Estadística descriptiva

TEMA 6

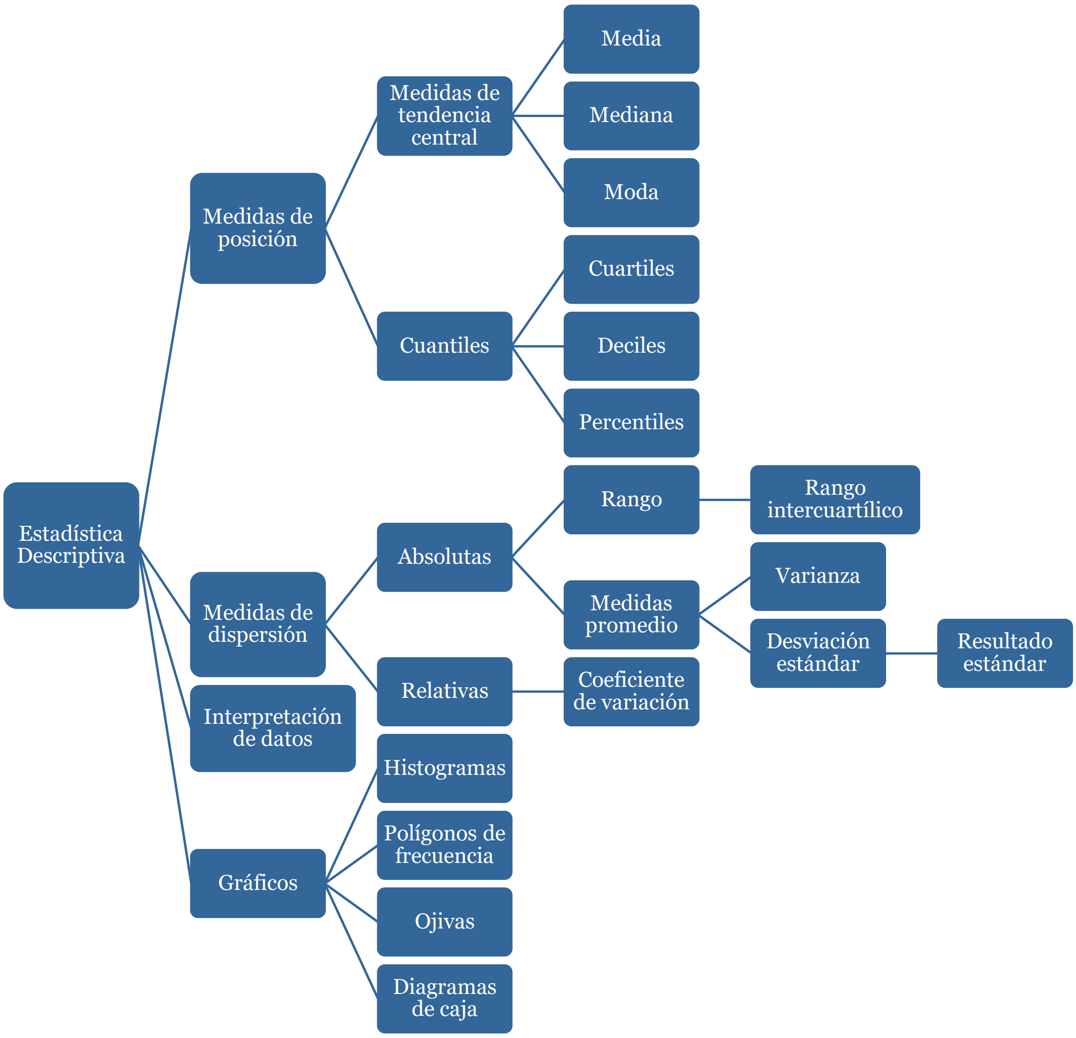
Ideas clave

6.1. ¿Cómo estudiar este tema?

En este primer tema se verán **las diferentes medidas de posición y dispersión que más se utilizan.** Además, se verá **la interpretación que le vamos a dar a los datos obtenidos.**

Por último, se hará un especial hincapié en **los gráficos que vamos a obtener, bien a mano o utilizando herramientas informáticas, y en la interpretación de los mismos.**

Para estudiar este tema **deberás comprender las Ideas clave** expuestas en este documento y que han sido elaboradas por el profesor de la asignatura. Estas ideas se van a complementar con lecturas y otros documentos para que puedas ampliar los conocimientos sobre el mismo. También es importante que **veas la lección magistral** grabada para afianzar los conocimientos expuestos en este tema.



6.2. Medidas de posición

La interpretación de los datos es crucial en estadística, por ello es preciso saber cómo se pueden organizar los datos para hacer comparaciones y establecer la fiabilidad de los mismos, entre otros fines.

Las medidas de posición son parámetros que se obtienen de los datos que permiten establecer un resumen de los mismos. Se diferencian en medidas de tendencia central y medidas de tendencia no central. Las medidas de tendencia central son las que proporcionan información sobre los valores medios a los que tienden un conjunto de datos. Tales medidas son la media, la mediana y la moda. Por otro lado, las medidas de tendencia no central son las que dividen el conjunto de datos en partes iguales. Se llaman cuantiles y se distinguen los cuartiles, deciles y percentiles.

1. Medidas de tendencia central

Las medidas de tendencia central se basan en el punto medio de una distribución y, tal y como se ha mencionado, son tres: la media, la mediana y la moda.

