Tema 1. Descripción de una variable.

Eva Romero Ramos
evarom03@ucm.es

Universidad Complutense de Madrid

Outline

Outline

Conceptos básicos

- **Frecuencia absoluta (n_i)**: el número de veces que se repite un valor en la muestra.
- **Frecuencia Relativa (f_i)**: cociente entre la frecuencia absoluta y el tamaño muestral (N).
- **Frecuencia absoluta acumulada (N_i)**: el número de observaciones con valore menor o igual al valor considerado.
- **Frecuencia relativa acumulada (F_i)**: cociente entre la frecuencia absoluta acumulada y el tamaño muestral.

La suma de todas las frecuencias relativas es 1.

La última frecuencia relativa acumulada es igual a 1.

Distribución de frecuencias de una variable

x_i	n_i	f_i	N_i	F_i
300	20	$20/200 = 0.1$	20	0.1
600	40	$40/200 = 0.2$	60	0.3
900	60	$60/200 = 0.3$	120	0.6
1200	50	$50/200 = 0.25$	170	0.85
1500	30	$30/200 = 0.15$	200	1
	$N = 200$	1		

Distribución de frecuencias de una variable

x_i	n_i	f_i	N_i	F_i
300	20	$20/200 = 0.1$	20	0.1
600	40	$40/200 = 0.2$	60	0.3
900	60	$60/200 = 0.3$	120	0.6
1200	50	$50/200 = 0.25$	170	0.85
1500	30	$30/200 = 0.15$	200	1
	$N = 200$	1		

Dos distribuciones son iguales si todas sus x_i y sus frecuencias relativas f_i son iguales.

Organización del conjunto de datos

- Distribución de datos sin agrupar
- Datos agrupados: se agrupan los valores en **intervalos** or **clases**.
  - Tenemos la máxima información del conjunto de datos cuando se no se agrupan. Con los agrupamientos se pierde información.
  - Los intervalos son *arbitrarios*, es decir, que son definidos por el investigador.

Distribución de frecuencias con datos agrupados

- L_i: Límite superior del intervalo.
- L_{i-1}: Límite inferior del intervalo.
- Rango de la variable: $\boxed{Re = \max_i x_i - \min_i x_i}$
- Amplitud del intervalo: $\boxed{c_i = L_i - L_{i-1}}$
  - Puede ser constante constante (más fácil de manejar) o variable.
  - Si es constante, se puede calcular el número de intervalos fijando la amplitud, o la amplitud fijando el número de intervalos, teniendo en cuenta que:

$$\boxed{Re = \text{Número de intervalos} \cdot c_i}$$

Frequency distribution with grouped data

- Para crear los intervalos debemos tener en cuenta que cada dato debe estar necesariamente en un solo intervalo, es decir, que un mismo dato no puede estar en dos intervalos a la vez y debe aparecer en alguno.
- Denominaremos $L_i - 1$ al límite inferior del intervalo y L_i al límite superior.
- **Marca de clase (x_i)**: es el punto de referencia de cada intervalo que se usará en el cálculor de medidas para las que sea necesario.

$$x_i = \frac{L_{i-1} + L_i}{2}$$

Ejercicio 1

Un inversor pretende invertir en un activo. Antes de hacerlo observa el histórico de rendimientos para hacerse una idea que cómo será la inversión, para los que obtiene la siguiente muestra de los rendimientos del activo durante 25 días:

-2.5	0.6	-0.1	-1.2	-2.5
1.7	0.6	-1.7	0.9	-3.9
0.6	1.7	0.0	2.3	0.6
2.8	1.2	0.9	1.7	0.6
-1.7	0.6	-0.9	-3.6	-1.7

Ejercicio 1

(a) Obtenga la distribución de frecuencias sin agrupar. (b) Obtenga la distribución de frecuencias agrupada en intervalos de longitud 1%, y calcule la marca de clase de cada intervalo.

Ejercicio 1a

(a) Obtenga la distribución de frecuencias sin agrupar.

x_i	n_i	N_i	f_i	F_i	x_i	n_i	N_i	f_i	F_i
-3.9	1	1	0.04	0.04	0.6	6	17	0.24	0.68
-3.6	1	2	0.04	0.08	0.9	2	19	0.08	0.76
-2.5	2	4	0.08	0.16	1.2	1	20	0.04	0.80
-1.7	3	7	0.12	0.28	1.7	3	23	0.12	0.92
-1.2	1	8	0.04	0.32	2.3	1	24	0.04	0.96
-0.9	1	9	0.04	0.36	2.8	1	25	0.04	1
-0.1	1	10	0.04	0.40		25		1	
0.0	1	11	0.04	0.44					

Ejercicio 1b

(b) Obtenga la distribución de frecuencias agrupada en intervalos de longitud 1%, y calcule la marca de clase de cada intervalo.

L_{i-1}	L_i	x_i	n_i	N_i	f_i	F_i
-4	-3	-3.5	2	2	0.08	0.08
-3	-2	-2.5	2	4	0.08	0.16
-2	-1	-1.5	4	8	0.16	0.32
-1	0	-0.5	3	11	0.12	0.44
0	1	0.5	8	19	0.32	0.76
1	2	1.5	4	23	0.16	0.92
2	3	2.5	2	25	0.08	1
			25		1	

Outline

Diagrama de sectores

El diagrama de sectores es una circunferencia sobre la que se representan cada categoría como un sector de dicha circunferencia, de forma que tanto el ángulo como el área contenida en dicho sector es proporcional a la frecuencia correspondiente.

