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.

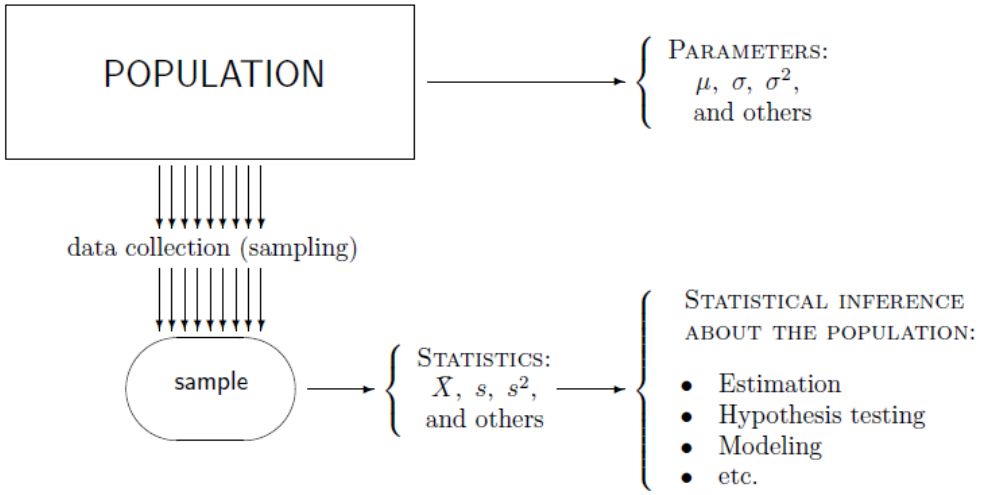


FIGURE 8.1: *Population parameters and sample statistics.*

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 numéros
- 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**

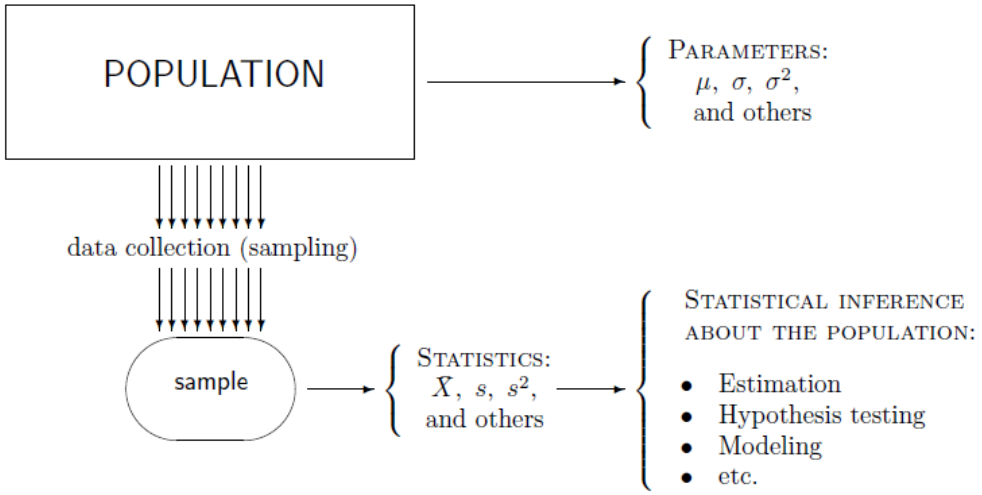


FIGURE 8.1: *Population parameters and sample statistics.*

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 la distribución de los valores es simétrica
  - Hay un número similar de valores inferiores y superiores a la media
- Si la mediana es menor que la media, la distribución de los valores es right-skewed
  - Asimétrica (sesgada) hacia la derecha
  - Hay más valores por debajo de la media
- Si la mediana es menor que la media, la distribución de los valores es left-skewed
  - Asimétrica (sesgada) hacia la izquierda
  - Hay más valores por encima de la media