» Media

La media es un parámetro muy común en nuestras vidas cotidianas. También conocida como promedio, es la medida de tendencia central que puede representar a un conjunto de datos por sí misma. Hay diferentes tipos de medias las cuales se muestran a continuación.

○ Media aritmética

La media aritmética es la media que se emplea para calcular la estatura promedio, el número de asistentes promedio para ver una película, el precio promedio de algún producto, etc.

Cuando se trata de datos que no se repiten, la media aritmética se calcula como la suma de todos ellos dividida entre el total de datos. La fórmula para calcular la media aritmética, cuando los datos están presentados en la tabla de distribuciones de frecuencia, es la siguiente:

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

n es el número total de datos de la muestra y x_i y f_i son los valores y frecuencias absolutas de cada dato hasta el dato k, tomado como el último.

1.1.

A partir de las frecuencias relativas también se puede calcular la media aritmética con la siguiente fórmula:

$$\bar{x} = \sum_{i=1}^{k} x_i \cdot h_i$$

h_i es la frecuencia relativa de cada dato.

1.2.

Cuando en vez de la muestra se trata de la población entera, la media aritmética se calcula de la misma manera, pero se simboliza con la letra griega μ. Cuando los datos se están en intervalos, la media aritmética se calcula con los valores de las marcas de clase.

La media aritmética es la que más se emplea en estadística para saber cómo es una distribución de datos y para hacer comparaciones entre distribuciones. De hecho, cuando se habla de media, siempre hace referencia a la media aritmética. Sin embargo, cuando dentro del conjunto de datos hay algún dato anómalo que se aleja demasiado de los demás, la media no proporciona información sobre ese dato. Más aún, es preciso quitar tal dato antes de hacer la media, ya que si no se quita, no será una buena medida de tendencia central para el resto de datos. Cabe destacar que para aquellas agrupaciones en las que haya una clase del tipo 'mayor o menor que', la media aritmética no se puede calcular ya que no se conoce el dato.

Ejemplo 1. Cálculo de la media aritmética con datos agrupados en intervalos

Calcular la media aritmética a partir de los siguientes datos, agrupados en intervalos, de faltas de asistencia trimestrales de alumnos de cuatro cursos del mismo año académico. ¿Qué promedio de faltas de asistencia ha habido?

I_i	f_i
4-8	25
9-15	18
16-20	12
21-27	15
28-32	5

Tabla 1. Distribución de frecuencias absolutas de faltas de asistencia.

Primero hay que calcular las marcas de clase para cada intervalo (nótese que la amplitud de los intervalos no es la misma):

I_i	Marca de clase x_i	f_i
4-8	6	25
9-15	12	18
16-20	18	12
21-27	24	15
28-32	30	5
Total		75

Tabla 2. Marcas de clase.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i \cdot f_i = \frac{6 \cdot 25 + 12 \cdot 18 + 18 \cdot 12 + 24 \cdot 15 + 30 \cdot 5}{75} = 14{,}56$$

- **Media ponderada**

La media ponderada es el tipo de media que se calcula cuando se pretende saber información sobre un valor en concreto con respecto al total de valores. Por ejemplo, si se precisa saber la calificación total de un alumno para una evaluación suponiendo cada examen, trabajo, etc. cada uno de estos resultados tendrá un peso para la calificación total y se trabaja con la media ponderada.

Para calcularla hay que multiplicar cada dato por una ponderación, que es el valor del peso de cada dato respecto al total, hacer el sumatorio y dividir entre el total de las ponderaciones tal y como se muestra a continuación:

$$\bar{x}_p = \frac{\sum_{i=1}^{k} x_i \cdot w_i}{\sum_{i=1}^{k} w_i}$$

w es el peso para unidad de análisis.

1.3.

Prestando un poco de atención, esta fórmula es la misma que la de la media aritmética cuando se trabaja con frecuencias relativas. Los pesos son las frecuencias relativas y la suma de todas es 1 o el 100%.

Ejemplo 2. Calcular la media ponderada

A partir de los datos del ejemplo anterior, calcula la media ponderada:

Calculamos las frecuencias relativas, cuya suma ha de ser la unidad.

I_i	Marca de clase x_i	f_i	h_i
4-8	6	25	0,333
9-15	12	18	0,240
16-20	18	12	0,160
21-27	24	15	0,200
28-32	30	5	0,066
Total		75	1

Tabla 3. Cálculo de la media ponderada.

$$\bar{x}_p = \frac{\sum_{i=1}^{k} x_i \cdot w_i}{\sum_{i=1}^{k} w_i} = \frac{6 \cdot 0,333 + 12 \cdot 0,240 + 18 \cdot 0,160 + 24 \cdot 0,200 + 30 \cdot 0,066}{0,333 + 0,240 + 0,160 + 0,200 + 0,066} = 14,56$$

- **Media geométrica**

La media geométrica de un conjunto de datos es la raíz enésima del producto de cada dato elevado a la frecuencia absoluta.

$$\bar{x}_g = \sqrt[n]{\prod_{i=1}^{k} {x_i}^{f_i}}$$

1.4.

Para los cálculos se toman logaritmos

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

1.5.

Con lo que

$$\bar{x}_g = 10^{\left(\frac{1}{n}\sum_{i=1}^{k} f_i \cdot log x_i\right)}$$

1.6.

La media geométrica se usa principalmente para hacer promedio de índices, calcular las tasas de variaciones promedio de ventas, calcular el factor de crecimiento de un capital con el tiempo, etc.

