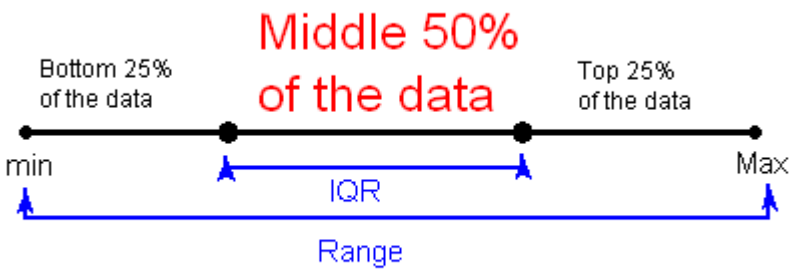
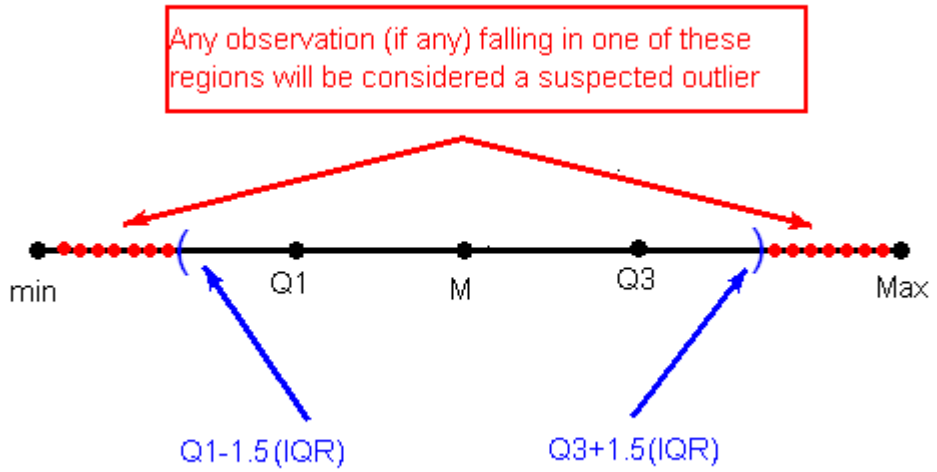


Imagen: libretexts.org



3. Cálculo de medidas estadísticas
Medidas de dispersión. Ejemplos 8.16 y 8.18

- Para evaluar la efectividad de un procesador para un determinado tipo de tareas, registramos el tiempo de CPU para una muestra de n = 30 trabajos elegidos aleatoriamente (en segundos)
- $S = (9, 15, 19, 22, 24, 25, 30, 34, 35, 35, 36, 36, 37, 38, 42, 43, 46, 48, 54, 55, 56, 56, 59, 62, 69, 70, 82, 82, 89, 139)$
- Calcular las medidas de dispersión para la muestra
  a) Rango
  b) Varianza (cuasivarianza muestral)
  c) Desviación estándar
  d) Coeficiente de variación (no está en el libro)
  e) Rango intercuartílico
  f) Detectar si hay datos atípicos
  g) Si hay algún dato atípico, comparar las medidas estadísticas si se eliminasen de la muestra los datos atípicos (no está en el libro)

3. Cálculo de medidas estadísticas
Medidas de dispersión. Ejemplos 8.16 y 8.18 (solución)(I)

- La muestra ya está ordenada de menor a mayor
  - $S = (9, 15, 19, 22, 24, 25, 30, 34, 35, 35, 36, 36, 37, 38, 42, 43, 46, 48, 54, 55, 56, 56, 59, 62, 69, 70, 82, 82, 89, 139)$
- La media muestral se calculó en el ejemplo 8.12 → $\bar{X} = 48.23\ seg.$
- a) Rango
  - $rango = X_n - X_1 = X_{30} - X_1 = 139 - 9 = 130\ seg.$
- b) Varianza (cuasivarianza)
  - $s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} = \frac{\sum_{i=1}^{30}(X_i - \bar{X})^2}{30-1} = \frac{(9-48.23)^2 + \cdots + (139-48.23)^2}{29} = 703.15\ seg^2$
- c) Desviación estándar
  - $s = \sqrt{s^2} = \sqrt{703.15} = 26.52\ seg$
- d) Coeficiente de variación
  - $CV = \frac{s}{\bar{X}} = \frac{26.52}{48.23} = 0.55$
  - Como $CV > 0.3$ puede suponerse que hay una gran dispersión y, por tanto, la media no es representativa del conjunto de datos

3. Cálculo de medidas estadísticas
Medidas de dispersión. Ejemplos 8.16 y 8.18 (solución)(II)

- Los cuartiles se calcularon en el ejemplo 8.14 → $\hat{Q}_1 = 34\text{s}$, $\hat{Q}_3 = 59\text{s}$
- e) Rango intercuartílico
  - $\widehat{IQR} = \hat{Q}_3 - \hat{Q}_1 = 59 - 34 = 25\ seg$

- f) Detectar si hay datos atípicos
  - $\left[\hat{Q}_1 - 1.5 \cdot \widehat{IQR}, \hat{Q}_3 + 1.5 \cdot \widehat{IQR}\right] = [34 - 1.5 \cdot 25, 59 + 1.5 \cdot 25] = [-3.5, 96.5]$
  - El valor 139 seg es atípico porque está fuera del intervalo

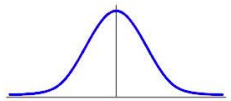
3. Cálculo de medidas estadísticas
Medidas de dispersión. Ejemplos 8.16 y 8.18 (solución)(III)

- g) Comparar las medidas estadísticas si se eliminasen de la muestra los datos atípicos

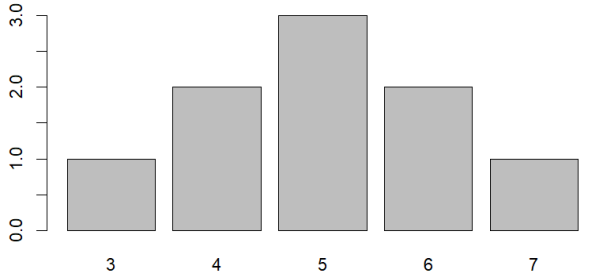
Medida	Muestra original (n=30)	Muestra sin el dato atípico (n=29)
Media	48.23s	45.1s
Mediana	42.5s	42s
Primer cuartil	34s	34s
Tercer cuartil	59s	56s
Rango	130s	80s
Varianza	$703.15s^2$	$423.88s^2$
Desviación estándar	26.52s	20.59s
Coeficiente de variación	0.55	0.46
Rango intercuartílico	25s	22s

3. Cálculo de medidas estadísticas
Medidas de forma

- Permiten conocer la forma que tiene la curva que representa la distribución de las frecuencias (repeticiones) de los valores de las observaciones, normalmente un histograma o un diagrama de barras.
- Se pueden utilizar para comparar con un posible conjunto de datos con la misma media y varianza, pero considerado como “Normal”.
- Se considera “Normal” un conjunto de datos cuyo diagrama de frecuencias se ajusta aproximadamente a la curva conocida como campana de Gauss:
  - $f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$



  - Ejemplo: $S = (3,4,4,5,5,5,6,6,7)$, la media es 5
- Algunas medidas de forma son las siguientes
  - Sesgo o coeficiente de asimetría
  - Curtosis o coeficiente de apuntamiento



3. Cálculo de medidas estadísticas
Medidas de forma (cálculo)

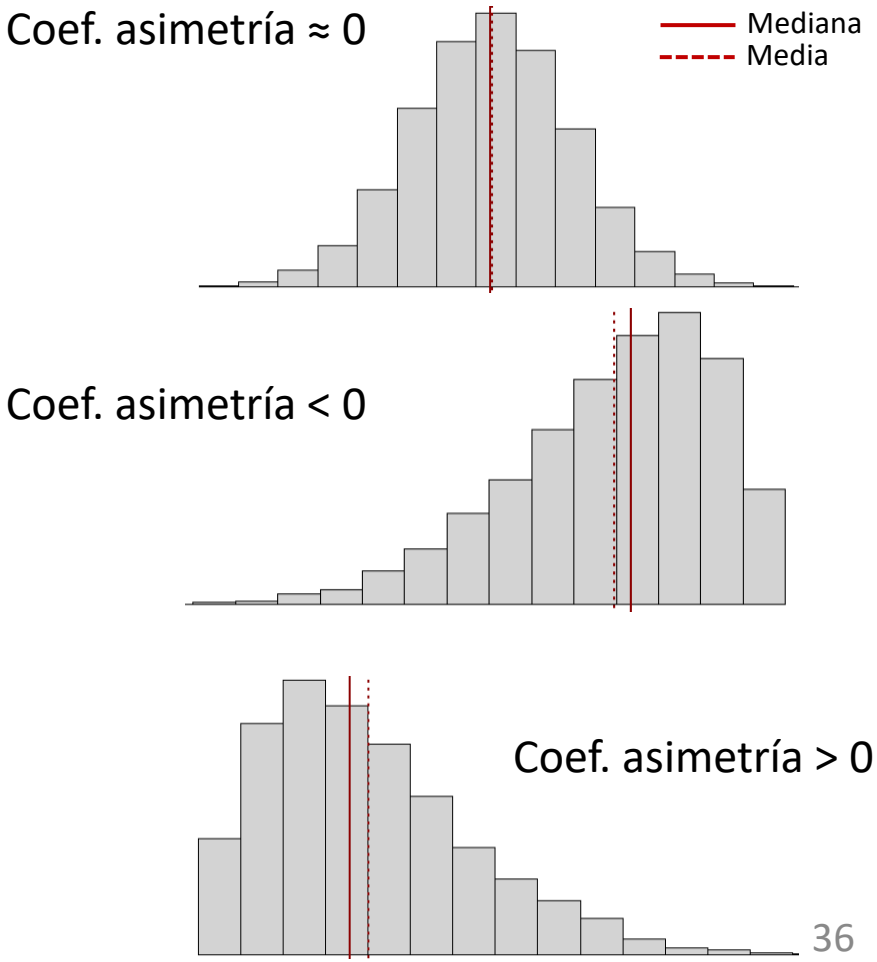
Medida	Poblacional	Muestral
Sesgo o coeficiente de asimetría	$A = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^3}{\sigma^3}$	$\hat{A} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3}{s^3}$
Curtosis o coeficiente de apuntamiento	$K = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^4}{\sigma^4} - 3$	$\widehat{K} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^4}{s^4} - 3$

NOTA: Existen otras fórmulas alternativas para el cálculo de las medidas de forma.