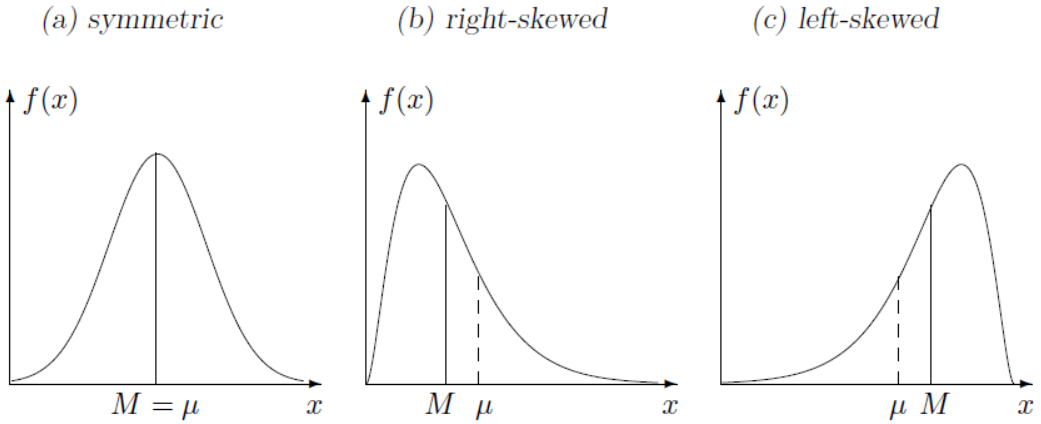


FIGURE 8.2: *A mean μ and a median M for distributions of different shapes.*

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 que la dividen en partes iguales, es decir en intervalos (cuantiles) 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)

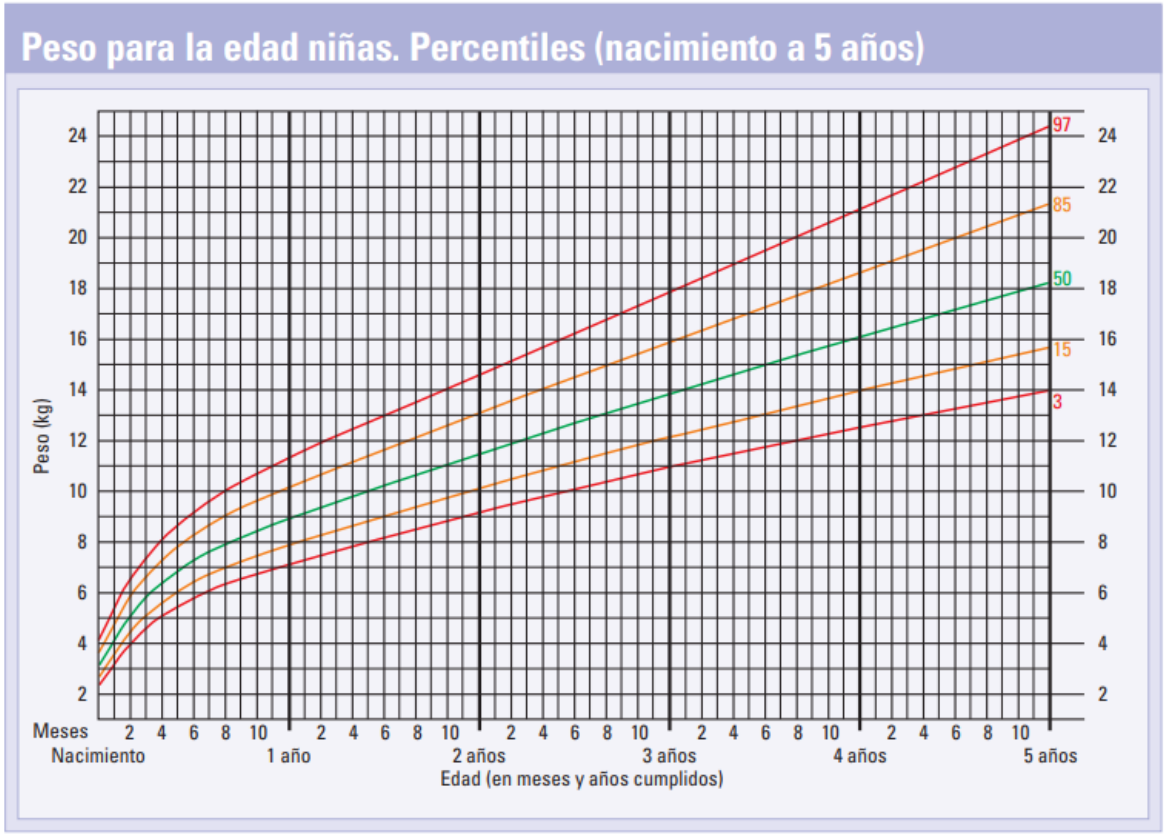
Medida	Poblacional	Muestral
Primer cuartil	$Población\ ordenada = (x_1, x_2, .., x_N)$ Si $N/4$ no es un número entero: $Q_1 = x_{\lfloor N/4 \rfloor + 1}$ Si $N/4$ es un número entero: $Q_1 = \frac{x_{N/4} + x_{N/4+1}}{2}$	$Muestra\ ordenada = (X_1, X_2, .., X_n)$ Si $n/4$ no es un número entero: $\hat{Q}_1 = X_{\lfloor N/4 \rfloor + 1}$ 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
  - $[Q_1 - 1.5 \cdot IQR,\ Q_3 + 1.5 \cdot IQR]$