En tal caso se aplica la fórmula de la media geométrica y se calcula tal y como se muestra a continuación donde los valores de x_i hacen referencia a los factores de crecimiento.

$$\bar{x}_g = \sqrt[n]{x_1 \cdot x_1 \cdot \dots \dots x_2}$$

1.7.

Ejemplo 3. Calcular la media geométrica

Si se invierten 100.000 euros al 2%, 4% y 7% durante tres años, ¿cuál es el valor medio al que se está invirtiendo tal capital?

Se calcula el factor de crecimiento en cada año con la siguiente fórmula:

$$1 + \%\ ingresos$$

Año	% de ingresos	Factor de crecimiento
1	2%	1,02
2	4%	1,04
3	7%	1,07

Tabla 4. Crecimiento de un capital en 3 años.

Aplicando la fórmula 1.7:

$$\bar{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \dots x_k} = \sqrt[3]{1{,}02 \cdot 1{,}04 \cdot 1{,}07} = 1{,}043$$

Por lo tanto el rédito medio es **4,3%.**

Si los cálculos se hubiesen hecho con la media aritmética como (1,02+1,04+1,07)/3 el resultado hubiese sido también 1,043. Los valores obtenidos son iguales por lo que se puede pensar que no tiene sentido usar la media geométrica. El problema es cuando se trata de porcentajes elevados tal y como se muestra en el ejemplo siguiente:

Ejemplo 4. Cálculo de la media geométrica con valores de porcentaje elevados

A continuación se muestran los datos que pagaron los bancos en cinco años en los que el índice de inflación fue elevado. Calcula el factor de crecimiento promedio.

Año	% (tasa de interés)	Factor de crecimiento
1	75%	1,75
2	100%	2
3	200%	3
4	250%	3,5
5	350%	4,5

Tabla 5. Tasas de interés en 5 años.

Se calcula el factor de crecimiento como 1 + tasa de interés y a continuación se calcula la media geométrica.

$$\bar{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \ldots x_k} = \sqrt[5]{1{,}75 \cdot 2 \cdot 3 \cdot 3{,}5 \cdot 4{,}5} = 2{,}77$$

El resultado obtenido empleando la media aritmética es 2,95. Este resultado difiere más que la media geométrica, por lo que sería menos apropiado su uso.

- **Media armónica**

En aquellas situaciones en las que se promedian magnitudes relativas, respecto al tiempo, al área, etc., se emplea la media armónica.

Se calcula como el inverso de la media aritmética de los valores inversos de los datos multiplicados por las frecuencias. Se considera más precisa que la media aritmética, aunque cuando se trata de valores muy próximos a cero es preferible no emplearla. La fórmula para calcularla es:

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

1.8.

Ejemplo 5. Cálculo de la media armónica

Un automóvil realiza 20 km de un trayecto a 90 km/h, 100 km a 120 km/h y 10 km a 70 km/h, ¿Cuál es la velocidad promedio de todo el trayecto?

Como se trata de variaciones con respecto al tiempo se emplea la media armónica:

$$\bar{x}_a = \frac{n}{\sum_{i=1}^{k} \frac{f_i}{x_i}} = \frac{3}{\frac{1}{90} + \frac{1}{120} + \frac{1}{70}} = 88{,}94\ km/h$$

- **Media cuadrática**

La media cuadrática o valor cuadrático medio es muy empleada cuando se precisa calcular el valor medio de valores negativos y positivos puesto que se basa en elevar al cuadrado tales valores y hacerlos positivos. Se suele emplear para calcular el valor medio de errores en medición tal y como se estudiará más adelante. La fórmula empleada es:

$$\bar{x}_c = \sqrt{\frac{1}{n} \sum_{i=1}^{k} x_i^2 \cdot f_i}$$

1.9.

Para un mismo conjunto de datos positivos se cumple que: $\bar{x}_a \leq \bar{x}_g \leq \bar{x} \leq \bar{x}_c$

Ejemplo 6. Cálculo de la media cuadrática

A continuación se muestran los errores que se han obtenido en la medición de una cantidad de sustancia con un aparato, calcular el error cuadrático medio.

Medida	Error (E)
3,15	-0,15
5,20	0,65
3,65	-0,07
4,90	0,35
4,14	0,12

Tabla 6. Errores en la medición.

$$\bar{x}_c = \sqrt{\frac{1}{n}\sum_{i=1}^{k} x_i^2 \cdot f_i} = \sqrt{\frac{1}{5}\cdot[(-0{,}15)^2 + (0{,}65)^2 + (-0{,}07)^2 + (0{,}35)^2 + (0{,}12)^2]} = 0{,}34$$

» Mediana

La mediana es el valor central de un conjunto de datos, los cuales han de estar ordenados, y su posición se calcula a partir de la siguiente fórmula:

$$Posición\ de\ la\ mediana = \frac{n+1}{2}$$

1.10.

De esta manera, cuando se trata de un número impar de datos, la mediana será el valor central que deje el mismo número de datos a la izquierda y a la derecha. Cuando el número de datos es par, la mediana será la media aritmética de los valores centrales.

Ejemplo 7. Cálculo de la mediana de un número impar e impar de datos

Calcular la mediana de los siguientes datos:

A. 1,2 2,3 3,6 4,7 5,3 6,9 8,1
B. 100 267 340 456 549 678 750 811

A. En este caso se observa fácilmente que la mediana es el valor central 4,7 ya que deja 3 valores a cada lado. Se puede comprobar mediante la fórmula:

$$Posición\ de\ la\ mediana = \frac{n+1}{2} = \frac{7+1}{2} = 4$$

El dato en posición 4 es 4,7.