3. Cálculo de medidas estadísticas Medidas de forma (interpretación)(I)

Coeficiente de asimetría

- Si es = 0, la distribución de los datos es simétrica
  - Hay un número similar de valores inferiores y superiores a la media
  - La mediana es igual que la media
- Si < 0, es asimétrica hacia la izquierda (*left-skewed*)
  - Hay más valores por encima de la media
  - La mediana es mayor que la media
- Si es > 0, la distribución de los valores es asimétrica hacia la derecha (r*ight-skewed*)
  - Hay más valores por debajo de la media
  - La mediana es menor que la media

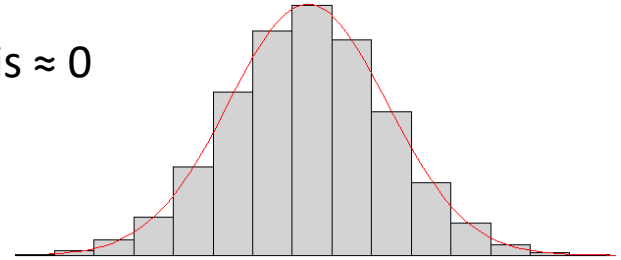


3. Cálculo de medidas estadísticas Medidas de forma (interpretación)(II)

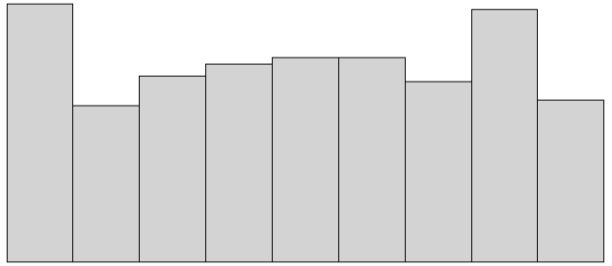
Curtosis (Coeficiente de apuntamiento)

- Si = 0, histograma con una pendiente similar a la distribución normal
  - Se dice que es una distribución de datos mesocúrtica
- Si <0, histograma con pendiente menos apuntada (más aplanada) que la normal
  - Se dice que es platicúrtica
- Si > 0, histograma con pendiente más abrupta (apuntada) que la normal
  - Se dice que es leptocúrtica

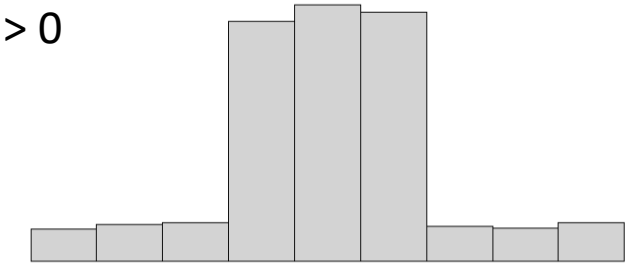
Curtosis ≈ 0



Curtosis < 0



Curtosis > 0



3. Cálculo de medidas estadísticas
Medidas de forma. 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)
  - $S = (9, 15, 19, 22, 24, 25, 30, 34, 35, 35, 36, 36, 37, 38, 42, 43, 46, 48, 54, 55, 56, 56, 59, 62, 69, 70, 82, 82, 89, 139)$
  - $n = 30$
- Calcular las siguientes medidas de forma para la muestra
  a) Sesgo o coeficiente de asimetría (no está en el libro)
  b) Curtosis o coeficiente de apuntamiento (no está en el libro)
  c) Comparar si se eliminan los datos atípicos (no está en el libro)

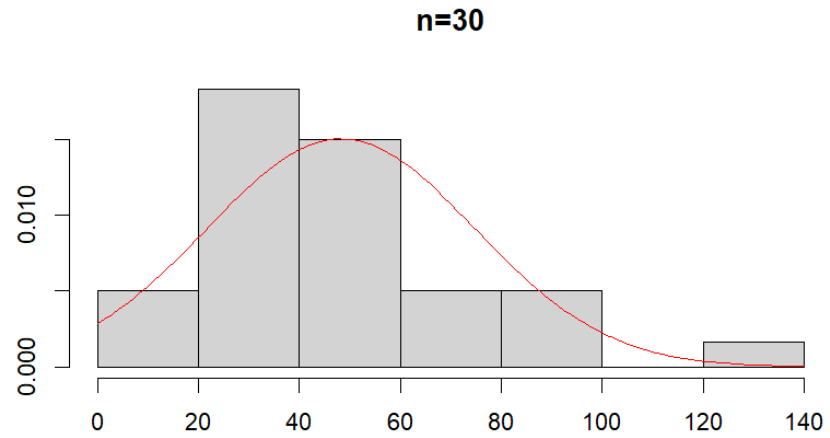
3. Cálculo de medidas estadísticas Medidas de forma. Ejemplo 8.12 (solución)(I)

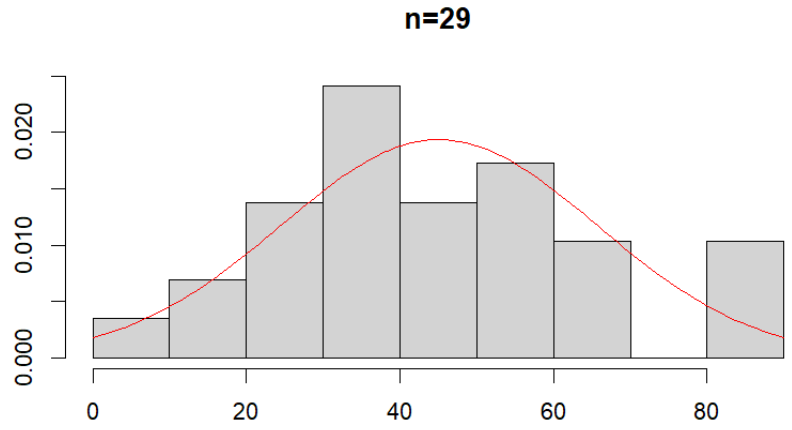
- $S = (9, 15, 19, 22, 24, 25, 30, 34, 35, 35, 36, 36, 37, 38, 42, 43, 46, 48, 54, 55, 56, 56, 59, 62, 69, 70, 82, 82, 89, 139)$
- $n = 30, \bar{X} = 48.23s, s = 26.52s$
- a) Sesgo o coeficiente de asimetría
  - $\hat{A} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3}{s^3} = \frac{\frac{1}{30}\sum_{i=1}^{30}(X_i - 48.23)^3}{26.52^3} = \frac{\frac{1}{30}(9-48.23)^3 + \cdots + (139-48.23)^3}{26.52^3} = 1.31$
  - Como $\hat{A} > 0$, es una distribución de datos asimétrica a la derecha (*right-skewed*)
- b) Curtosis o coeficiente de apuntamiento
  - $\widehat{K} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^4}{s^4} - 3 = \frac{\frac{1}{30}\sum_{i=1}^{30}(X_i - 48.23)^4}{26.52^4} = \frac{\frac{1}{30}(9-48.23)^4 + \cdots + (139-48.23)^4}{26.52^4} = 2.35$
  - Como $\widehat{K} > 0$, es una distribución de datos leptocúrtica

3. Cálculo de medidas estadísticas Medidas de forma. Ejemplo 8.12 (solución)(II)

- c) Comparar las medidas de forma si se eliminasen de la muestra los datos atípicos

Medida	Muestra original (n=30)	Muestra sin el dato atípico (n=29)
Sesgo o coeficiente de asimetría	1.31 Asimétrica a la derecha	0.37 Asimétrica a la derecha
Curtosis o coeficiente de apuntamiento	2.35 Leptocúrtica	-0.69 Platicúrtica





3. Cálculo de medidas estadísticas
Proporción

- Una proporción representa la relación de elementos de una población o muestra que cumplen una determinada condición sobre el total de elementos de la población
- Se mide entre 0 y 1.
  - Si se multiplica por 100 se convierte en porcentaje
- Ejemplo
  - Proporción de personas que pesan más de 50 kilos
  - Proporción de coches con 3 puertas

3. Cálculo de medidas estadísticas
Proporción (cálculo)

Medida	Poblacional	Muestral
Proporción de elementos que cumplen una condición	$Población = (x_1, x_2, .., x_N)$ $p = \frac{nº\ elementos\ que\ cumplen\ la\ condición}{N}$	$Muestra\ = (X_1, X_2, .., X_n)$ $\hat{p} = \frac{nº\ elementos\ que\ cumplen\ la\ condición}{n}$

3. Cálculo de medidas estadísticas Proporción. 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)
  - $S = (9, 15, 19, 22, 24, 25, 30, 34, 35, 35, 36, 36, 37, 38, 42, 43, 46, 48, 54, 55, 56, 56, 59, 62, 69, 70, 82, 82, 89, 139)$
  - $n = 30$
- Calcular la proporción de trabajos que se ejecutan en menos de un minuto (no está en el libro)

3. Cálculo de medidas estadísticas Proporción. Ejemplo 8.12 (solución)

- $S = (9, 15, 19, 22, 24, 25, 30, 34, 35, 35, 36, 36, 37, 38, 42, 43, 46, 48, 54, 55, 56, 56, 59, 62, 69, 70, 82, 82, 89, 139)$
- $n = 30, \bar{X} = 48.23s, s = 26.52s$
- Calcular la proporción de trabajos que se ejecutan en menos de un minuto
  - $\hat{p} = \frac{nº\ de\ trabajos\ que\ se\ ejecutan\ en\ menos\ de\ 60\ segundos}{30} = \frac{23}{30} = 0.77$

3. Cálculo de medidas estadísticas Ejercicios propuestos

- Ejercicios 8.1, 8.2, 8.8, 8.9 del libro
  - NOTA: Cuando el enunciado de un ejercicio pide “compute the five-point summary”, se refiere a calcular las cinco medidas siguientes: valor mínimo, primer cuartil, mediana, tercer cuartil y valor máximo
  - Calcular todas las medidas posibles en cada ejercicio, aunque no se pidan en el libro
  - Las respuestas de 8.1, 8.2, 8.6, 8.8 están disponibles en el libro

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): wpd.ugr.es

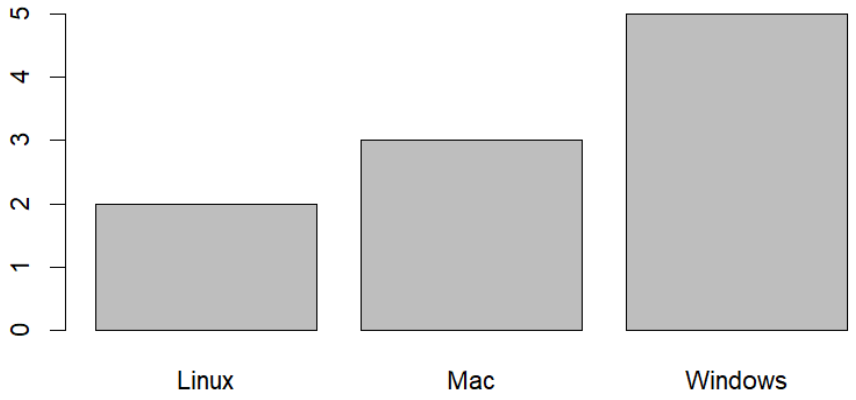
5. Gráficos estadísticos

- 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

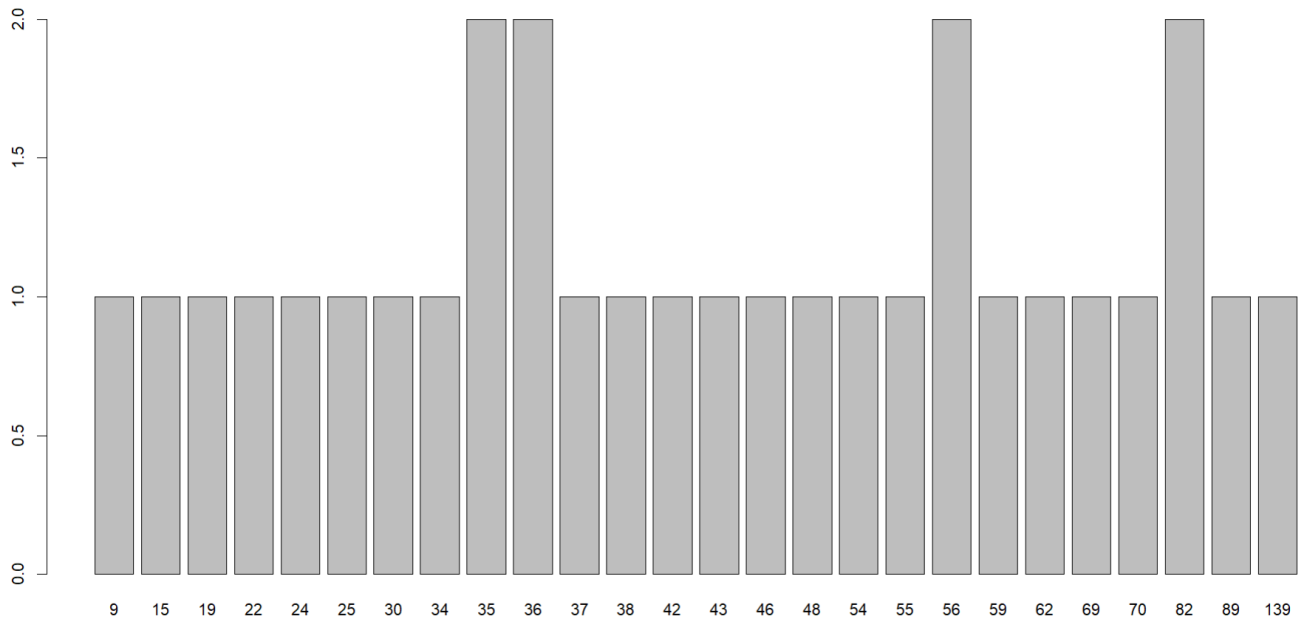
$Fila$	X	f	h
1	Linux	2	$2/10 = 0.2$
2	Mac	3	$3/10 = 0.3$
3	Windows	5	$5/10 = 0.5$
		10	1