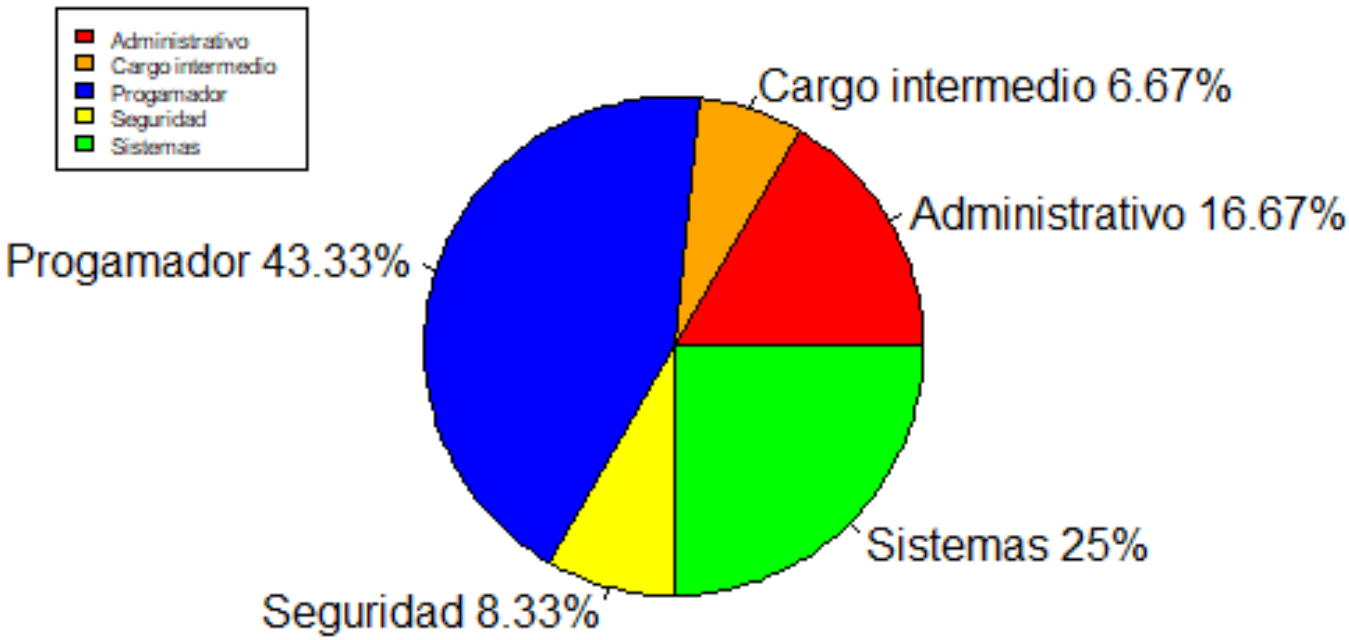
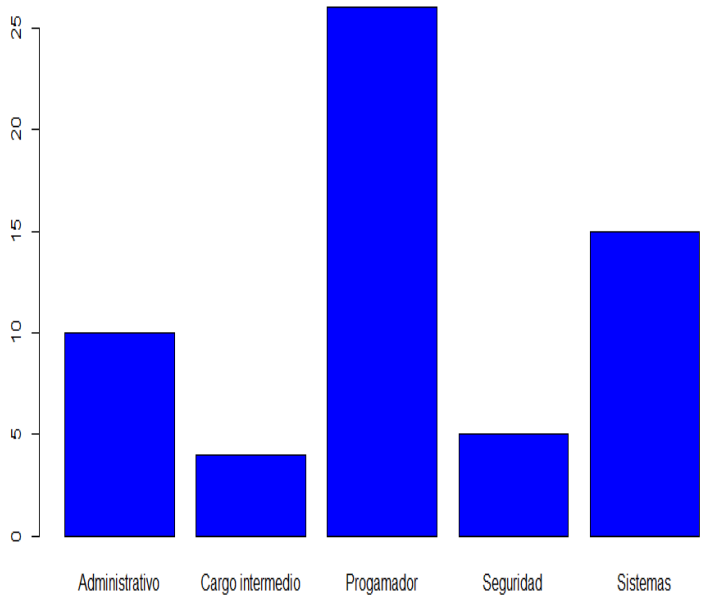


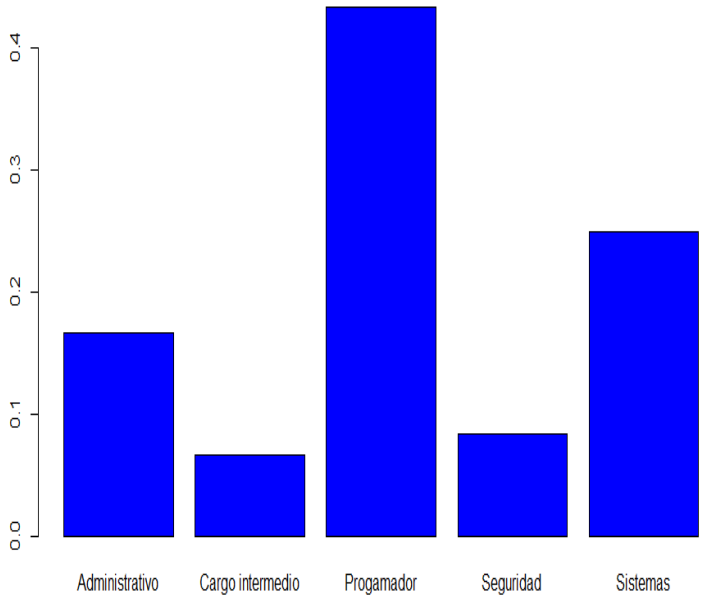
Diagrama de barras

El diagrama de barras es un gráfico que representa sobre el horizontal las diferentes categorías o valores de la variable y para cada una de ellas levanta una barra de longitud igual a la frecuencia absoluta o relativa. De este modo la altura de las barra se puede apreciar sobre el eje y.