B. En este caso la mediana se halla entre los datos en posiciones 4 y 5 que son los que dejan 3 datos a cada lado. Empelando la fórmula resulta:

$$Posición\ de\ la\ mediana = \frac{n+1}{2} = \frac{8+1}{2} = 4{,}5$$

Como no es número entero hay que hacer el promedio entre los datos de las posiciones 4 y 5 que son los enteros más próximos.

$$\bar{x} = \frac{456+549}{2} = 502{,}5$$

Ahora bien, cuando se trata de datos agrupados en intervalos lo que hay que saber es cuál de dichos intervalos contiene a la mediana. El procedimiento es similar a cuando se trata de datos no agrupados para sacar la posición pero para calcular la mediana se emplea la siguiente fórmula:

$$Mediana = \left(\frac{\left(\frac{n+1}{2}\right) - (F_{m-1}+1)}{f_m}\right) \cdot a + L_m$$

n es el total de datos, F_{m-1} es la frecuencia acumulada hasta el intervalo donde se halla la mediana sin incluir, f_m la frecuencia del intervalo de la mediana, a la amplitud del intervalo y L_m el límite inferior del intervalo de la mediana.

1.11.

En contraposición a la media, la mediana es un valor que no varía aunque se adicionen datos a los extremos y aunque tales extremos sean del tipo 'menor o mayor que'. Además, se puede calcular aunque se trate de variables cualitativas. Por el contario, su cálculo puede resultar más complejo que el de la media y precisa una ordenación previa.

Ejemplo 8. Cálculo de la mediana

A partir de los datos de la tabla 1 del ejemplo 1 calcular la mediana.

I_i	f_i
4-8	25
9-15	18
16-20	12
21-27	15
28-32	5
Total	75

Datos de la tabla 1.

Se calcula la posición de la mediana:

$$Posición\ de\ la\ mediana = \frac{n+1}{2} = \frac{75+1}{2} = 38$$

Se hace una tabla con las frecuencias acumuladas:

I_i	f_i	F_i
4-8	25	25
9-15	18	43
16-20	12	55
21-27	15	70
28-32	5	75
Total	75	

Tabla 7. Distribuciones de frecuencia con las frecuencias acumuladas.

Comparando con la columna de frecuencias acumuladas, el valor 38 se halla en el intervalo [9-15). Aplicando la fórmula 1.11. se obtiene la mediana:

$$Mediana = \left(\frac{\left(\frac{n+1}{2}\right) - (F_{m-1}+1)}{f_m}\right) \cdot a + L_m = \left(\frac{\left(\frac{75+1}{2}\right) - (25+1)}{18}\right) \cdot 6 + 9 = 13$$

» Moda

La moda es el valor que más se repite de una serie de datos. Cuando se trata de datos agrupados en intervalos se suele hablar de intervalos modales. La fórmula para calcular la moda es:

$$Moda = L_{mo} + \left[\frac{f_{mo} - f_{mo-1}}{(f_{mo} - f_{mo-1}) + (f_{mo} - f_{mo+1})}\right] \cdot a$$

L_{mo} es el límite inferior del intervalo modal, f_{mo} es la frecuencia del intervalo modal, f_{mo-1} es la frecuencia del intervalo inferior al intervalo modal, f_{mo+1} es la frecuencia del intervalo por encima del intervalo modal y a es la amplitud del intervalo.

1.12.

La moda presenta las mismas ventajas que las comentadas para la mediana anteriormente, es decir, no se ve afectada si se introducen datos en los extremos, se puede calcular si extremos son del tipo 'mayor o menor que' y si se trata de variables cualitativas. Pero su uso es menor que el de la mediana en aquellas situaciones en las que no se repitan datos y cuando se trata de distribuciones multimodales (se llama 'distribución multimodal' a aquella que tiene más de una moda o intervalo modal).

Ejemplo 9. Cálculo de la moda en datos agrupados

Calcular la moda para los datos de la tabla 1.

I_i	f_i
4-8	25
9-15	18
16-20	12
21-27	15
28-32	5
Total	75

Datos de la tabla 1.

El primer intervalo es el que más datos tiene.

$$Moda = L_{mo} + \left[\frac{f_{mo} - f_{mo-1}}{(f_{mo} - f_{mo-1}) + (f_{mo} - f_{mo+1})}\right] \cdot a = 4 + \left[\frac{25 - 0}{(25 - 0) + (25 - 18)}\right] \cdot 4 = 7{,}125$$

2. Medidas de tendencia no central

Las medidas de tendencia no central, también llamadas cuantiles, se emplean para calcular otros puntos del conjunto de datos. Tales valores hacen que el conjunto de datos quede dividido en secciones constantes. Dentro de los cuantiles se hallan los cuartiles, deciles y percentiles.

» Cuartiles

Los cuartiles son las medidas de tendencia no central que dividen al conjunto de datos en 4 partes iguales que albergan el 25% de los datos cada una. El primer cuartil, Q_1, corresponde al 25% de los datos; el segundo cuartil, Q_2, al 50% de los datos, por lo que es la mediana; y el tercer cuartil, Q_3, al 75% de los datos.