5. Gráficos estadísticos
Diagrama de barras. Ejemplo 8.12

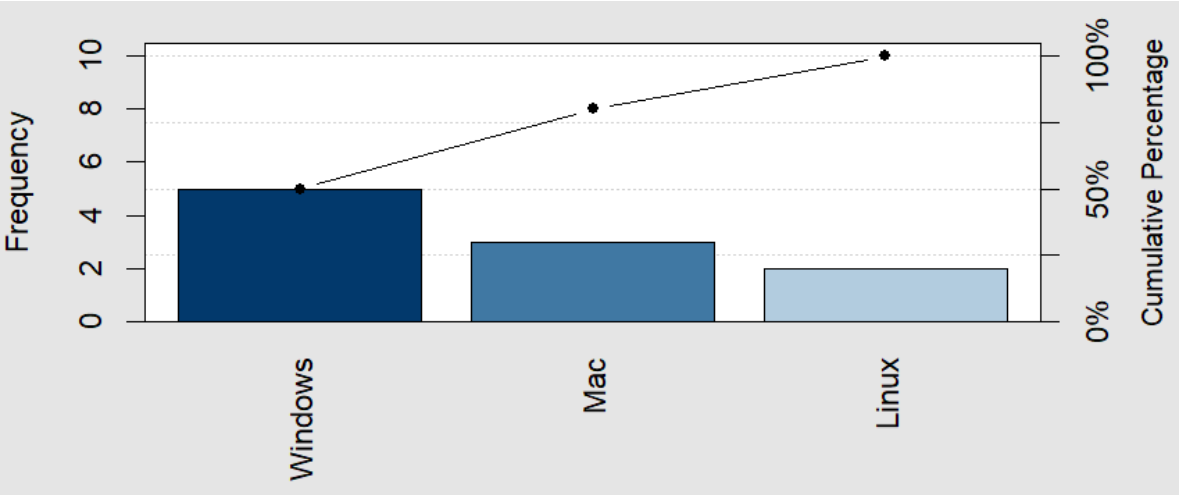
- Para evaluar la efectividad de un procesador para un determinado tipo de tareas, registramos el tiempo de CPU para una muestra de 30 trabajos elegidos aleatoriamente (en segundos)
- S=(9, 15, 19, 22, 24, 25, 30, 34, 35, 35, 36, 36, 37, 38, 42, 43, 46, 48, 54, 55, 56, 56, 59, 62, 69, 70, 82, 82, 89,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

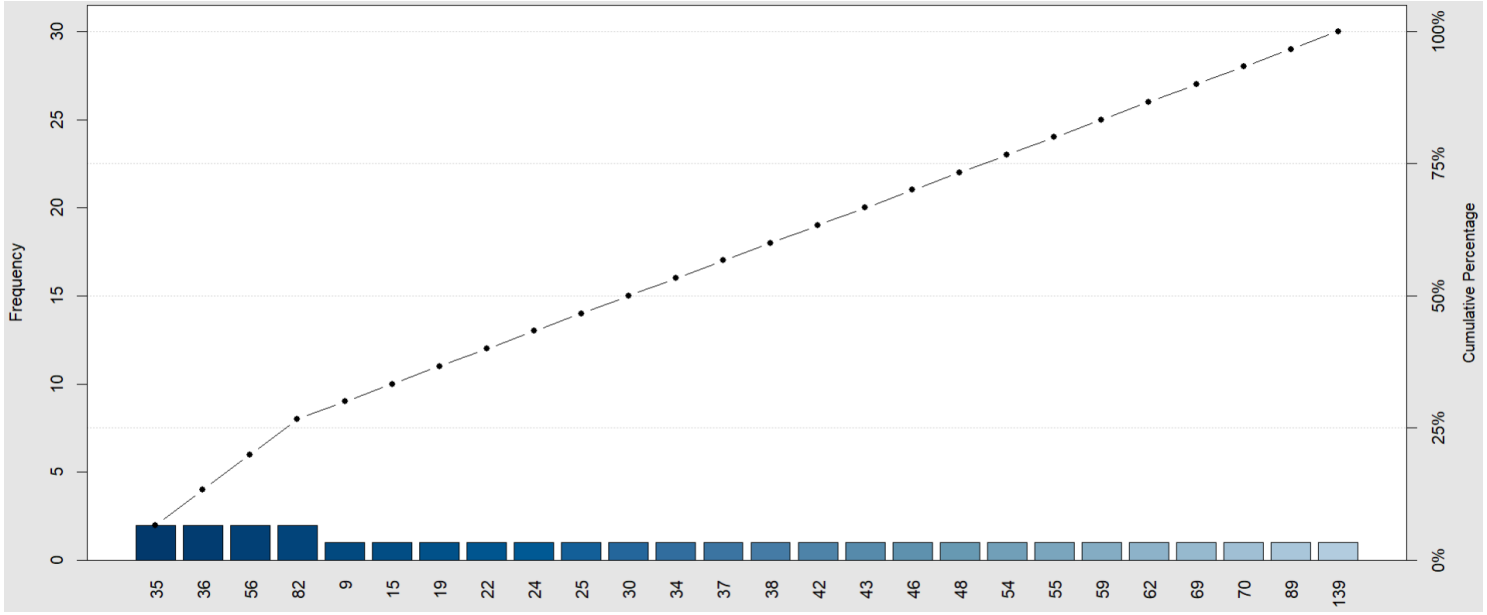


$Fila$	X	f	F
1	Linux	2	2
2	Mac	3	5
3	Windows	5	10

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

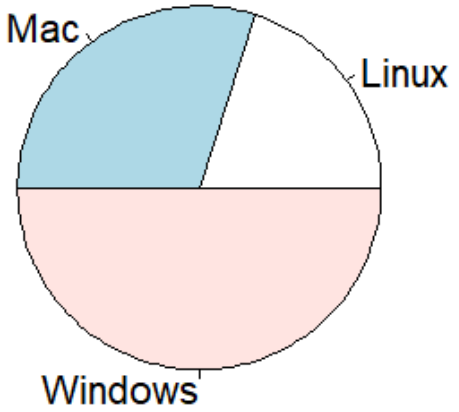
Fila	*X*	*f*	*F*
1	9	1	1
2	15	1	2
…	…	…	…
24	82	2	28
25	89	1	29
26	139	1	30



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

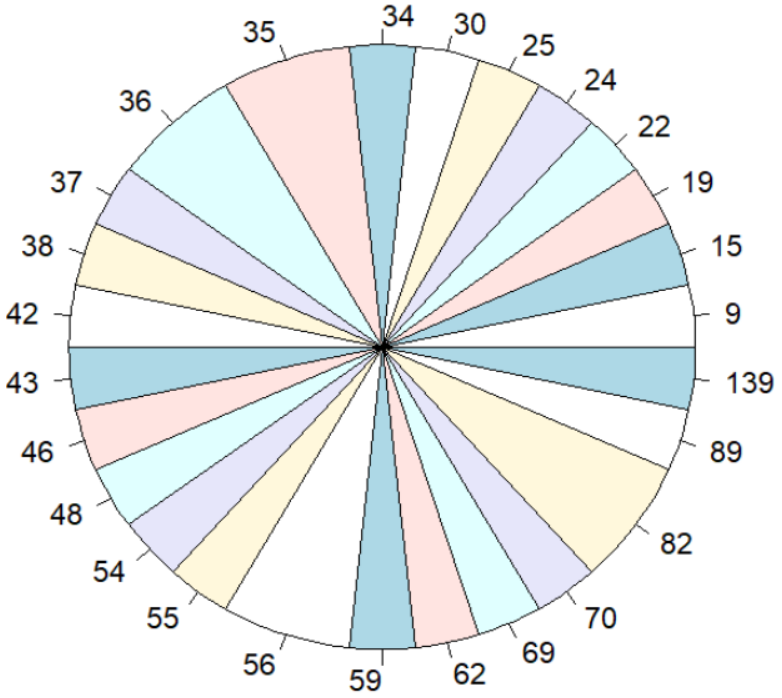
$Fila$	X	f	h
1	Linux	2	$2/10 = 0.2$
2	Mac	3	$3/10 = 0.3$
3	Windows	5	$5/10 = 0.5$
		10	1



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

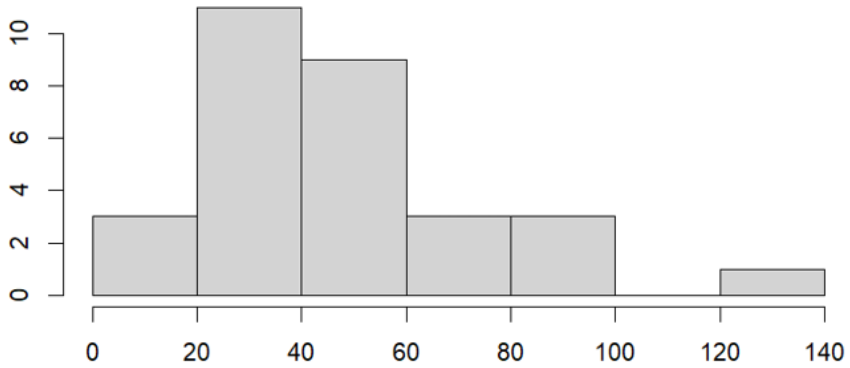
Fila	*X*	*f*
1	9	1
2	15	1
…	…	…
24	82	2
25	89	1
26	139	1
		30