Con frecuencias absolutas

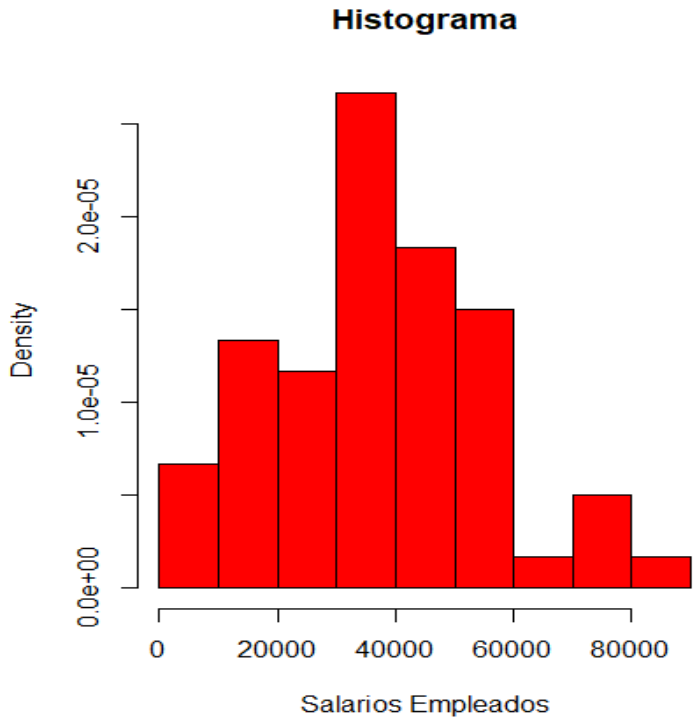


Con frecuencias relativas



Histograma

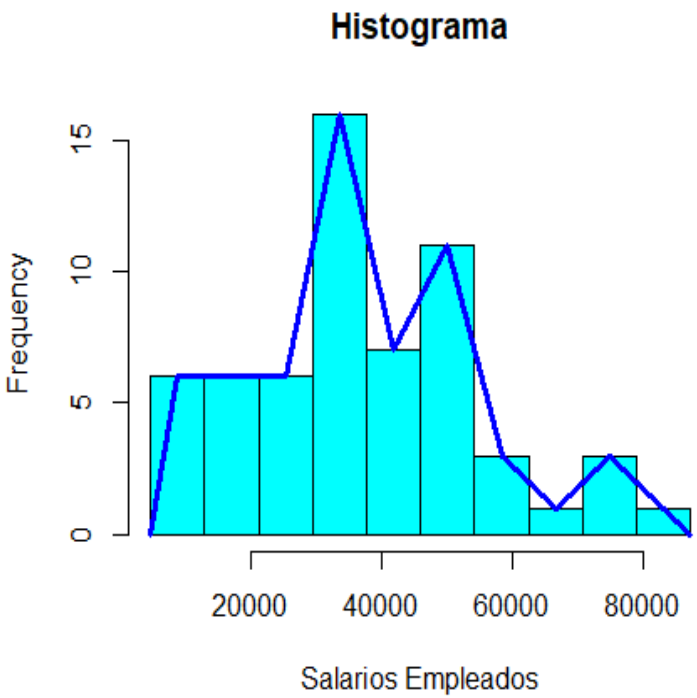
- Para realizar un histograma se deben agrupar los datos en intervalos.
- Se sitúan los diferentes intervalos en el eje x y se levanta sobre cada uno un rectángulo cuya base es el intervalo y cuyo área es proporcional a la frecuencia absoluta o relativa asociada al mismo.
- Como el área de cada intervalo es $A_i = c_i \cdot d_i = n_i$, entonces la altura es $d_i = n_i / c_i$.
- Si los intervalos tienen la misma anchura ($c_i = L,\ \forall i$), entonces la altura d_i es proporcional a la frecuencia.



Poligono de frecuencias

Se puede hacer con frecuencias **absolutas** o absolutas **acumuladas**.
El polígono de frecuencias se puede hacer a partir del histograma, uniendo los extremos superiores de las barras con líneas rectas, de forma que área que queda por debajo del polígono de frecuencias es igual al área contenida dentro del correspondiente histograma, si se hace con frecuencias absolutas.

Con frecuencias absolutas



Con frecuencias relativas

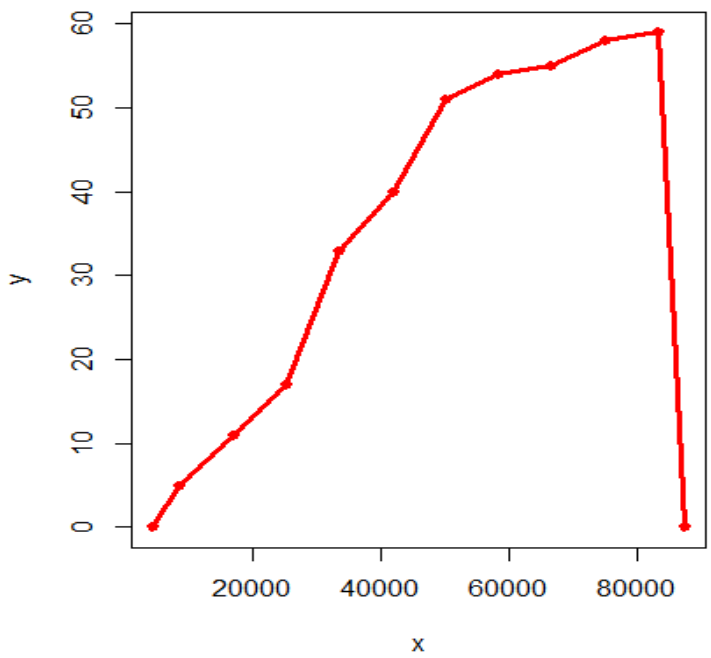
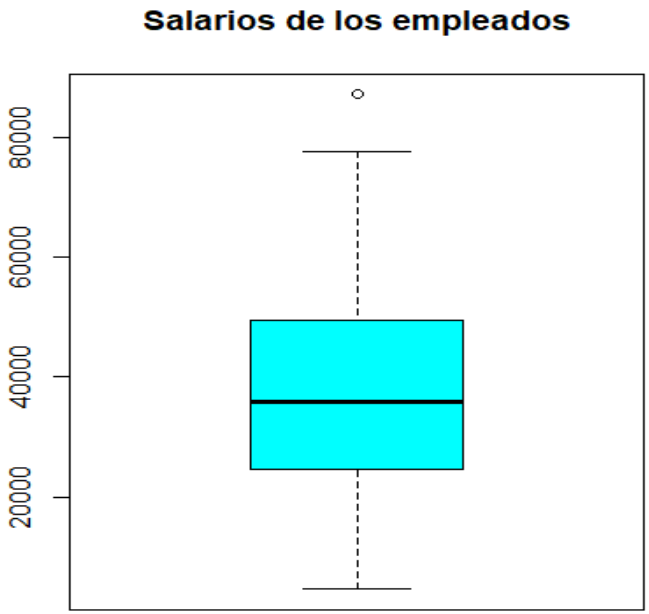


Diagrama de cajas

- El diagrama de cajas es una representación de una serie de carácterísticas importantes de la distribución como son la dispersión y la simetría.
- En la caja se situan los cuantiles.
- El final de las líneas se corresponde con el máximo entre 3 veces la desviación típica y el máximo o mínimo de la distribución.
- Si se observan valores más allá de las líneas se trata de elementos outlier.



Outline

Introducción

- La distribución de frecuencias refleja toda la información disponible, lo que en general es demasiada información. Es, por tanto, necesario resumir la información disponible.
- Existen diferentes medidas que proporcionan una descripción global de la variable.
- Son características deseables para las medidas de posición:
  - Que utilicen todas y cada una de las observaciones disponibles.
  - Que sean sencillas de calcular.
  - Que sean fáciles de interpretar.
  - Que sean únicas para cada distribución de frecuencias.

Media Aritmética: $\overline{x}$

- La media aritmética es la suma de todas las observaciones, dividida entre el tamaño muestral, es decir:

$$\overline{x} = \frac{x_1 n_1 + x_2 n_2 + \cdots + x_n n_n}{N} = \frac{1}{N} \sum_{i=1}^{n} x_i n_i$$

- La media aritmética es el valor que tendría cada observación si todas las observaciones tuviesen el mismo valor.
- Para datos agrupados, se usan las marcas de clase para calcular la media en vez de los valores individuales, que son desconocidos.

Media Aritmética: $\overline{x}$

Es la medida de posición central más precisa y resume información para observaciones medidas en escala de intervalo o de razón.

Ventajas:

- Considera toda la información disponible;
- Es fácil de calcular;
- Es única;
- Es el centro de gravedad (la suma de todas las desviaciones con respecto a ella es igual a cero);
- Si minimizamos las desviaciones al cuadrado de las observaciones respecto a una constante k el valor mínimo se obtiene cuando k es la media aritmética;
- La media de la variable $ax + b$ es $a\overline{x} + b$

Desventajas:

- En distribuciones con datos extremos (outliers), la media puede resultar poco representativa, al verse fuertemente afectada por estos valores.