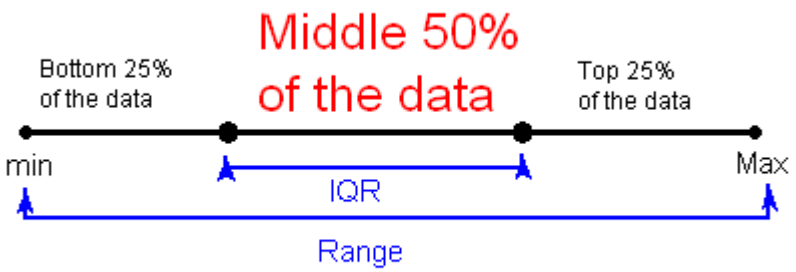
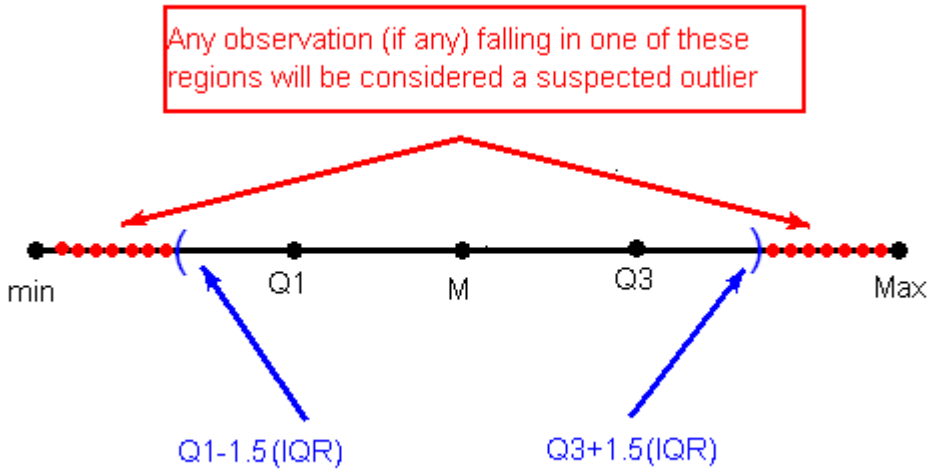


Imagen: libretexts.org



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
- 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
  - $\widehat{IQR} = \hat{Q}_3 - \hat{Q}_1 = 59 - 34 = 25\ seg$

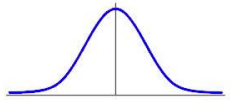
- f) Detectar si hay datos atípicos
  - $\left[\hat{Q}_1 - 1.5 \cdot \widehat{IQR}, \hat{Q}_3 + 1.5 \cdot \widehat{IQR}\right] = [34 - 1.5 \cdot 25, 59 + 1.5 \cdot 25] = [-3.5, 96.5]$
  - El valor 139 seg es atípico porque está fuera del intervalo

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 la frecuencia de los valores de las observaciones, normalmente un histograma.
- 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 histograma se ajusta aproximadamente a la curva conocida como campana de Gauss:
  - $f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$ 
- 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)