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



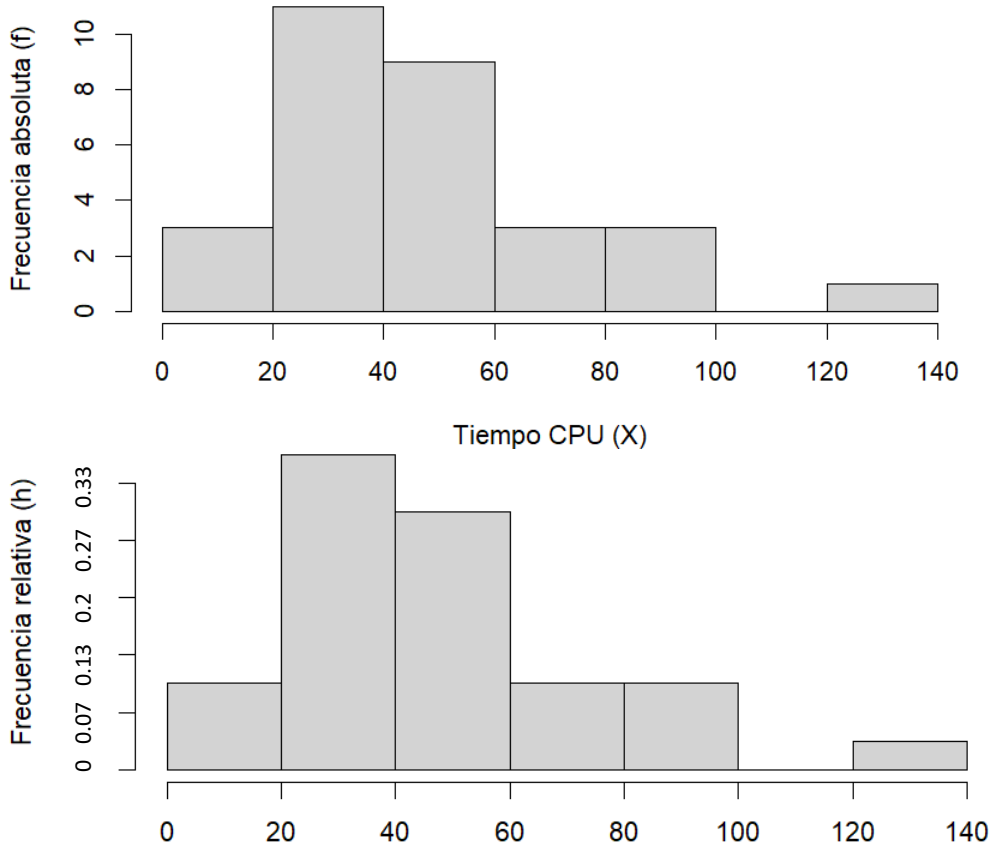
5. Gráficos estadísticos
Histograma. Ejemplo 8.12

- 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)

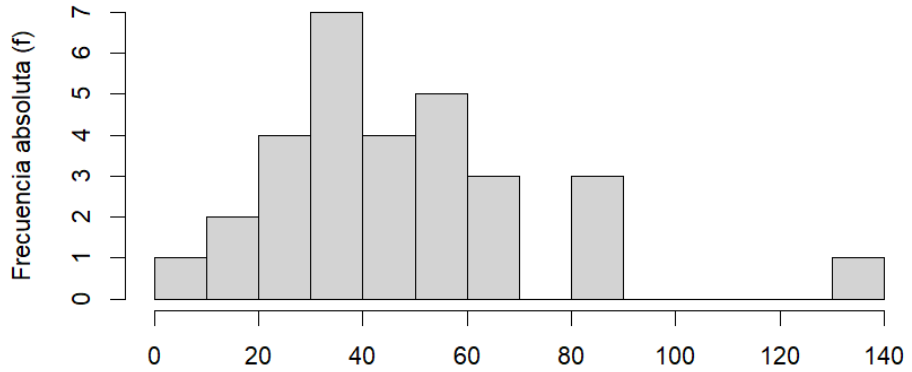


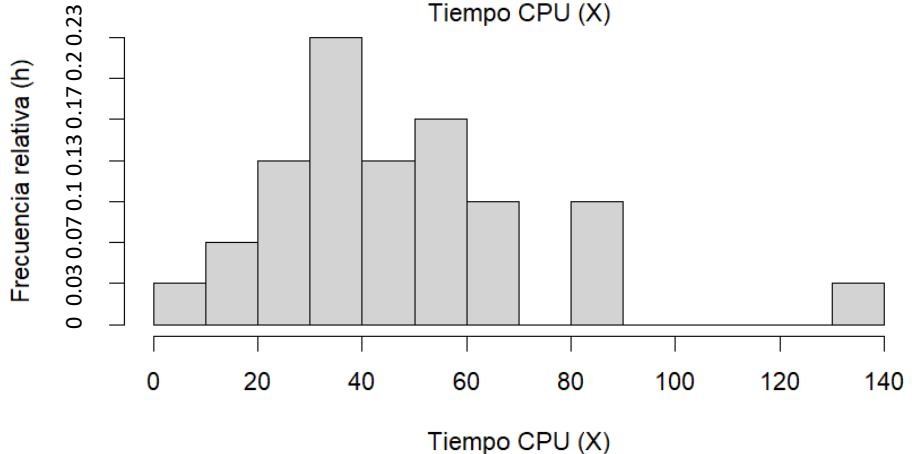
Fila	***X***	***f***	***h***
1	$[0, 20)$	3	$3/30 = 0.10$
2	$[20, 40)$	11	$11/30 = 0.37$
3	$[40, 60)$	9	$9/30 = 0.30$
4	$[60, 80)$	3	$3/30 = 0.10$
5	$[80, 100)$	3	$3/30 = 0.10$
6	$[100, 120)$	0	$0/30 = 0$
7	$[120,140]$	1	$1/30 = 0.03$
		30	1

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)

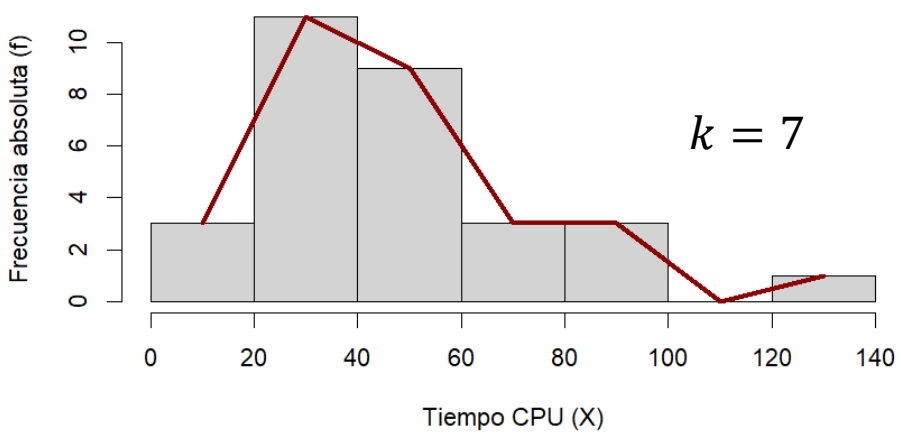


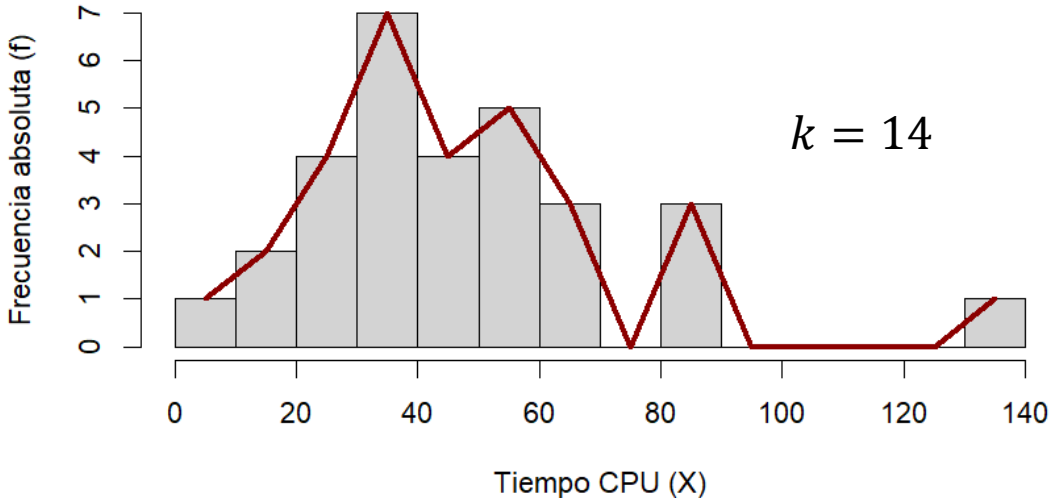


Fila	*X*	*f*	*h*
1	$[0, 10)$	1	$1/30 = 0.03$
2	$[10, 20)$	2	$2/30 = 0.07$
3	$[20, 30)$	3	$3/30 = 0.10$
4	$[30, 40)$	8	$8/30 = 0.27$
5	$[40, 50)$	4	$4/30 = 0.13$
6	$[50, 60)$	5	$5/30 = 0.17$
7	$[60,70]$	2	$2/30 = 0.07$
8	$[70, 80)$	1	$3/30 = 0.03$
9	$[80, 90)$	3	$3/30 = 0.10$
10	$[90, 100)$	0	$0/30 =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

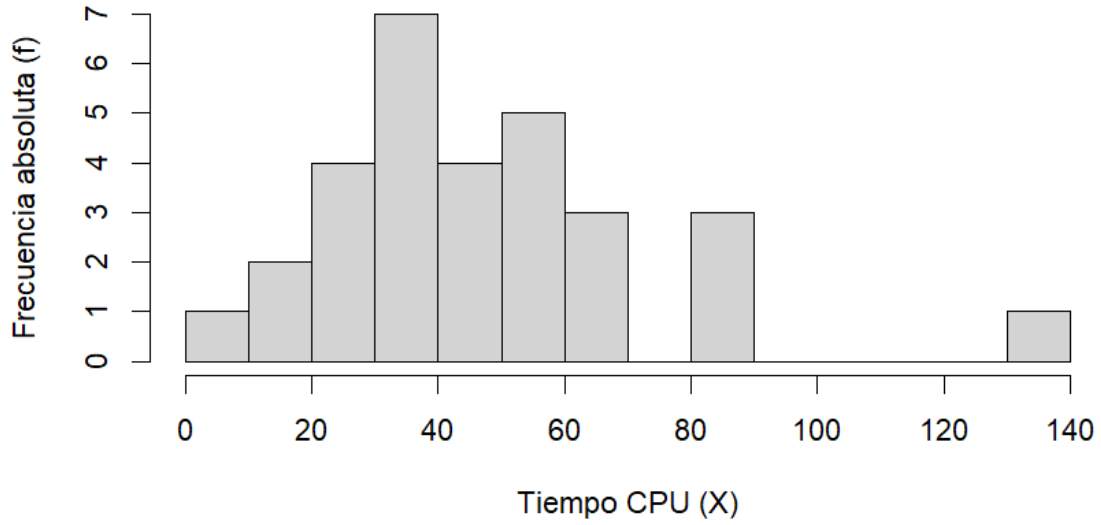
5. Gráficos estadísticos
Diagrama de tallo y hojas. Escala

- 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

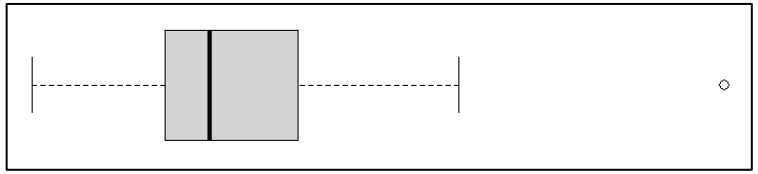
- Si se gira un diagrama de tallo y hojas, se asemeja a un histograma
- Pero el diagrama de tallo y hojas tiene más información → se pueden ver todos los valores, no sólo las frecuencias absolutas