Ejemplo 1: $\overline{x}$

Calcula la media aritmética de la siguiente distribución.

x_i	n_i	$x_i n_i$
300	20	6000
600	40	24000
900	60	54000
1200	50	60000
1500	30	45000
	$N = 200$	189000

Ejemplo 1: $\overline{x}$

Calcula la media aritmética de la siguiente distribución.

x_i	n_i	$x_i n_i$
300	20	6000
600	40	24000
900	60	54000
1200	50	60000
1500	30	45000
	$N = 200$	189000

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{n} x_i n_i = \frac{189000}{200} = 945$$

Mediana

- Es el valor de la distribución, ordenado de forma ascendente, que deja a la mitad de las observaciones a su izquierda y a la otra mitad a su derecha.
- La forma de obtenerla es ordenar los datos, y la mediana será el valor que ocupe la posición central si el número de observaciones es impar. Si el número de observaciones es par, entonces no existe un único valor en la posición central; entonces se suele definir la mediana como la media de los dos valores de las posiciones centrales.
- También puede definirse como el valor de la distribución cuya frecuencia absoluta acumulada es $N/2$ (o su frecuencia relativa acumulada es 50%).

Mediana

- Con los datos agrupados en intervalos, en lugar de un valor mediano hallaremos primero un intervalo mediano. Luego elegimos un valor representativo de este intervalo al que llamaremos mediana. Obtendremos este valor mediante:

$$Me = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} c_i$$

- No importa si los intervalos tiene la misma longitud o no.

Mediana

Propiedad:

- La mediana mínimiza la media del valor absoluto de las desviaciones:

$$\arg\min_{k} \sum_{i=1}^{n} |x_i - k|\, n_i = \sum_{i=1}^{n} |x_i - Me|\, n_i$$

Usos:

- La mediana es la medida de posición más representativa en el caso de datos en escala ordinal.
- Cuando la media aritmética resulta poco representativa a causa de la presencia de ouliers en la distribución, se suele tener en cuenta la mediana.

Mediana

Ventajas:

- Los valores exremos no affectan a la mediana tanto como a la media;
- Es fácil de calcular e interpretar;
- Se puede calcular incluso con datos cualitativos.

Desventajas:

- Es necesario ordenar la muestra, por lo que solo se puede usar con datos cualitativos en escala ordinal.
- No usa toda la información disponible en la muestra.

Ejemplo 2: Mediana

Calcule la mediana de la siguiente distribución.

x_i	n_i	N_i
300	20	20
600	40	60
900	60	120
1200	50	170
1500	30	200

Ejemplo 2: Mediana

Calcule la mediana de la siguiente distribución.

x_i	n_i	N_i
300	20	20
600	40	60
900	60	120
1200	50	170
1500	30	200

$$\frac{N}{2} = \frac{200}{2} = 100$$

Ejemplo 2: Mediana

Calcule la mediana de la siguiente distribución.

x_i	n_i	N_i
300	20	20
600	40	60
900	60	120
1200	50	170
1500	30	200

$$\frac{N}{2} = \frac{200}{2} = 100 \Longrightarrow \text{posición } 100 - 101$$

Ejemplo 2: Mediana

Calcule la mediana de la siguiente distribución.

x_i	n_i	N_i
300	20	20
600	40	60
900	60	120
1200	50	170
1500	30	200

$$\frac{N}{2} = \frac{200}{2} = 100 \Longrightarrow \text{posición } 100 - 101 \Longrightarrow Me = 900$$

Moda

- La **moda** es el valor con mayor frecuencia de la distribución.
- Para datos sin agrupar: Para obtener la moda en datos sin agrupar simplemente se busca el valor valor con mayor frecuencia.
- La moda puede no tener un único resultado por lo que existen distribuciones con varias modas (distribuciones multimodales)

Moda

- Para datos agrupados:
  1. *Intervalos de igual amplitud*. Buscamos el intervalo con la frecuencia más alta y elegimos en él la moda de acuerdo a alguno de los siguientes criterios:
     - Considerar el límite inferior del intervalo: $Mo = L_{i-1}$
     - Considerar el límite superior del intervalo: $Mo = L_i$
     - Considerar el punto medio del intervalo: $Mo = x_i$
     - Suponiendo que los valores del intervalo se distribuyen de forma análoga a como se distribuyen los datos en la muestra, usar:

$$Mo = L_{i-1} + \frac{n_{i+1}}{n_{i-1} + n_{i+1}} c_i$$

Moda

2. *Intervalos de distinta amplitud*. En este caso el intervalo modal no será el de mayor frecuencia sino el de mayor densidad de frecuencia.
   - Densidad de frecuencia:
   $$d_i = \frac{n_i}{c_i}$$
   - Una vez determinado el intervalo, se puede aplicar cualquiera de los criterios presentados anteriormente. El más adecuado sería:
   $$Mo = L_{i-1} + \frac{d_{i+1}}{d_{i-1} + d_{i+1}} c_i$$

Moda

Ventajas:

- Es la medida de posición más representativa para datos en escala nominal.
- Excepto en el caso de muestras muy pequeñas, la moda no se ve afectada por elementos outliers.
- Se puede calcular incluso con intervalos abiertos.

Desventajas:

- No se usa toda la información disponible en la muestra.
- A veces el hecho de que un elemento se repita más que el resto es casualidad. Por eso no es suele utilizar en el caso de variables numéricas.
- Si la distribución es multimodal se hace difícil de interpretar la moda.

Ejemplo 3: Moda

Calcula la moda de la siguiente distribución:

L_{i-1}	L_i	n_i
300	600	20
600	900	40
900	1200	60
1200	1500	50
1500	1800	30

Ejemplo 3: Moda

Calcula la moda de la siguiente distribución:

L_{i-1}	L_i	n_i
300	600	20
600	900	40
900	1200	60
1200	1500	50
1500	1800	30

Intervalo modal: $(900, 1200]$

Ejemplo 3: Moda

Calcula la moda de la siguiente distribución:

L_{i-1}	L_i	n_i
300	600	20
600	900	40
900	1200	60
1200	1500	50
1500	1800	30

Intervalo modal: $(900, 1200]$

$$Mo = L_{i-1} + \frac{n_{i+1}}{n_{i-1} + n_{i+1}} c_i = 900 + \frac{50}{40 + 50} 300 = 1066.67$$

Ejercicio 1c

(c) Obtenga la media aritmética, la mediana y la moda para la distribución de frecuencias de los datos agrupados y sin agrupar.