Ejemplo 10. Cálculo de los cuartiles de un conjunto de datos

Calcular los cuartiles de este conjunto de datos de faltas de asistencia a un curso:

xi	f_i	F_i
0	3	3
2	12	15
4	21	36
6	34	70
8	22	92
10	3	95
Total	95	

Tabla 8. Distribución de frecuencias de datos no agrupados.

El segundo cuartil, Q_2, es la mediana:

$$Posición\ de\ la\ mediana = \frac{n+1}{2} = \frac{95+1}{2} = 48$$

Comparando con la frecuencia absoluta acumulada, el segundo cuartil, Q_2, corresponde al dato que está en posición 4. El 50% de los asistentes faltaron 6 veces.

La posición del primer cuartil, Q_1, se puede calcular como la mitad de la posición de la mediana, es decir, la que corresponde a un cuarto de los datos:

$$Posición\ de\ Q_1 = \frac{n+1}{4} = \frac{95+1}{4} = 24$$

Comparando con la frecuencia absoluta acumulada, el primer cuartil, Q_1, corresponde al dato que está en posición 3. El 25% de los asistentes faltaron 4 veces.

La posición del tercer cuartil, Q_3, se puede calcular como la que corresponde a las tres cuartas partes de los datos:

$$Posición\ de\ Q_3 = \frac{3(n+1)}{4} = \frac{3(95+1)}{4} = 72$$

Comparando con la frecuencia absoluta acumulada, el tercer cuartil, $Q3$, corresponde al dato en posición 5. El 75% de los asistentes faltaron 8 veces.

Cuando se trata de datos agrupados en intervalos se emplea la fórmula 1.11.

Ejemplo 11. Cálculo de cuartiles de una distribución de datos agrupados en intervalos

A partir de los datos de la tabla 8 calcular los cuartiles.

I_i	f_i	F_i
4-8	25	25
9-15	18	43
16-20	12	55
21-27	15	70
28-32	5	75
Total	75	

Datos de la tabla 7.

El segundo cuartil Q_2 es la mediana que ya se ha calculado en el ejemplo 8 y es 13. El 50% de los datos es menor o igual a 13.

El primer cuartil Q1 corresponde a la cuarta parte de los datos:

$$Posición\ de\ Q_1 = \frac{n+1}{4} = \frac{75+1}{4} = 19$$

Comparando con las frecuencias acumuladas, tal posición está en el primer intervalo.

$$Q_1 = \left(\frac{\left(\frac{n+1}{4}\right) - (F_{m-1}+1)}{f_m}\right) \cdot a + L_m = \left(\frac{\left(\frac{75+1}{4}\right) - (0+1)}{25}\right) \cdot 4 + 4 = 6{,}88$$

El 25% de los datos es menor o igual que 6,88.

El tercer cuartil Q3 corresponde a las tres cuartas partes de los datos:

$$Posición\ de\ Q_3 = \frac{3(n+1)}{4} = \frac{3(75+1)}{4} = 57$$

Comparando con las frecuencias acumuladas, tal posición está en el intervalo [21-27)

$$Q_3 = \left(\frac{\frac{3(n+1)}{4} - (F_{m-1}+1)}{f_m}\right) \cdot a + L_m = \left(\frac{\frac{3(75+1)}{4} - (55+1)}{15}\right) \cdot 6 + 21 = 21{,}4$$

El 75% de los datos es menor o igual que 21,4.

» **Deciles**

Los deciles dividen la distribución en diez partes iguales en las que cada una concentra el 10% de los resultados.

Ejemplo 12. Calcular los deciles de datos no agrupados

A partir de los datos de la Tabla 10 calcular los deciles 2, 4 y 9.

xi	f_i	F_i
0	3	3
2	12	15
4	21	36
6	34	70
8	22	92
10	3	95
Total	95	

Datos de la tabla 8.

Posición	Valor
$\frac{2(95+1)}{10} = 19{,}2$	$D_2 = 4$
$\frac{4(95+1)}{10} = 38{,}4$	$D_4 = 6$
$\frac{9(95+1)}{10} = 86{,}4$	$D_9 = 10$

Tabla 9. Cálculo de los deciles 1, 4 y 9.

El 20% de los asistentes faltaron 4 veces o menos, El 40% de los asistentes faltaron 6 veces o menos y el 90% de los asistentes faltaron 10 veces o menos.

Ejemplo 13. Cálculo de deciles en datos agrupados

A partir de los datos de la tabla 8 calcular los deciles 1,5, y 7:

I_i	f_i	F_i
4-8	25	25
9-15	18	43
16-20	12	55
21-27	15	70
28-32	5	75
Total	75	

Datos de la tabla 7.

Posición del decil	Intervalo	Valor del decil
$\frac{(75+1)}{10} = 7{,}6$	[4-8)	$D_1 = \left(\frac{\frac{(75+1)}{10} - (0+1)}{25}\right) \cdot 4 + 4 = 5{,}06$
$\frac{5(75+1)}{10} = 38$	[9-15)	$D_5 = \left(\frac{\frac{5(75+1)}{10} - (25+1)}{18}\right) \cdot 6 + 9 = 13$
$\frac{7(75+1)}{10}$ $= 53{,}2$	[16-20)	$D_7 = \left(\frac{\frac{7(n+1)}{10} - (F+1)}{f_m}\right) \cdot w + L_m$ $= \left(\frac{\frac{7(75+1)}{10} - (43+1)}{12}\right) \cdot 4 + 16 = 19{,}06$

Tabla 10. Deciles 1, 5 y 7 para datos agrupados.