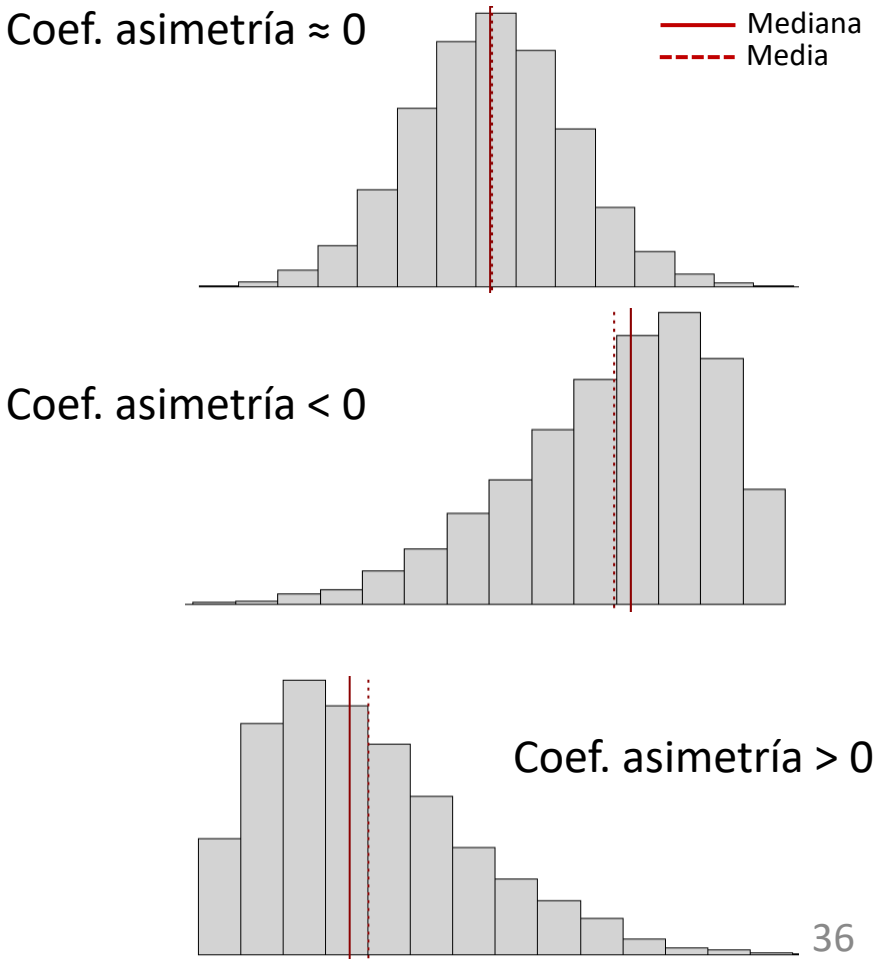
Medida	Poblacional	Muestral
Sesgo o coeficiente de asimetría	$A = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^3}{\sigma^3}$	$\hat{A} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3}{s^3}$
Curtosis o coeficiente de apuntamiento	$K = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^4}{\sigma^4} - 3$	$\widehat{K}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

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

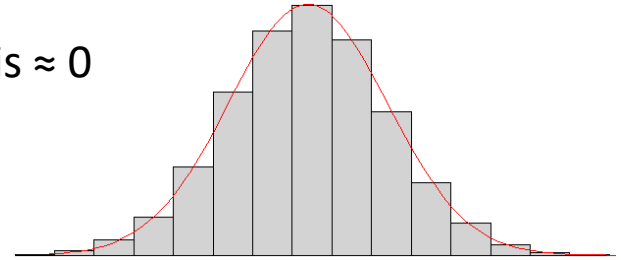


3. Cálculo de medidas estadísticas Medidas de forma (interpretación)(II)

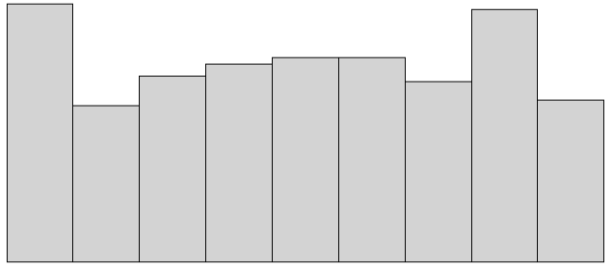
Curtosis (Coeficiente de apuntamiento)