Ejercicio 1c

(c) Obtenga la media aritmética, la mediana y la moda para la distribución de frecuencias de los datos agrupados y sin agrupar.

Media aritmética

Ejercicio 1c

(c) Obtenga la media aritmética, la mediana y la moda para la distribución de frecuencias de los datos agrupados y sin agrupar.

Media aritmética

Sin agrupar: $\overline{x} = \frac{1}{N}\sum_{i=1}^{n} x_i n_i = \frac{-3}{25} = -0.12$

Ejercicio 1c

(c) Obtenga la media aritmética, la mediana y la moda para la distribución de frecuencias de los datos agrupados y sin agrupar.

Media aritmética

Sin agrupar: $\overline{x} = \frac{1}{N}\sum_{i=1}^{n} x_i n_i = \frac{-3}{25} = -0.12$

Agrupados: $\overline{x} = \frac{1}{N}\sum_{i=1}^{n} x_i n_i = \frac{-4.5}{25} = -0.18$

Ejercicio 1c

(c) Obtenga la media aritmética, la mediana y la moda para la distribución de frecuencias de los datos agrupados y sin agrupar.

Media aritmética

Sin agrupar: $\overline{x} = \frac{1}{N}\sum_{i=1}^{n} x_i n_i = \frac{-3}{25} = -0.12$

Agrupados: $\overline{x} = \frac{1}{N}\sum_{i=1}^{n} x_i n_i = \frac{-4.5}{25} = -0.18$

Mediana

Ejercicio 1c

(c) Obtenga la media aritmética, la mediana y la moda para la distribución de frecuencias de los datos agrupados y sin agrupar.

Media aritmética

Sin agrupar: $\overline{x} = \frac{1}{N}\sum_{i=1}^{n} x_i n_i = \frac{-3}{25} = -0.12$

Agrupados: $\overline{x} = \frac{1}{N}\sum_{i=1}^{n} x_i n_i = \frac{-4.5}{25} = -0.18$

Mediana

Sin agrupar: $\frac{N}{2} = 12.5 \Longrightarrow$ position 13 $\Longrightarrow Me = 0.6$

Ejercicio 1c

(c) Obtenga la media aritmética, la mediana y la moda para la distribución de frecuencias de los datos agrupados y sin agrupar.

Media aritmética

Sin agrupar: $\overline{x} = \frac{1}{N}\sum_{i=1}^{n} x_i n_i = \frac{-3}{25} = -0.12$

Agrupados: $\overline{x} = \frac{1}{N}\sum_{i=1}^{n} x_i n_i = \frac{-4.5}{25} = -0.18$

Mediana

Sin agrupar: $\frac{N}{2} = 12.5 \Longrightarrow$ position 13 $\Longrightarrow Me = 0.6$

Agrupar: intervalo de la mediana: $(0, 1]$, mediana:

$$Me = 0 + \frac{12.5 - 11}{8} \cdot 1 = 0.188$$

Ejercicio 1c

(c) Obtenga la media aritmética, la mediana y la moda para la distribución de frecuencias de los datos agrupados y sin agrupar.

Media aritmética

Sin agrupar: $\overline{x} = \frac{1}{N}\sum_{i=1}^{n} x_i n_i = \frac{-3}{25} = -0.12$

Agrupados: $\overline{x} = \frac{1}{N}\sum_{i=1}^{n} x_i n_i = \frac{-4.5}{25} = -0.18$

Mediana

Sin agrupar: $\frac{N}{2} = 12.5 \Longrightarrow$ position 13 $\Longrightarrow Me = 0.6$

Agrupar: intervalo de la mediana: $(0, 1]$, mediana:

$$Me = 0 + \frac{12.5 - 11}{8} \cdot 1 = 0.188$$

Moda

Ejercicio 1c

(c) Obtenga la media aritmética, la mediana y la moda para la distribución de frecuencias de los datos agrupados y sin agrupar.

Media aritmética

Sin agrupar: $\overline{x} = \frac{1}{N}\sum_{i=1}^{n} x_i n_i = \frac{-3}{25} = -0.12$

Agrupados: $\overline{x} = \frac{1}{N}\sum_{i=1}^{n} x_i n_i = \frac{-4.5}{25} = -0.18$

Mediana

Sin agrupar: $\frac{N}{2} = 12.5 \Longrightarrow$ position 13 $\Longrightarrow Me = 0.6$

Agrupar: intervalo de la mediana: $(0, 1]$, mediana:

$$Me = 0 + \frac{12.5 - 11}{8} \cdot 1 = 0.188$$

Moda

Sin agrupar: $Mo = 0.6$

Ejercicio 1c

(c) Obtenga la media aritmética, la mediana y la moda para la distribución de frecuencias de los datos agrupados y sin agrupar.

Media aritmética

Sin agrupar: $\overline{x} = \frac{1}{N}\sum_{i=1}^{n} x_i n_i = \frac{-3}{25} = -0.12$

Agrupados: $\overline{x} = \frac{1}{N}\sum_{i=1}^{n} x_i n_i = \frac{-4.5}{25} = -0.18$

Mediana

Sin agrupar: $\frac{N}{2} = 12.5 \Longrightarrow$ position 13 $\Longrightarrow Me = 0.6$

Agrupar: intervalo de la mediana: $(0, 1]$, mediana:

$$Me = 0 + \frac{12.5 - 11}{8} \cdot 1 = 0.188$$

Moda

Sin agrupar: $Mo = 0.6$

Agrupados: el intervalo modal es $(0, 1]$. Y la moda es:

$$Mo = 0 + \frac{4}{3+4} \cdot 1 = 0.571$$

Outline

Los cuantiles

Los cuantiles son medidas que dividen en partes iguales la distribución.
Los más utilizados son:

- Los cuartiles: Son tres valores que dividen la distribución en cuatro partes iguales, es decir, en cuatro intervalos dentro de cada cual están incluidos el 25% de los valores de la distribución.
- Los deciles: Son los nueve valores que dividen la distribución en diez partes que incluyen al 10% de los valores cada una.
- Los percentiles: Son los noventa y nueve puntos que dividen la distribución en cien partes iguales.

Cuantiles

El cálculo de los cuantiles es similar al calculo de la mediana.

De hecho la mediana es el cuartil 2, el décil 5 o el percentil 50.

Primero debemos calcular las frecuencias acumuladas y en ellas buscar el valor que ocupe la posición $\frac{r}{k}N$, teniendo en cuenta que k el número total de partes en que divido la distribución y r la parte a calcular.

Cuartiles	Deciles	Percentiles
K = 4	k = 10	K = 100
r = 1, 2 o 3	r = 1, 2, $\cdots$, 9	r = 1, 2, $\cdots$, 99

Una vez encontrada la posición el valor que contenga será el cuantil buscado.