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

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

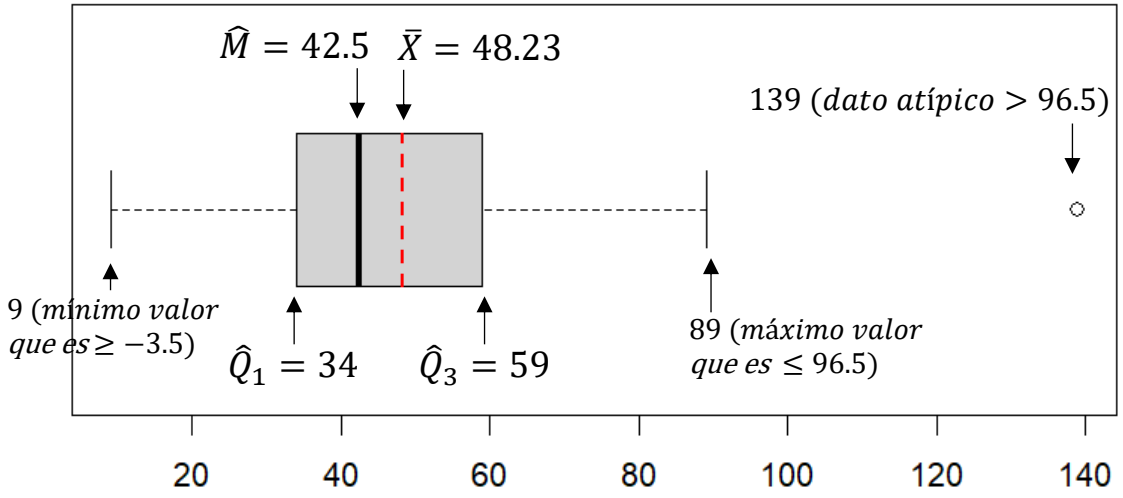
5. Gráficos estadísticos
Diagrama de caja. Ejemplo 8.12

- 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)

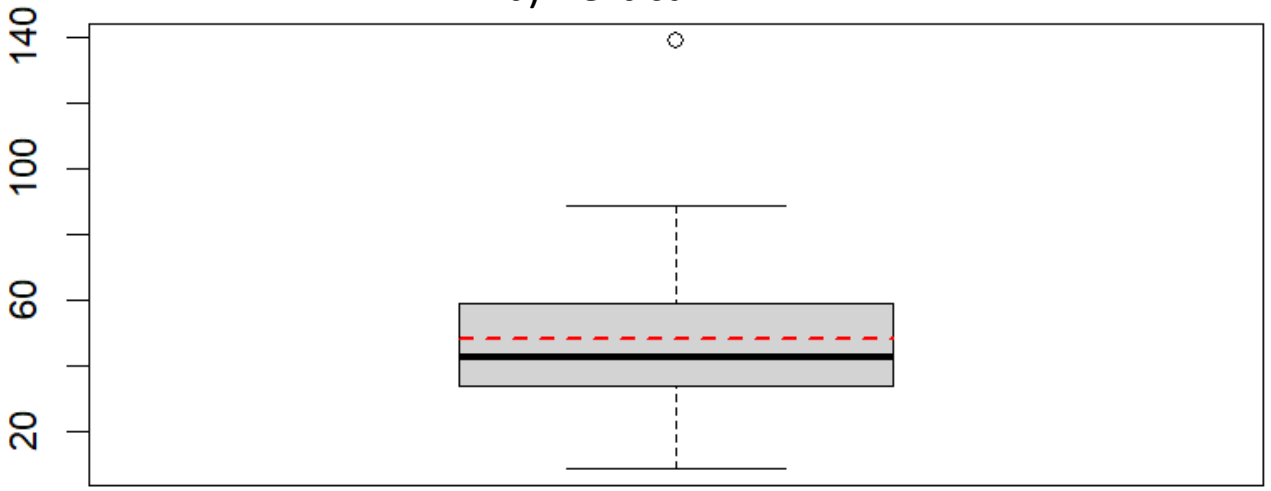
5. Gráficos estadísticos
Diagrama de caja. Ejemplo 8.12 (solución)

- *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

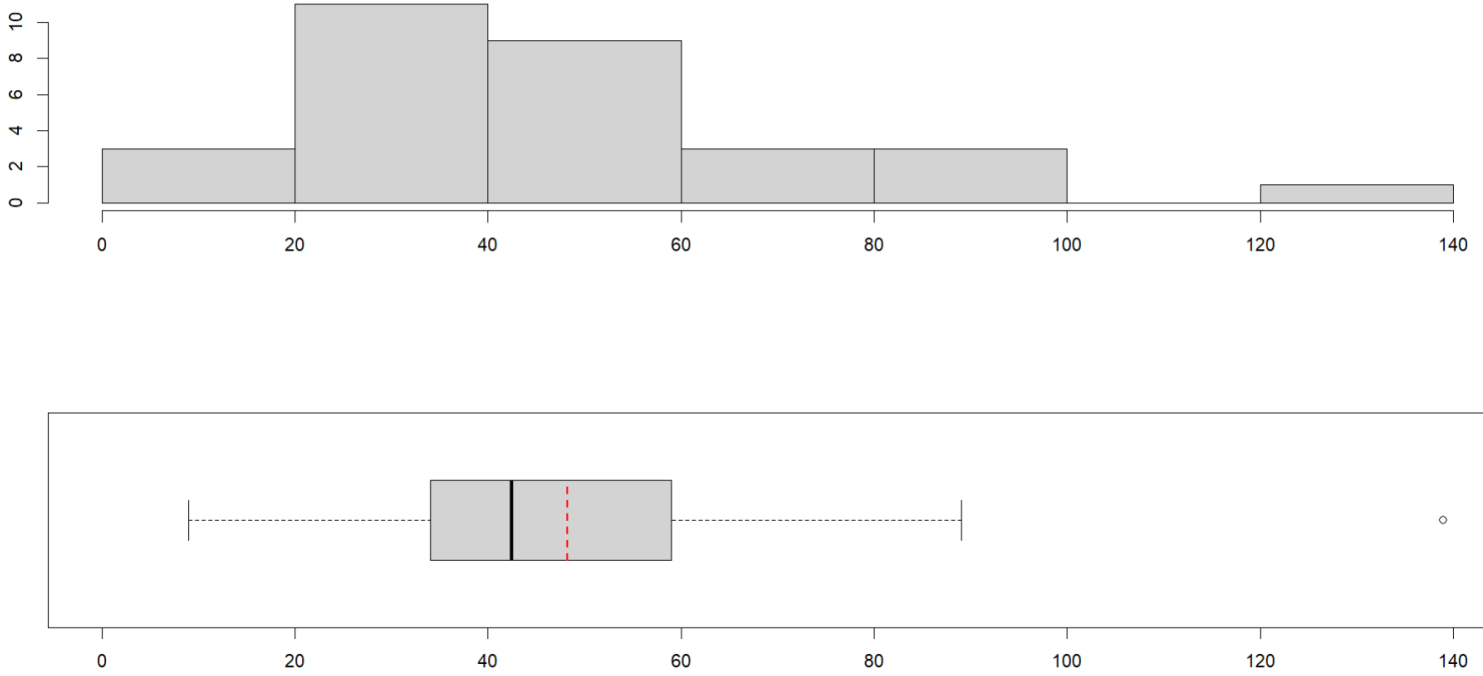


b) Vertical



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?

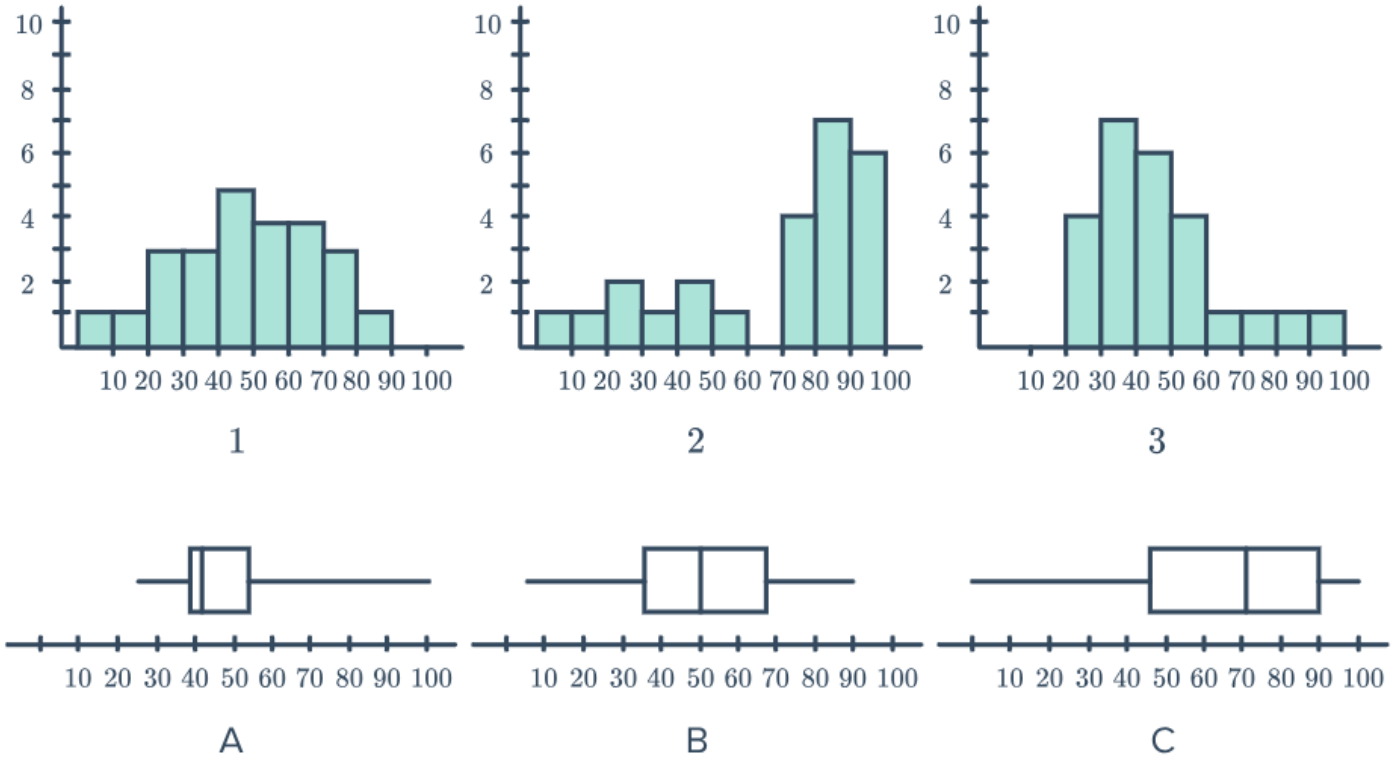
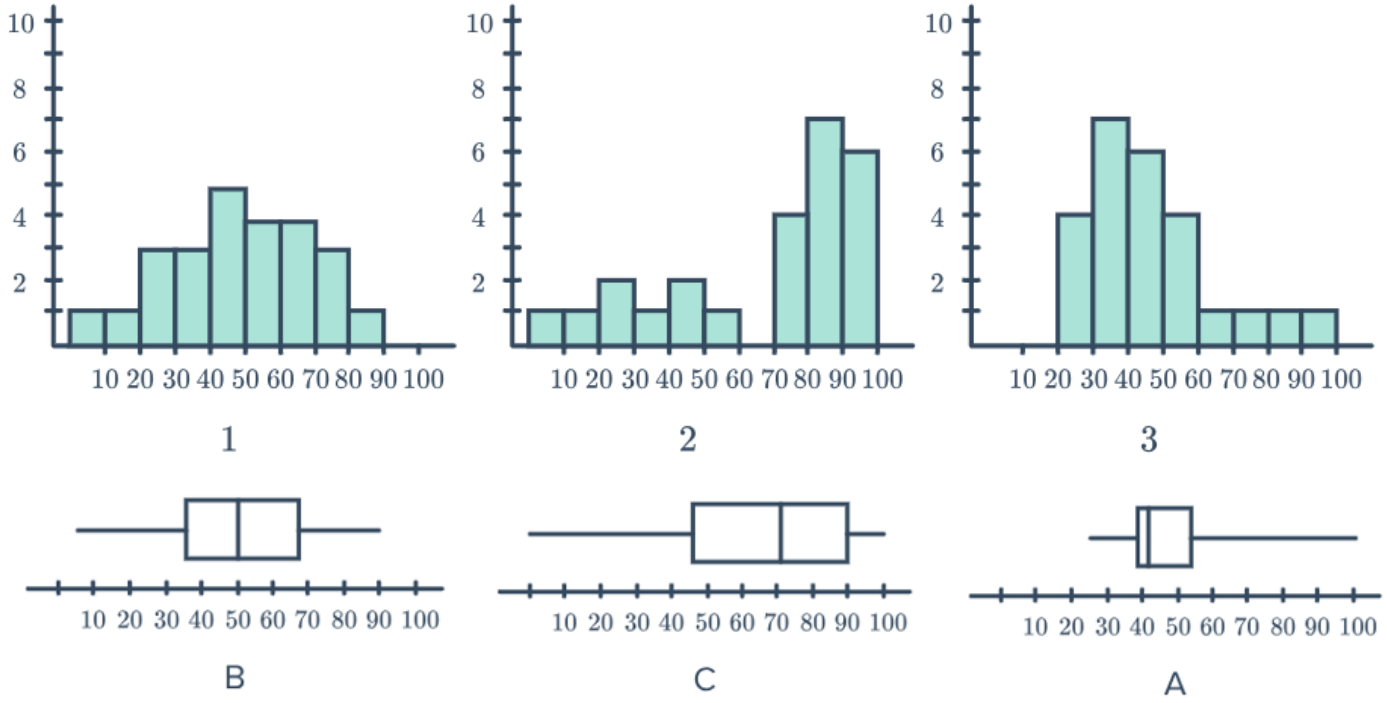