El 10% de los datos incluyen hasta el dato 5.06, el 50% de los datos incluyen hasta el dato 13, la mediana, y el 70% de los datos incluyen hasta el dato 19,06.

» Percentiles

Los percentiles dividen el conjunto de datos en 100 partes de manera que cada dato concentra el 1% de los valores.

Ejemplo 14. Cálculo de percentiles en datos no agrupados

Calcular el percentil 50 con los datos de la tabla 8.

xi	f_i	F_i
0	3	3
2	12	15
4	21	36
6	34	70
8	22	92
10	3	95
Total	95	

Datos de la tabla 8.

La posición del percentil se calcula de la siguiente manera:

$$Posición\ del\ percentil = \frac{50(n+1)}{100} = \frac{50(95+1)}{100} = 48$$

Este valor pertenece al cuarto dato. El 50% de los asistentes faltaron 6 veces o menos.

Si se calculase el percentil 70 con los datos agrupados de la tabla 7, el valor sería el mismo que el decil 7 calculado en el ejemplo 13.

6.3. Medidas de dispersión

A la hora de analizar un conjunto de datos, es preciso saber cuánto difieren los datos entre sí. Tal magnitud se conoce como dispersión y es de gran importancia para la toma de decisiones en estadística. Imaginemos que dos conjuntos de datos tienen la misma media, la forma de saber cómo están distribuidos tales datos es mediante el estudio de la dispersión de los mismos.

En la figura que se muestra del libro *Estadística para Administración y Economía,* se recogen tres distribuciones que tienen la misma media pero su forma es diferente, por lo que los datos presentan diferente dispersión.

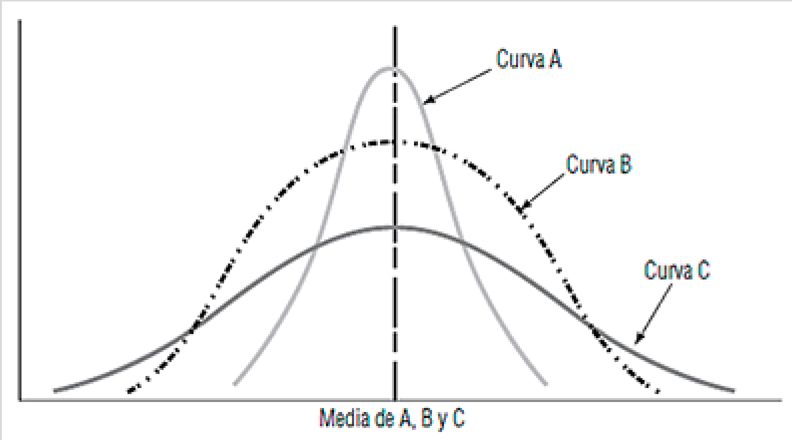


Figura 1. Distribuciones de valores con la misma media pero diferente dispersión.

1. Medidas de dispersión absolutas

Las medidas de dispersión absolutas nos dan información sobre la dispersión de una distribución sin permitir la comparación entre distribuciones. Se trata del rango y las medidas de desviación promedio.

» Rango

El rango se calcula como la diferencia entre los valores más grande y más pequeño de una serie de datos.

$$Rango = valor\ más\ grande - valor\ más\ pequeño$$

1.13.

Su valor se calcula fácilmente pero sus limitaciones son muchas ya que no da información de los datos interiores, de cómo se dispersan. Por tal motivo, si un valor central se aleja de los demás el rango no dará información sobre esa dispersión. Además, cuando se trata de extremos del tipo 'mayor o menor que' el rango no se puede calcular. Hay dos tipos de rangos que eliminan tales inconvenientes, el rango interfractil y el rango intercuartil o intercuartílico.

El rango interfractil se calcula entre dos medidas de cuantiles de una distribución. Por ejemplo, entre los deciles 3 y 7, los percentiles 40 y 65, etc. y da información de la dispersión entre tales cuantiles.

El rango intercuartil es un tipo de rango interfractil definido como la diferencia entre el primer cuartil y el tercer cuartil. La mitad del rango intercuartil se llama desviación intercuartil.

$$Rango\ intercuartil = Q_3 - Q_1$$

1.14.

$$Desviación\ intercuartil = \frac{Q_3 - Q_1}{2}$$

1.15.

El rango intercuartil tiene importancia para medir cuánto hay que desplazarse de la mediana hacia valores mayores y menores antes de recorrer la mitad de los mismos, tal y como se muestra en la siguiente figura del manual *Estadística para Administración y Economía*.

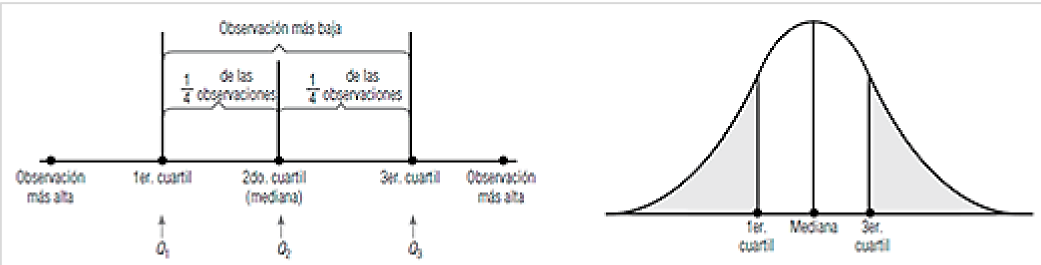


Figura 2. Explicación del rango intercuartil.