Si los datos están agrupados en intervalos, sobre el intervalo encontrado aplicaremos la siguiente fórmula:

$$Q_{r/k} = L_{i-1} + \frac{r/k \cdot N - N_{i-1}}{n_i} \cdot c_i$$

Outline

Definición

- Los momentos de una distribución son valores que caracterizan a la distribución midiendo diferentes caracterí de la misma.
- Diremos que dos distribuciones son idénticas si todos sus momentos son iguales, y cuantos más momentos coincidan o se parezcan, más parecidas serán las distribuciones.
- El *r*-ésimo momento con respecto a algún origen arbitrario O_t se define como:

$$M_r = \frac{1}{N} \sum_{i=1}^{n} (x_i - O_t)^r \, n_i$$

Momentos con respecto a cero

Se denotan por α_r y se calculan tomando como origen $O_t = 0$:

$$\alpha_r = \frac{1}{N} \sum_{i=1}^{n} (x_i - 0)^r \, n_i = \frac{1}{N} \sum_{i=1}^{n} x_i^r \, n_i$$

$$\alpha_0 = \frac{N}{N} = 1$$

$$\alpha_1 = \frac{1}{N} \sum_{i=1}^{n} x_i \, n_i = \overline{x}$$

Momentos centrales o momentos con respecto a la media

Se denotan por m_r y se obtienen tomando como origen, $O_t = \overline{x}$:

$$\boxed{m_r = \frac{1}{N}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^r n_i}$$

$$\begin{aligned} m_0 &= \frac{N}{N} = 1 \\ m_1 &= \frac{1}{N}\sum_{i=1}^{n}\left(x_i - \overline{x}\right) n_i = \overline{x} - \overline{x} = 0 \\ m_2 &= \frac{1}{N}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 n_i = S^2 \end{aligned}$$

Todos los momentos centrales se pueden calcular en función de momentos con respecto a cero.

Outline

Motivación

- Dada una distribución de frecuencias, ¿hasta qué punto las medidas de tendencia central son representativas o respresentan adecuadamente la información de la muestra?
- Cuanto más cerca estén las observaciones de la medida de tendencia central, más representativa será.

  Será menos representativa si se observa mucha dispersión a su alrededor.

Dispersión o variabilidad

- Nos centraremos en observar la desviación de cada valor con respecto a una media de posición central.
- Si todos los valores están cerca de ese valor, diremos que la medida es representativa.
- Estas desviaciones se llaman **dispersion** or **variabilidad**.
- Si la dispersión o variabilidad es muy grande, la medida de tendencia central no será representativa.
- Ejemplo: considere dos muestras:

$$(3, 4, 5, 6, 7) \qquad (5, 5, 5, 5, 5)$$

- Ambas tienen la misma media (5), pero presentan diferente dispersión.

Recorrido y Recorrido intercuartílico

Recorrido

Definimos el recorrido de una muestra como la diferencia entre el valor máximo y el mínimo:

$$\text{Re}=x_n - x_1$$

Recorrido Intercuartílico

Definimos el recorrido intercuartílico como la distancia entre el primer y el tercer cuartil.

$$\text{Ri}=Q_{3/4} - Q_{1/4}$$

El recorrido intercuartílico nos indica la longitud del intervalo en el que están incluidos el 50% de los valores centrales de la muestra.
Así, si R_i es pequeño podemos intuir que la muestra presentará poca dispersión.
Estás dos medidas nos dan una idea de la dispersión de la muestra pero no utilizan ninguna medida de posición central, por lo que no pueden utilizarse para analizar la representatividad de ninguna medida en concreto.

Desviación respecto a la media aritmética

La desviación media respecto a la media aritmética se define como la media de las distancias en a la media aritmética en valor absoluto, es decir,

$$D_{\overline{x}} = \frac{\sum_{i=1}^{N} |x_i - \overline{x}| \cdot n_i}{N}$$

Un valor grande de esta medida de dispersión nos indicará una gran dispersión en la distribución y una media aritmética poco representativa.

Desviación respecto a la mediana

La desviación media respecto a la mediana se obtiene como la media entre el valor absoluto de las distancias a la mediana, es decir,

$$D_{Me} = \frac{\sum_{i=1}^{N} |x_i - Me| \cdot n_i}{N}$$

Un valor grande para esta medida, indicará al igual que en el caso anterior, gran dispersión en la muestra, y en este caso podremos afirmar que la mediana no es representativa.
Como comentamos en las propiedades de la mediana, la desviación media se hace mínima al calcularla con la mediana, por lo que:

$$D_{Me} < D_{\overline{x}}$$

Las medidas de desviación medias presentan el inconveniente de utilizar el valor absoluto, función que no es derivable y no resulta muy adecuada para determinados cálculos.

Varianza

- Entre todas las medidas de desviación con respecto a la media aritmética, la **varianza** y su raíz cuadrada, la **desviación típica**, son con mucho las más importantes y ampliamente utilizadas.
- Definición:

$$S^2 = m_2 = \frac{1}{N}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 n_i$$

- Cuanto mayor sea la dispersión de las observaciones en torno a su media, mayor será la varianza y menos representativa será la media.

Ejemplo 4: Varianza

Calcule la varianza de la siguiente distribución:

x_i	n_i	$(x_i - \overline{x})^2 n_i$
300	20	8320500
600	40	4761000
900	60	121500
1200	50	3251250
1500	30	9240750
	$N = 200$	25695000

Ejemplo 4: Varianza

Calcule la varianza de la siguiente distribución:

x_i	n_i	$(x_i - \overline{x})^2 n_i$
300	20	8320500
600	40	4761000
900	60	121500
1200	50	3251250
1500	30	9240750
	$N = 200$	25695000

$$S^2 = m_2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \overline{x})^2 n_i = \frac{25695000}{200} = 128475$$

Desviación típica

- La varianza se expresa en las unidades de la variable elevadas al cuadrado y esto dificulta su interpretación.
- La Desviación típica, se denota por S y es la raíz cuadrada positiva de la varianza:

$$S = \sqrt{S^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(x_i - \overline{x})^2 n_i}$$

- La desviación típica se mide en las mismas unidades que las observaciones, lo que la hace más sencilla de interpretar.
- Los valores extremos tienen una fuerte influencia tanto en la desviación típica como en la varianza, ya que su desviación respecto a la media se eleva al cuadrado.