Imagen: mathspace.co

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

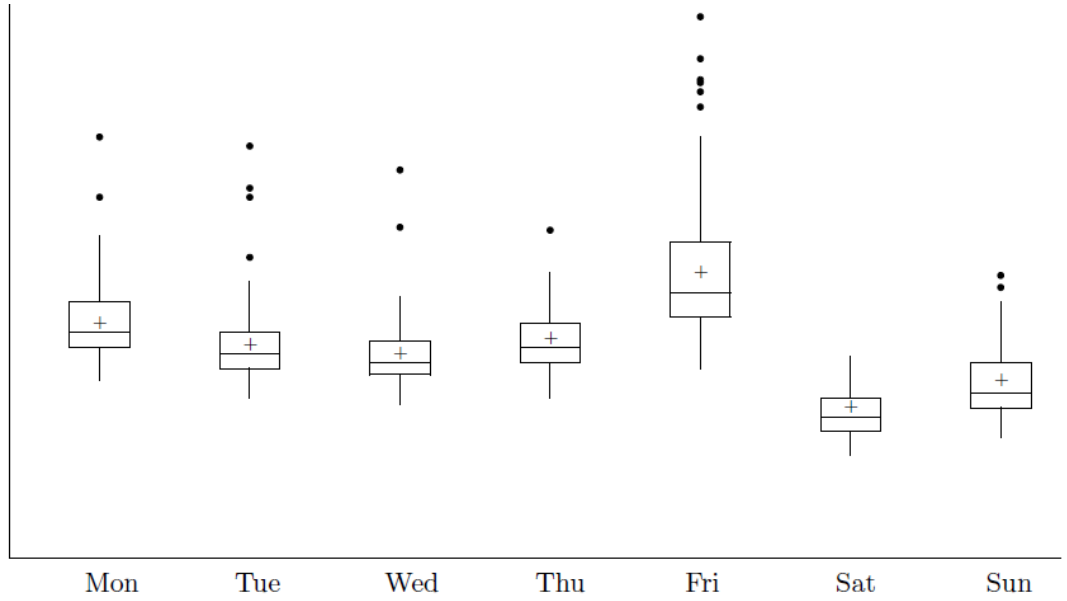


FIGURE 8.10: *Parallel boxplots of internet traffic.*

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 (II)

- 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

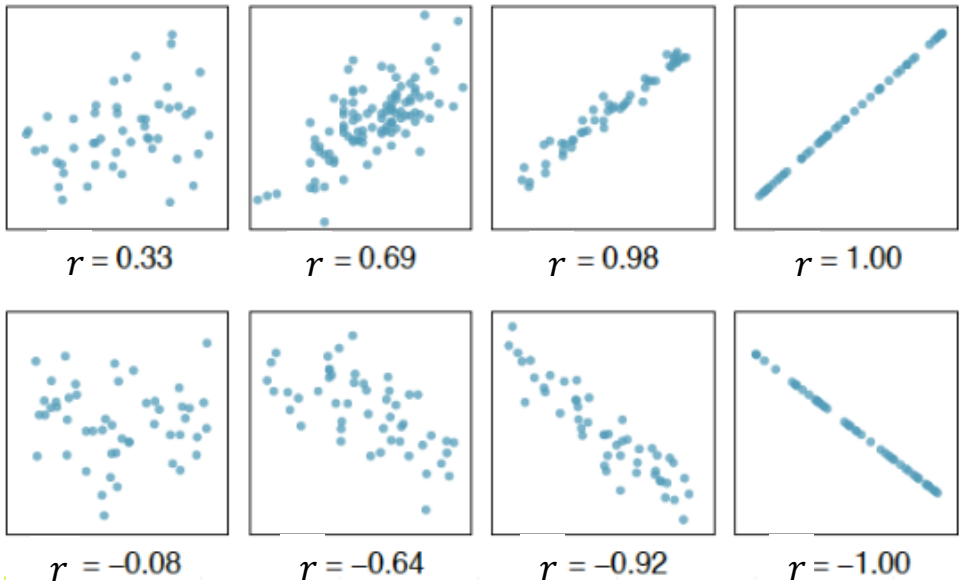


Imagen: RPubs

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 o tabla de contingencia
- 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**
  - $h_{v_i} = \frac{f_{v_i}}{N \text{ ó } n}$
  - $h_{w_j} = \frac{f_{w_j}}{N \text{ ó } n}$

X/Y	w_1	w_2	…	w_j	…	w_l	f_X	h_X
v_1	$f_{1,1}$	$f_{1,2}$	…	$f_{1,j}$	…	$f_{1,l}$	f_{v_1}	h_{v_1}
v_2	$f_{2,1}$	$f_{2,2}$	…	$f_{2,j}$	…	$f_{2,l}$	f_{v_2}	h_{v_2}
…	…	…	…	…	…	…	…	…
v_i	$f_{i,1}$	$f_{i,2}$	…	$f_{i,j}$	…	$f_{i,l}$	f_{v_i}	h_{v_i}
…	…	…	…	…	…	…	…	…
v_k	$f_{k,1}$	$f_{k,2}$	…	$f_{k,j}$	…	$f_{k,l}$	f_{v_k}	h_v
f_Y	f_{w_1}	f_{w_2}	…	f_{w_j}	…	f_{w_l}	N ó n	
h_Y	h_{w_1}	h_{w_2}	…	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 o clases: $c_i = [a_i, b_i)$
- Para realizar cálculos se utiliza el valor medio del intervalo, denominado marca de clase
  - $m_i = \frac{a_i + b_i}{2}$

X/Y	w_1	w_2	…	w_j	…	w_l	f_X	h_X
$c_1 = [a_1, b_1)$ m_1	$f_{1,1}$	$f_{1,2}$	…	$f_{1,j}$	…	$f_{1,l}$	f_{c_1}	h_{c_1}
$c_2 = [a_2, b_2)$ m_2	$f_{2,1}$	$f_{2,2}$	…	$f_{2,j}$	…	$f_{2,l}$	f_{c_2}	h_{c_2}
…	…	…	…	…	…	…	…	…
$c_i = [a_i, b_i)$ m_i	$f_{i,1}$	$f_{i,2}$	…	$f_{i,j}$	…	$f_{i,l}$	f_{c_i}	h_{c_i}
…	…	…	…	…	…	…	…	…
$c_k = [a_k, b_k]$ m_k	$f_{k,1}$	$f_{k,2}$	…	$f_{k,j}$	…	$f_{k,l}$	f_{c_k}	h_{c_k}
f_Y	f_{w_1}	f_{w_2}	…	f_{w_j}	…	f_{w_l}	N ó n	
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 (I)

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

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

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

6. Estadística descriptiva bidimensional
Tabla de contingencia. Medidas marginales (II)

- Cuantiles marginales

Medida	Poblacional	Muestral
Cuantil marginal p de x (ó X)	Siendo v_i el primer valor que cumple: $F_{v_i} \geq p_x \cdot N$ Si $F_{v_i} > p_x \cdot N$: $q_{p_x} = v_i$ Si $F_{v_i} = p_x \cdot N$: $q_{p_x} = \frac{v_i + v_{i+1}}{2}$ Donde $F_{v_i} = \sum_{r=1}^{i} f_{v_i}$	Siendo v_i el primer valor que cumple: $F_{v_i} \geq p_X \cdot n$ Si $F_{v_i} > p_X \cdot n$: $\hat{q}_{p_X} = v_i$ Si $F_{v_i} = p_X \cdot n$: $\hat{q}_{p_X} = \frac{v_i + v_{i+1}}{2}$ Donde $F_{v_i} = \sum_{r=1}^{i} f_{v_i}$

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

6. Estadística descriptiva bidimensional
Medidas marginales para datos agrupados

Medida	Poblacional	Muestral
Media marginal de x (ó X)	$\mu_x = \frac{\sum_{i=1}^{k} f_{c_i} m_i}{N}$	$\bar{X} = \frac{\sum_{i=1}^{k} f_{c_i} m_i}{n}$
Varianza marginal de x (ó X)	${\sigma_x}^2 = \frac{\sum_{i=1}^{k} f_{c_i} (m_i - \mu_x)^2}{N}$	${s_X}^2 = \frac{\sum_{i=1}^{k} f_{c_i} (m_i - \bar{X})^2}{n-1}$
Cuantil marginal p de x (ó X)	Siendo $[a_i, b_i)$ el primer intervalo o clase c_i que cumple: $F_{c_i} \geq p_x \cdot N$ $q_{p_x} = a_i + (b_i - a_i) \cdot \frac{p_x \cdot N - F_{c_{i-1}}}{f_{c_i}}$ Donde $F_{c_i} = \sum_{r=1}^{i} f_{c_i}$	Siendo $[a_i, b_i)$ el primer intervalo o clase c_i que cumple: $F_{c_i} \geq p_X \cdot n$ $\hat{q}_{p_X} = a_i + (b_i - a_i) \cdot \frac{p_X \cdot n - F_{c_{i-1}}}{f_{c_i}}$ Donde $F_{c_i} = \sum_{r=1}^{i} f_{c_i}$

X/Y	w_1	w_2	…	w_j	…	w_l	f_X	h_X
$c_1 = [a_1, b_1)$ m_1	$f_{1,1}$	$f_{1,2}$	…	$f_{1,j}$	…	$f_{1,l}$	f_{c_1}	h_{c_1}
$c_2 = [a_2, b_2)$ m_2	$f_{2,1}$	$f_{2,2}$	…	$f_{2,j}$	…	$f_{2,l}$	f_{c_2}	h_{c_2}
…	…	…	…	…	…	…	…	…
$c_i = [a_i, b_i)$ m_i	$f_{i,1}$	$f_{i,2}$	…	$f_{i,j}$	…	$f_{i,l}$	f_{c_i}	h_{c_i}
…	…	…	…	…	…	…	…	…
$c_k = [a_k, b_k]$ m_k	$f_{k,1}$	$f_{k,2}$	…	$f_{k,j}$	…	$f_{k,l}$	f_{c_k}	h_{c_k}
f_Y	f_{w_1}	f_{w_2}	…	f_{w_j}	…	f_{w_l}	N ó n	
h_Y	h_{Y_1}	h_{Y_2}	…	h_{Y_j}	…	h_{Y_l}		1