- 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

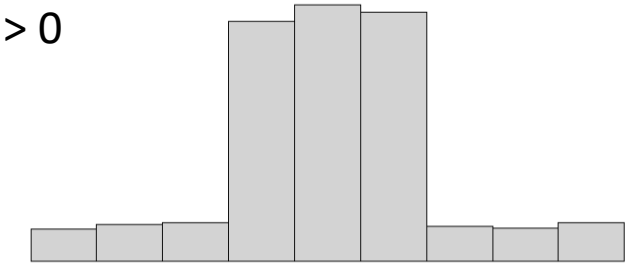
Curtosis ≈ 0



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)

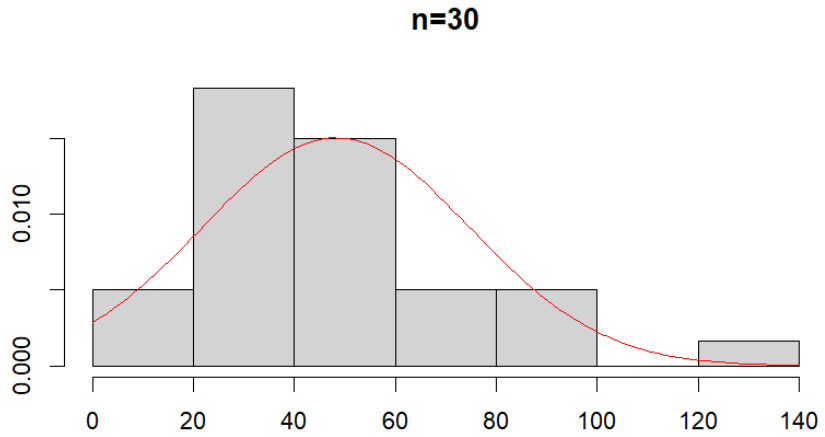
3. Cálculo de medidas estadísticas Medidas de forma. Ejemplo 8.12 (solución)(I)

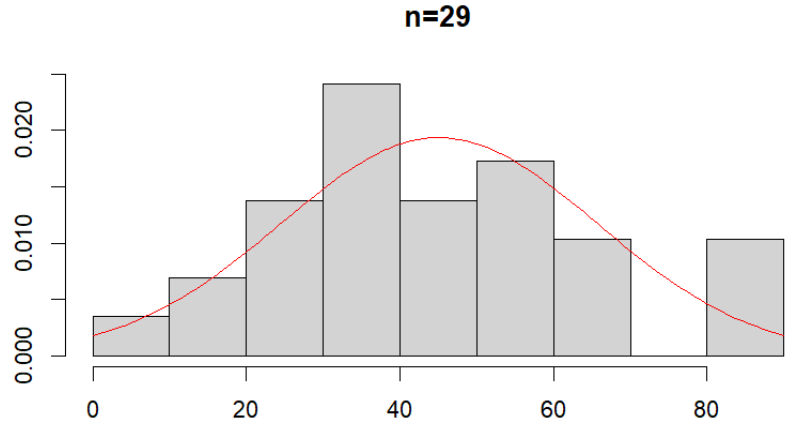
- $S = (9, 15, 19, 22, 24, 25, 30, 34, 35, 35, 36, 36, 37, 38, 42, 43, 46, 48, 54, 55, 56, 56, 59, 62, 69, 70, 82, 82, 89, 139)$
- $n = 30, \bar{X} = 48.23s, s = 26.52s$
- a) Sesgo o coeficiente de asimetría
  - $\hat{A} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3}{s^3} = \frac{\frac{1}{30}\sum_{i=1}^{30}(X_i - 48.23)^3}{26.52^3} = \frac{\frac{1}{30}(9-48.23)^3 + \cdots + (139-48.23)^3}{26.52^3} = 1.31$
  - 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)

- g) 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
Propoción. 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