Ejemplo 5: Desviación típica

Calcule la desviación típica para la siguiente distribución.

x_i	n_i	$(x_i - \overline{x})^2 n_i$
300	20	8320500
600	40	4761000
900	60	121500
1200	50	3251250
1500	30	9240750
	$N = 200$	25695000

Ejemplo 5: Desviación típica

Calcule la desviación típica para la siguiente distribución.

x_i	n_i	$(x_i - \overline{x})^2 n_i$
300	20	8320500
600	40	4761000
900	60	121500
1200	50	3251250
1500	30	9240750
	$N = 200$	25695000

$$S = \sqrt{S^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(x_i - \overline{x})^2 n_i} = \sqrt{128475} = 358.43$$

Propiedades de la varianza

1. La varianza nunca es negativa: $\boxed{S^2 \geq 0}$
2. La varianza es la desviación cuadrática óptima, ya que

$$\boxed{S^2 = \frac{1}{N}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 n_i < \frac{1}{N}\sum_{i=1}^{n}\left(x_i - k\right)^2 n_i \quad \forall k \neq \overline{x}}$$

3. La varianza se puede obtener en función de los momentos con respecto al origen, haciendo uso de la siguiente expresión: $\boxed{S^2 = m_2 = \alpha_2 - \alpha_1^2}$
4. La varianza permanece invariante ante cambios de origen: $\boxed{S^2\left(x + a\right) = S^2\left(x\right)}$
5. En cambios de escala, la varianza quedará multiplicada por el cuadrado de la constante que define el cambio de escala: $\boxed{S^2\left(kx\right) = k^2 S^2\left(x\right)}$

Coeficiente de variación

- Si queremos comparar distribuciones en términos de dispersión necesitamos una medida de variación relativa.
- Esta medida debe ser adimensional, es decir, no debe tener unidades de medida.
- El coeficiente de variación es la relación entre la desviación estándar y la media aritmética:

$$V = \frac{S}{\overline{x}}$$

- Al utilizar el coeficiente de variación nos libramos de las unidades de medida.
- V representa el número de veces S contiene $\overline{x}$. Cuanto mayor sea V más veces S contiene $\overline{x}$, por lo que un mayor V muestra menos representatividad de $\overline{x}$.
- Se utilizan todas las observaciones, es decir, toda la información disponible.
- Problema: No está definido si la media es cero.

Ejercicio 2

Para comparar los rendimientos entre empresas españolas y norteamericanas, se seleccionaron 20 empresas de características similares de ambas regiones, obteniéndose los siguientes resultados:

Empresas españolas		Empresas norteamericanas	
Beneficios (en euros)	n_i	Beneficios (en \$)	n_i
1000000	4	10000	2
1100000	6	11000	2
1200000	6	12000	4
1300000	2	13000	4
1400000	2	14000	4
		15000	2
		16000	2

Ejercicio 2 - Solución

Empresas españolas:

$$
\begin{aligned}
\overline{x} &= 1160000 \\
S^2 &= 14400000000 \\
S &= 120000 \\
V &= \frac{S}{\overline{x}} = \frac{120000}{1160000} = 0.103
\end{aligned}
$$

Ejercicio 2 - Solución

Empresas españolas:

$$\begin{aligned} \overline{x} &= 1160000 \\ S^2 &= 14400000000 \\ S &= 120000 \\ V &= \frac{S}{\overline{x}} = \frac{120000}{1160000} = 0.103 \end{aligned}$$

Empresas norteamericanas:

$$\begin{aligned} \overline{x} &= 13000 \\ S^2 &= 3000000 \\ S &= 1732.05 \\ V &= \frac{S}{\overline{x}} = \frac{1732.05}{13000} = 0.133 \end{aligned}$$

Outline

Motivación

- Es importante analizar la forma de la distribución, para entender mejor el comportamiento de la variable.
- La asimetría mide si la distribución es simétrica y si no lo es cuanto dista de serlo.
- La curtosis mide la concentración de valores alrededor de la media artmética.

Asimetría

- Es un indicador que permite evaluar el grado de simetría (o asimetría) de la distribución sin representarlas gráficamente.
- Si una distribución es simétrica, la distancia media de los valores a la media para los valores que inferiores a esta, es igual a las distacia media de los valores superiores a la media.
- Asimetría negativa: la cola izquierda es más larga y la masa de la distribución se concentra a la derecha de la figura.
- Asimetría positiva: la cola derecha es más larga y la masa de la distribución se concentra a la izquierda de la figura.

Momento de orden 3 con respecto a la media

- Para construir una medida de asimetría, necesitamos mantener los signos de las desviaciones. Por esto se emplea una potencia impar de las desviaciones en torno a la media ($r > 1$). La más sencilla es la forma cúbica:

$$m_3 = \frac{1}{N} \sum_{i=1}^{n} (x_i - \overline{x})^3 n_i$$

- Si la distribución es simétrica $\Longrightarrow m_3 = 0$.
- Si la distribución es asimétrica hacia la derecha $\Longrightarrow m_3 > 0$.
- Si la distribución es asimétrica hacia la izquierda $\Longrightarrow m_3 < 0$.
- El tercer momento con respecto a la media se expresa en unidades cúbicas, por lo que no permanecerá invariante ante cambios de escala.
- Otra forma de calcular m_3 es:

$$m_3 = \alpha_3 - 3\alpha_2\alpha_1 + 2\alpha_1^3$$

Coeficiente de asimetría de R.A.Fisher (g_1)

- El coeficiente de asimetría de R.A.Fisher es una medida de asimetría adimensional definida por:

$$g_1 = \frac{m_3}{S^3} = \frac{\frac{1}{N}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^3 n_i}{\left[\frac{1}{N}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 n_i\right]^{3/2}}$$

- S es siempre positive, por lo que el signo de g_1 será el mismo que el de m_3 de forma que:
  - Si la distribución es simétrica $\Longrightarrow g_1 = 0$.
  - Si la distribución es asimétrica a la derecha $\Longrightarrow g_1 > 0$.
  - Si la distribución es asimétrica a la izquierda $\Longrightarrow g_1 < 0$.
- Una distribución simétrica tiene $g_1 = 0$, pero eso no significa que una distribución con $g_1 = 0$ sea necesariamente simétrica.

Coeficiente de asimetría de Pearson (A_p)

Si la distribución es campaniforme y moderadamente asimetrícа se puede utilizar para analizar su simetría el coeficiente de asimetría de Pearson, definido por:

$$\boxed{A_p = \frac{\overline{X} - Mo}{S}}$$

Esta medida se puede aproximar mediante la siguiente expresión:

$$A_p \approx \frac{\overline{X} - Me}{S}$$

El coeficiente de asimetría de Pearson está basado en que si una distribución campaniforme es simétrica se cumple que, su media, su moda y su mediana coinciden, $\overline{X} = Me = Mo$.

Si la distribución es asimétrica positiva la media se sitúa por encima de la moda y Ap>0. Si la distribución es asimétrica negativa el efecto será el contrario.

Coeficiente de asimetría de Bowley (A_B)

El coeficiente de asimetría de Bowley está basado en la posición de los cuartiles y la mediana, y se calcula mediante la siguiente expresión:

$$A_B = \frac{Q_{1/4} + Q_{3/4} - 2Me}{Q_{3/4} - Q_{1/4}}$$

De nuevo, será cero en distribuciones simétricas, ya que en estas distribuciones el primer cuartil estará a la misma distancia de la media que el tercero.