El rango intercuartil se emplea para hacer diagramas de caja que se estudian más adelante en este mismo tema.

Medidas de desviación promedio

Las medidas de desviación promedio permiten calcular la dispersión de los datos de manera más precisa, aunque más compleja. Se trata de la varianza y la desviación estándar y estudian la dispersión de los datos respecto a la media.

- **Varianza**

La varianza permite identificar la media de las desviaciones de entre los datos de una variable. Es el sumatorio de los cuadrados de las diferencias entre los valores y la media dividido entre el total de datos. Cuando la muestra es la población entera, la varianza se simboliza como σ^2 (sigma cuadrada) y se calcula de la siguiente manera:

$$\sigma^2 = \frac{\sum(x-\mu)^2}{N} = \frac{\sum x^2}{N} - \mu^2$$

x es cada dato, μ es la media poblacional y N es el total de observaciones poblacionales.

1.16.

El hecho de elevar al cuadrado hace que diferencias sean positivas y se consigue dar más valor a aquellas que supongan más separación respecto a la media. La segunda parte de la igualdad permite calcular la varianza sin tener que hacer las diferencias individuales entre cada valor y la media.

Cuando los datos se repiten, se puede calcular la varianza a partir de las frecuencias:

$$\sigma^2 = \frac{\sum f(x-\mu)^2}{N} = \frac{\sum f x^2}{N} - \mu^2$$

f es la frecuencia de cada clase.

1.17.

Cuando se trata de muestras, mayoría de los casos de estudio, la varianza se expresa de la siguiente forma:

$$S^2 = \frac{\sum(x-\bar{x})^2}{n-1} = \frac{\sum x^2}{n-1} - \frac{n\bar{x}}{n-1}$$

S^2 es la varianza de la muestra, $\bar{x}$ es la media muestral y n el tamaño muestral. Al tomar varias muestras y comparar la suma de las varianzas con la varianza de la población total los datos se acercan más cuando se toma para las muestran un valor de datos de *n-1*.

1.18.

- **Desviación estándar**

La desviación estándar es la raíz cuadrada de la varianza y se emplea mucho más para hacer los cálculos, ya que permite obtener la dispersión en las mismas unidades que los datos, mientras que la varianza proporciona tal valor elevado al cuadrado. No tiene sentido si, por ejemplo, se está evaluando la dispersión entre diferentes medidas de longitud de unos tornillos y medimos tal dispersión en cm^2. Lo mejor es mostrar la dispersión en las unidades de medida.

La desviación estándar es la raíz cuadrada del sumatorio de los cuadrados de las distancias entre cada una de las observaciones y la media dividido entre el total de observaciones. Cuando se trata de la población la desviación estándar se representa como *σ:*

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum(x-\mu)^2}{N}} = \sqrt{\frac{\sum x^2}{N} - \mu^2}$$

1.19.

El resultado estándar permite saber lo que cada observación en particular se aleja de la media, por encima o por debajo:

$$Resultado\ estándar = \frac{x-\mu}{\sigma}$$

1.20.

Cuando hay información sobre las frecuencias absolutas, la desviación estándar se calcula como:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum f(x-\mu)^2}{N}} = \sqrt{\frac{\sum fx^2}{N} - \mu^2}$$

f es la frecuencia de cada clase.

1.21.

Para una muestra la fórmula es:

$$S = \sqrt{S^2} = \sqrt{\frac{\sum f(x-\bar{x})^2}{n-1}}$$

1.22.

Ejemplo 15. Calculo de la desviación estándar para una muestra

A partir de los datos de la siguiente tabla de salarios anuales de personas de una empresa, calcular la desviación estándar. Un individuo que cobre 22.000 euros anuales, ¿cuánto se aleja de la media?

Salarios	Empleados
8.000-10.000	25
10.000-15.000	32
15.000-20.000	12
20.000-30.000	16
30.000-50.000	7
50.000-100.000	2

Tabla 11. Salarios anuales de personas en una empresa.

Calculamos la media con las marcas de clase de cada intervalo:

$$\bar{x} = \frac{9.000 \cdot 25 + 12.500 \cdot 32 + 17.500 \cdot 12 + 25.000 \cdot 16 + 40.000 \cdot 7 + 75.000 \cdot 2}{94} = 17.712{,}8$$

Intervalo	x	f_i	$x - \bar{x}$	$(x - \bar{x})^2$	$f_i \cdot (x - \bar{x})^2$
8.000-10.000	9.000	25	-8.713	75912361,1	1897809027
10.000-15.000	12.500	32	-5.213	27172971,1	869535074
15.000-20.000	17.500	12	-213	45271,07	543252,875
20.000-30.000	25.000	16	7.287	53103721,1	849659537
30.000-50.000	40.000	7	22.287	496720621	3477044348
50.000-100.000	75.000	2	57.287	3281826721	6563653442
Total		94			13658244681

$$S = \sqrt{S^2} = \sqrt{\frac{\sum f(x - \bar{x})^2}{n - 1}} = \sqrt{\frac{13658244681}{94 - 1}} = 12.118,7$$

Para calcular lo que se aleja de la media alguien que cobre 22.000 euros anuales, se calcula el resultado estándar:

$$Resultado\ estándar = \frac{x - \bar{x}}{S} = \frac{22.000 - 17.712,8}{12.118,7} = 1,8$$

Ese individuo está alejado en 1,8 desviaciones estándar de la media.