6. Estadística descriptiva bidimensional
Tabla de contingencia. Medidas marginales. Ejemplo (I)

- 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: $20.5, 25.5, 30\ años$
- Media, varianza y mediana marginales de la edad de las personas de la muestra
  - $\bar{X} = \frac{(3)(20.5)+(3)(25.5)+(4)(30)}{10} = 25.8\ años$
  - $s^2 = \frac{(3)(20.5-25.8)^2+(3)(25.5-25.8)^2+(4)(30-25.8)^2}{10-1} = 17.2\ años^2$

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 marginales. Ejemplo (II)

- Primer cuartil marginal de la edad
  - $F_{c_1} = f_{c_1} = 3 \geq 0.25 \cdot 10 = 2.5 \rightarrow i = 1$
  - $\hat{Q}_{1_X} = \hat{q}_{0.25_X} = 18 + (23 - 18) \cdot \frac{0.25 \cdot 10 - 0}{3} = 22.2\ años$
- Mediana marginal de la edad
  - $F_{c_2} = f_{c_1} + f_{c_2} = 3 + 3 = 6 \geq 0.5 \cdot 10 = 5 \rightarrow i = 2$
  - $\widehat{M}_X = \hat{q}_{0.5_X} = 23 + (28 - 23) \cdot \frac{0.5 \cdot 10 - 3}{3} = 24.3\ años$
- Tercer cuartil marginal de la edad
  - $F_{c_1} = f_{c_1} + f_{c_2} + f_{c_3} = 3 + 3 + 4 = 10 \geq 0.75 \cdot 10 = 7.5 \rightarrow i = 3$
  - $\hat{Q}_{3_X} = \hat{q}_{0.75_X} = 28 + (32 - 28) \cdot \frac{0.75 \cdot 10 - 6}{4} = 29.9\ años$
- Rango intercuartílico marginal de la edad
  - $\widehat{IQR}_X = \hat{Q}_{3_X} - \hat{Q}_{1_X} = 29.9 - 22.2 = 7.7\ años$

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 (I)

- 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	$\mu_y \mid v_i = \frac{\sum_{j=1}^{l} f_{i,j} w_j}{f_{v_i}}$	$\bar{Y} \mid v_i = \frac{\sum_{j=1}^{l} f_{i,j} w_j}{f_{v_i}}$
Varianza condicionada de x (ó X) cuando y (ó Y) = w_j	$\sigma_x{}^2 \mid w_j = \frac{\sum_{i=1}^{k} f_{i,j} \left(v_i - \mu_x \mid w_j\right)^2}{f_{w_j}}$	$s_X{}^2 \mid w_j = \frac{\sum_{i=1}^{k} f_{i,j} \left(v_i - \bar{X} \mid w_j\right)^2}{f_{w_j} - 1}$
Varianza condicionada de y (ó Y) cuando x (ó X) = v_i	$\sigma_y{}^2 \mid v_i = \frac{\sum_{j=1}^{l} f_{i,j} \left(w_j - \mu_y \mid v_i\right)^2}{f_{v_i}}$	$s_Y{}^2 \mid v_i = \frac{\sum_{j=1}^{l} f_{i,j} \left(w_j - \bar{Y} \mid v_i\right)^2}{f_{v_i} - 1}$

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

6. Estadística descriptiva bidimensional
Tabla de contingencia. Medidas condicionadas (II)

- Cuantiles condicionados

Medida	Poblacional	Muestral
Cuantil p de x (ó X) cuando y (ó Y) = w_j	Siendo v_i el primer valor que cumple: $F_{i,j} \geq p_x \cdot f_{w_j}$ Si $F_{i,j} > p_x \cdot f_{w_j}$: $q_{p_x} \mid w_j = v_i$ Si $F_{i,j} = p_x \cdot f_{w_j}$: $q_{p_x} \mid w_j = \frac{v_i + v_{i+1}}{2}$ Donde $F_{i,j} = \sum_{r=1}^{i} f_{r,j}$	Siendo v_i el primer valor que cumple: $F_{i,j} \geq p_X \cdot f_{w_j}$ Si $F_{i,j} > p_X \cdot f_{w_j}$: $\hat{q}_{p_X} \mid w_j = v_i$ Si $F_{i,j} = p_X \cdot f_{w_j}$: $\hat{q}_{p_X} \mid w_j = \frac{v_i + v_{i+1}}{2}$ Donde $F_{i,j} = \sum_{r=1}^{i} f_{r,j}$

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

6. Estadística descriptiva bidimensional
Medidas condicionadas para datos agrupados

Medida	Poblacional	Muestral
Media condicionada de x (ó X) cuando y (ó Y) = w_j	$\mu_x \left\| w_j = \frac{\sum_{i=1}^{k} f_{i,j} m_i}{f_{w_j}}$	$\bar{X} \left\| w_j = \frac{\sum_{i=1}^{k} f_{i,j} m_i}{f_{w_j}}$
Varianza condicionada de x (ó X) cuando y (ó Y) = w_j	$\sigma_x{}^2 \left\| w_j = \frac{\sum_{i=1}^{k} f_{i,j} \left(m_i - \mu_x \left\| w_j\right.\right)^2}{f_{w_j}}$	$s_X{}^2 \left\| w_j = \frac{\sum_{i=1}^{k} f_{i,j} \left(m_i - \bar{X} \left\| w_j\right.\right)^2}{f_{w_j} - 1}$
Cuantil condicionado p de x (ó X) cuando y (ó Y) = w_j	Siendo $[a_i, b_i)$ el primero que cumple: $F_{i.j} \geq p_x \cdot f_{w_j}$ $q_{p_x} \left\| w_j \right.$ $= a_i + (b_i - a_i) \cdot \frac{p_x \cdot f_{w_j} - F_{i-1,j}}{f_{i,j}}$ Donde $F_{i,j} = \sum_{r=1}^{i} f_{r,j}$	Siendo $[a_i, b_i)$ el primero que cumple: $F_{i,j} \geq p_X \cdot f_{w_j}$ $\hat{q}_{p_X} \left\| w_j \right.$ $= a_i + (b_i - a_i) \cdot \frac{p_X \cdot f_{w_j} - F_{i-1,j}}{f_{i,j}}$ Donde $F_{i,j} = \sum_{r=1}^{i} f_{r,j}$

X/Y	w_1	w_2	…	w_j	…	w_l	f_X	h_X
$c_1 = [a_1, b_1)$ m_1	$f_{1,1}$	$f_{1,2}$	…	$f_{1,j}$	…	$f_{1,l}$	f_{c_1}	h_{c_1}
$c_2 = [a_2, b_2)$ m_2	$f_{2,1}$	$f_{2,2}$	…	$f_{2,j}$	…	$f_{2,l}$	f_{c_2}	h_{c_2}
…	…	…	…	…	…	…	…	…
$c_i = [a_i, b_i)$ m_i	$f_{i,1}$	$f_{i,2}$	…	$f_{i,j}$	…	$f_{i,l}$	f_{c_i}	h_{c_i}
…	…	…	…	…	…	…	…	…
$c_k = [a_k, b_k]$ m_k	$f_{k,1}$	$f_{k,2}$	…	$f_{k,j}$	…	$f_{k,l}$	f_{c_k}	h_{c_k}
f_Y	f_{w_1}	f_{w_2}	…	f_{w_j}	…	f_{w_l}	N ó n	
h_Y	h_{Y_1}	h_{Y_2}	…	h_{Y_j}	…	h_{Y_l}		1

6. Estadística descriptiva bidimensional
Tabla de contingencia. Medidas condicionadas. Ejemplo (I)

- 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
Tabla de contingencia. Medidas condicionadas. Ejemplo (II)

- Primer cuartil de la edad de las personas con Windows
  - $F_1 = 2 \geq 0.25 \cdot 5 = 1.25 \rightarrow i = 1$
  - $\hat{Q}_{1_X}|Win = \hat{q}_{0.25_X}|Win = 18 + (23 - 18) \cdot \frac{0.25 \cdot 5 - 0}{2} = 21.1\ años$
- Mediana de la edad de las personas con Windows
  - $F_2 = 3 \geq 0.5 \cdot 5 = 2.5 \rightarrow i = 2$
  - $\hat{M}_X|Win = \hat{q}_{0.5_X}|Win = 23 + (28 - 23) \cdot \frac{0.5 \cdot 5 - 2}{1} = 25.5\ años$
- Tercer cuartil de la edad de las personas con Windows
  - $F_3 = 5 \geq 0.75 \cdot 5 = 3.75 \rightarrow i = 3$
  - $\hat{Q}_{3_X}|Win = \hat{q}_{0.75_X}|Win = 28 + (32 - 28) \cdot \frac{0.75 \cdot 5 - 3}{2} = 29.9\ años$
- Rango intercuartílico de las personas con Windows
  - $\widehat{IQR}_X|Win = \hat{Q}_{3_X}|Win - \hat{Q}_{1_X}|Win = 29.9 - 21.1 = 8.8\ años$

Windows			
Fila	X	f	F
1	$[18, 23)$ **20.5**	2	2
2	$[23, 28)$ **25.5**	1	3
3	$[28, 32]$ **30**	2	5

6. Estadística descriptiva bidimensional Independencia de variables

- Dos variables estadísticas X e Y son estadísticamente independientes cuando el valor de una de ellas no se ve afectado por los valores que toma la otra
- En ese caso se cumple que las frecuencias relativas condicionadas no se ven afectadas por la condición, y coinciden con las frecuencias relativas marginales
- Por tanto, para que X e Y sean independientes se cumple:
  - Datos no agrupados
    - $\frac{f_{i,j}}{N} = h_{v_i} \cdot h_{w_j}$ para todo i y j
  - Datos agrupados para X
    - $\frac{f_{i,j}}{N} = h_{c_i} \cdot h_{w_j}$ para todo i y j

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

6. Estadística descriptiva bidimensional Independencia de variables. Ejemplos

- Independientes

X/Y	$Rojo$	$Blanco$	f_X	h_X
$Ford$	2	6	8	0.4
$Renault$	3	9	12	0.6
f_Y	5	15	20	
h_Y	0.25	0.75		1

$$\frac{f_{1,1}}{20} = \frac{2}{20} = 0.1 = 0.4 \cdot 0.25 = 0.1$$

...

$$\frac{f_{2,2}}{20} = \frac{9}{20} = 0.45 = 0.6 \cdot 0.75 = 0.45$$

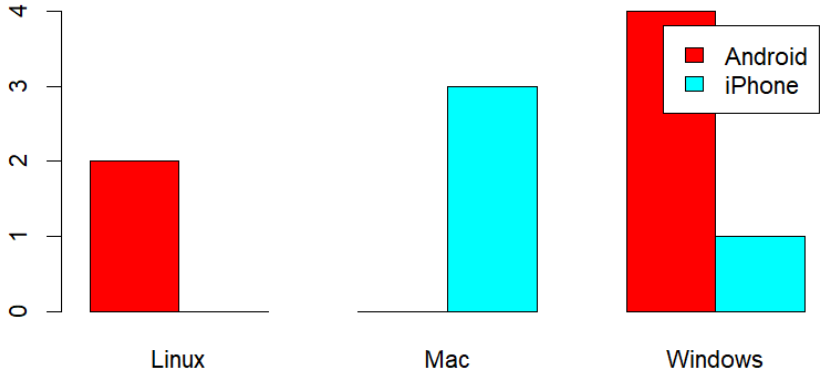
- No independientes

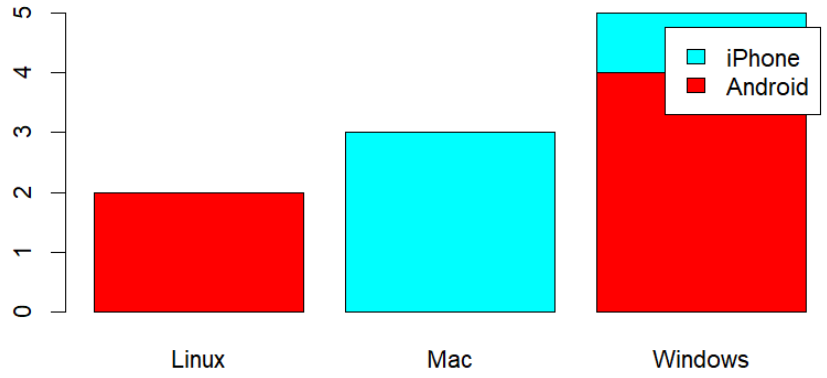
X/Y	$Windows$	Mac	$Linux$	f_X	h_X
$Android$	4	0	2	6	0.6
$iPhone$	1	3	0	4	0.4
f_Y	5	3	2	10	
h_Y	0.5	0.3	0.2		1