Curtosis

- Las medidas de curtosis se aplican a distribuciones con forma de campana, es decir, a distribuciones simétricas o ligeramente asimétricas y unimodales.
- Las medidas de curtosis se centran analizar la concentración de valores en la "zona central" de la distribución.
- Una mayor curtosis significa que una mayor parte de la varianza es el resultado de desviaciones extremas, en lugar de desviaciones frecuentes de tamaño modesto.

Curtosis

- Para analizar la curtosis de una distribución se toma como referencia la distribución normal.
- La **normal distribution**, es la distribución más utilizada de entre las distribuciones campaniformes, y se define mediante la siguiente función de densidad:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

(donde μ y σ son la media y la desviación típica respectivamente)

- En esta distribución la mayoría de los valores se encuentran alrededor de la media y a medida que nos alejamos de esta encontramos cada vez menos valores. Esta característica es muy intuitiva y puede observarse en muchas magnitudes estudiadas.

Curtosis

- Tomando como referencia la distribución normal tenemos que:
  - Una distribución más apuntada de lo normal es **leptocúrtica**
  - Una distribución menos apuntada de lo normal es **platicúrtica**
  - La distribución normal es **mesocúrtica**
- Para la distribución normal se cumple que $m_4 = 3S^4$, donde m_4 es el momento de orden 4 con respecto a la media. Eso implica que:

$$\beta_2 = \frac{m_4}{S^4} = 3 \qquad g_2 = \frac{m_4}{S^4} - 3 = 0$$

Coeficiente de curtosis (g_2)

Definiremos el coeficiente de curtosis g_2 mediante:

$$\boxed{g_2 = \frac{m_4}{S^4} - 3 = 0}$$

- Mesocúrtica (normal), si $g_2 = 0$
- Leptocúrtica si $g_2 > 0$
- Platicúrtica si $g_2 < 0$

m_4 también se puede calcular mediante:

$$\boxed{m_4 = \alpha_4 - 4\alpha_3\alpha_1 + 6\alpha_2\alpha_1^2 - 3\alpha_1^4}$$

Ejercicio 3

Dada la siguiente distribución, calcule:

x_i	n_i	$x_i n_i$	$x_i^2 n_i$	$x_i^3 n_i$	$x_i^4 n_i$	N_i
0	2	0	0	0	0	2
10	4	40	400	4000	40000	2
20	7	140	2800	56000	1120000	13
30	5	150	4500	135000	4050000	18
40	2	80	3200	128000	5120000	20
	20	410	10900	323000	10330000	

(a) Coeficiente de asimetría.
(b) Coeficiente de curtosis.

Ejercicio 3 - Solución

$$\alpha_1 = \frac{1}{N}\sum_{i=1}^{n} x_i n_i = 20.5$$

Ejercicio 3 - Solución

$$\alpha_1 = \frac{1}{N}\sum_{i=1}^{n} x_i n_i = 20.5$$

$$\alpha_2 = \frac{1}{N}\sum_{i=1}^{n} x_i^2 n_i = 545$$

Ejercicio 3 - Solución

$$\alpha_1 = \frac{1}{N}\sum_{i=1}^{n} x_i n_i = 20.5$$

$$\alpha_2 = \frac{1}{N}\sum_{i=1}^{n} x_i^2 n_i = 545$$

$$\alpha_3 = \frac{1}{N}\sum_{i=1}^{n} x_i^3 n_i = 16150$$

Ejercicio 3 - Solución

$$\alpha_1 = \frac{1}{N}\sum_{i=1}^{n} x_i n_i = 20.5 \qquad \alpha_3 = \frac{1}{N}\sum_{i=1}^{n} x_i^3 n_i = 16150$$

$$\alpha_2 = \frac{1}{N}\sum_{i=1}^{n} x_i^2 n_i = 545 \qquad \alpha_4 = \frac{1}{N}\sum_{i=1}^{n} x_i^4 n_i = 516500$$

Ejercicio 3 - Solución

$$\alpha_1 = \frac{1}{N}\sum_{i=1}^{n} x_i n_i = 20.5 \qquad \alpha_3 = \frac{1}{N}\sum_{i=1}^{n} x_i^3 n_i = 16150$$

$$\alpha_2 = \frac{1}{N}\sum_{i=1}^{n} x_i^2 n_i = 545 \qquad \alpha_4 = \frac{1}{N}\sum_{i=1}^{n} x_i^4 n_i = 516500$$

$$m_2 = \alpha_2 - \alpha_1^2 = 545 - 20.5^2 = 124.75$$

Ejercicio 3 - Solución

$$\alpha_1 = \frac{1}{N}\sum_{i=1}^{n} x_i n_i = 20.5 \qquad \alpha_3 = \frac{1}{N}\sum_{i=1}^{n} x_i^3 n_i = 16150$$

$$\alpha_2 = \frac{1}{N}\sum_{i=1}^{n} x_i^2 n_i = 545 \qquad \alpha_4 = \frac{1}{N}\sum_{i=1}^{n} x_i^4 n_i = 516500$$

$$m_2 = \alpha_2 - \alpha_1^2 = 545 - 20.5^2 = 124.75$$

$$m_3 = \alpha_3 - 3\alpha_2\alpha_1 + 2\alpha_1^3 = 16150 - 3 \cdot 545 \cdot 20.5 + 2 \cdot 20.5^3 = -137.25$$

Ejercicio 3 - Solución

$$\alpha_1 = \frac{1}{N}\sum_{i=1}^{n} x_i n_i = 20.5 \qquad \alpha_3 = \frac{1}{N}\sum_{i=1}^{n} x_i^3 n_i = 16150$$

$$\alpha_2 = \frac{1}{N}\sum_{i=1}^{n} x_i^2 n_i = 545 \qquad \alpha_4 = \frac{1}{N}\sum_{i=1}^{n} x_i^4 n_i = 516500$$

$$m_2 = \alpha_2 - \alpha_1^2 = 545 - 20.5^2 = 124.75$$

$$m_3 = \alpha_3 - 3\alpha_2\alpha_1 + 2\alpha_1^3 = 16150 - 3 \cdot 545 \cdot 20.5 + 2 \cdot 20.5^3 = -137.25$$

$$m_4 = \alpha_4 - 4\alpha_3\alpha_1 + 6\alpha_2\alpha_1^2 - 3\alpha_1^4 = 36587.31$$

(a) Coeficiente de asimetría:

$$g_1 = \frac{m_3}{S^3} = \frac{-137.25}{124.75^{3/2}} = -0.0985$$

(b) Coeficiente de curtosis:

$$g_2 = \frac{m_4}{S^4} - 3 = \frac{36587.31}{124.75^2} - 3 = -0.649$$