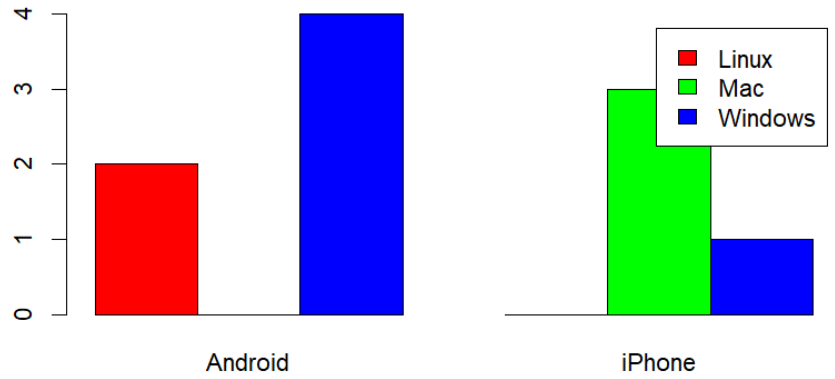
$$\frac{f_{1,1}}{10} = \frac{4}{10} = 0.4 \neq 0.6 \cdot 0.5 = 0.3$$

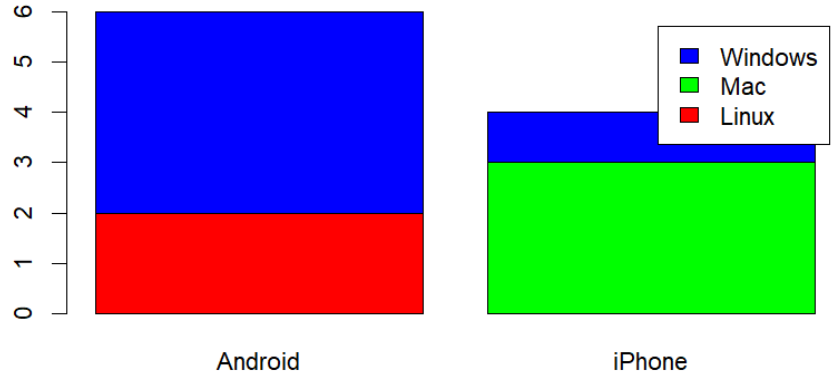
6. Estadística descriptiva bidimensional Gráficos. Diagrama de barras

- En el eje horizontal se representan los valores de una variable
- Se dibuja una barra para cada valor de la otra variable, sobre cada valor de la primera (o en versión apilada)
- En el eje vertical se representan las frecuencias conjuntas









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)

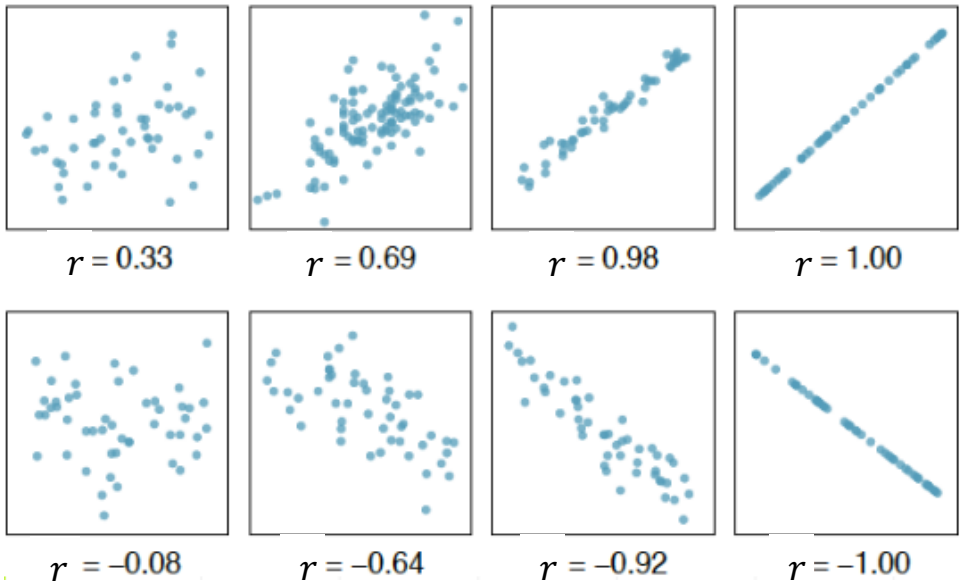


Imagen: RPubs

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

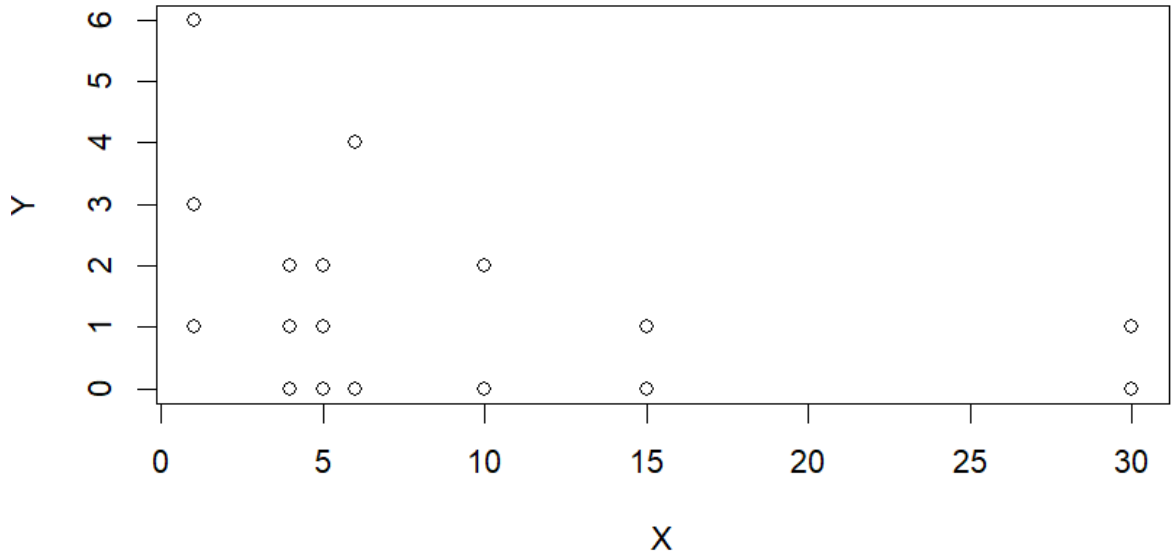
- Un administrador informático registra el número de veces que un software antivirus se ejecuta en cada ordenador de una empresa durante un mes (variable X) y el número de virus detectados (variable Y).
- Los datos de una muestra 30 ordenadores están en la siguiente tabla

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)

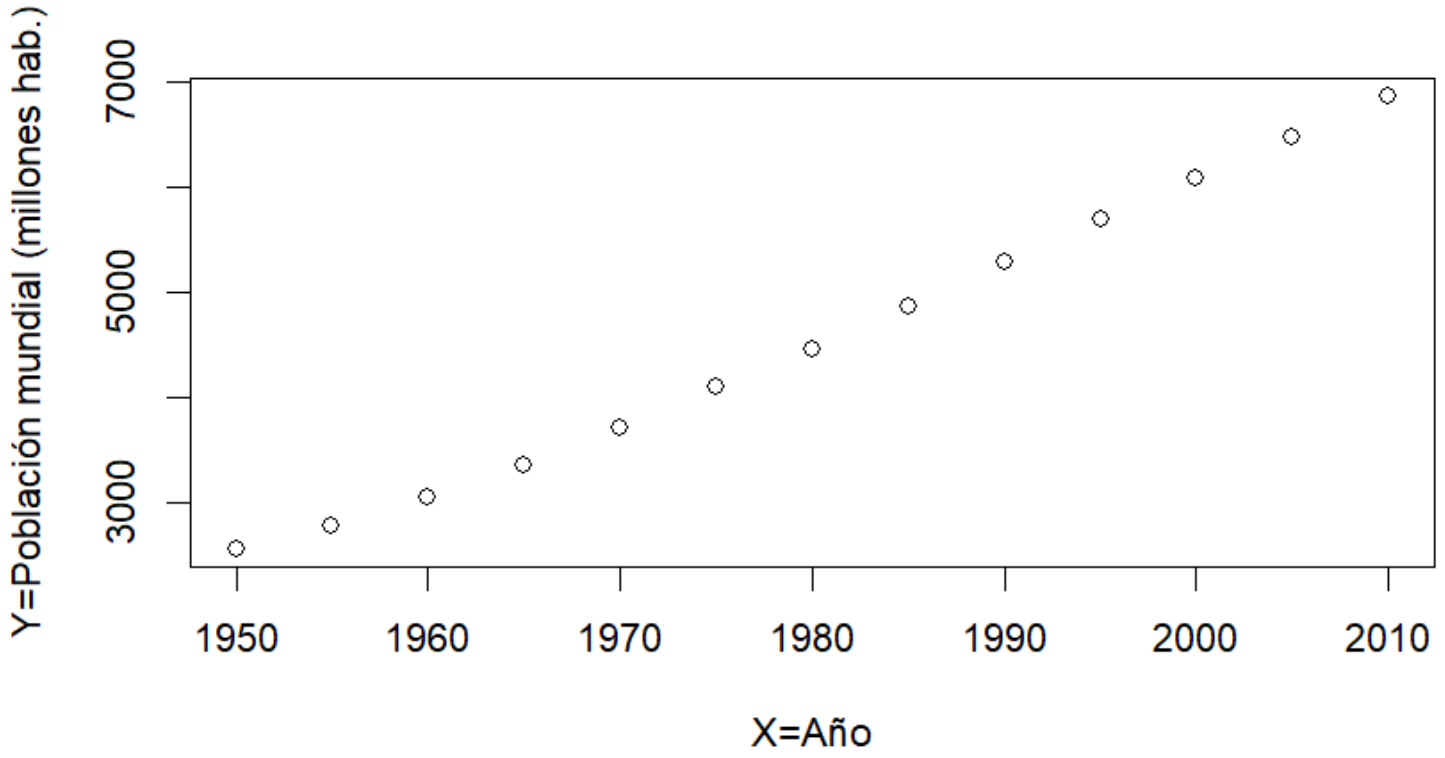
- Según la Base Internacional de Datos de la Oficina del Censo de los Estados Unidos, la población mundial entre 1950 y 2010 creció según esta tabla, en la que:
  - X representa un año
  - Y representa la población mundial en ese año (millones de personas)

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

- Otros ejercicios (resueltos): wpd.ugr.es

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