2. Medidas de dispersión relativas

En aquellas situaciones en las que se toman muestran iguales pero en diferentes intervalos de tiempo, o llevadas a cabo por diferentes personas o instrumentos para obtener algún tipo de información sobre un parámetro, las desviaciones estándar de cada muestra no nos permiten hacer una comparación entre todos los ensayos ya que cada ensayo tiene una media. Por ejemplo, si con un ensayo se obtiene una media de 2,78 y una desviación estándar de 0,14 y con otro ensayo una media de 3,15 y una desviación estándar de 0,78, al comparar las dos desviaciones estándar no se llega a ninguna conclusión. Es preciso saber cuánto vale cada desviación con respecto a su correspondiente media, es decir, lo que lo que 0,14 supone a 2,78 y lo que 0,78 supone a 3,15. A partir de ahí, es cuando se podrán comparar los ensayos.

El coeficiente de variación permite relacionar la desviación estándar con la media de un conjunto de datos en tanto por ciento.

$$C.V = \frac{\sigma}{\mu} \cdot 100$$

1.23.

Cuando se trate de una muestra el coeficiente de variación es:

$$C.V = \frac{S}{\bar{x}} \cdot 100$$

1.24.

Cálculo de la media aritmética y la desviación estándar con la calculadora

1. Borrar datos memorizados:

Pulsar *SHIFT* + *CLR* + tecla 1 (*SCL*) = 0 tecla 3 (*All*).

2. Introducir los datos:

Escribir el dato x_1 y pulsar *DT* (tecla M+). Hacer lo mismo para el resto de datos.

3. Escribir las frecuencias absolutas de cada dato:

Pulsar en la flecha hacia abajo en *REPLAY*. Aparecen los valores de x_1 y las frecuencias con valor 1. Escribir la frecuencia y la tecla =.

4. Calcular la media aritmética:

Pulsar *SHIFT* + *S-VAR* (tecla del número 2) + $\bar{x}$ (tecla del número 1)=.

5. Calcular la desviación estándar:

Pulsar *SHIFT* +*S-VAR* (tecla del número 2) + X_{n-1} (tecla del número 3)=.

6.4. Interpretación de los datos obtenidos

En este apartado se expondrán algunos ejemplos resueltos de aplicaciones de lo visto en este tema para aprender a interpretar los datos y saber qué parámetros debemos calcular en determinadas situaciones.

Ejemplo 16

Una empresa tiene un acuerdo de crédito revolvente con un banco. Los saldos mostrados del año anterior se muestran a continuación. La compañía puede tener una tasa de interés menor si el saldo mensual promedio es 70.000. ¿Califica la empresa para dicha tasa de interés menor?

Enero	Febrero	Marzo	Abril
110.000	75.000	50.000	72.000
Mayo	Junio	Julio	Agosto
86.000	120.000	90.000	85.000
Septiembre	Octubre	Noviembre	Diciembre
65.000	55.000	47.000	71.000

La media aritmética es 71.167 por lo que la empresa no califica para esa tasa de interés.

Ejemplo 17

En unos test realizados para medir la longitud de unos tubos de plástico, el analista 1 obtuvo una serie de medidas en las que la media de los valores fue 5,70 y la desviación estándar fue 0,05. Por otro lado, el Analista 2 obtuvo unos valores de resistencia de 6,9 y la desviación fue de 0,03. ¿Cuál de los dos obtuvo mayor variabilidad en los resultados?

Analista 1:

$$C.V = \frac{S}{\bar{x}} \cdot 100 = \frac{0,05}{5,70} \cdot 100 = \ 0,88\%$$

Analista 2:

$$C.V = \frac{S}{\bar{x}} = \frac{0,03}{6,9} \cdot 100 = \ 0,43\%$$

El Analista 1 obtuvo mayor variabilidad en los resultados.

Ejemplo 18

A continuación se muestra la producción de los trabajadores en cinco líneas de una empresa de embotado de encurtidos por hora, ¿Cuál es la productividad media? ¿Cuál es el tiempo medio para embotar un bote? En una jornada laboral de 8 horas, ¿cuántos botes se hacen?

Línea	Botes/hora
1	150
2	120
3	167
4	135
5	113

Hay que calcular la media armónica:

$$\bar{x}_a = \frac{n}{\sum_{i=1}^{k} \frac{f_i}{x_i}} = \frac{5}{\frac{1}{150} + \frac{1}{120} + \frac{1}{167} + \frac{1}{135} + \frac{1}{113}} = 134\ botes/hora$$

El tiempo empleado para cada bote es:

$$\bar{t} = \frac{1}{\bar{x}_a} = \frac{1}{134} = 0{,}0075 \frac{horas}{bote} = 0{,}45\ \frac{minutos}{bote}$$

En 8 horas se hacen:

$$134 \cdot 8 = 1072\ botes$$

Ejemplo 19

Para saber cuántas personas ven al día un canal televisivo se han tomados dos muestras estratificadas de una población y se ha calculado la media y la desviación estándar. ¿Cuál es la media y la varianza en la población?

	Individuos	Media	Varianza
Estrato 1	500	2	4
Estrato 2	160	5	9

El tamaño de cada muestra estratificada es:

Estrato 1:

$$f_1 = \frac{n_1}{n} = \frac{500}{660} = 0{,}76$$

Estrato 2:

$$f_2 = \frac{n_2}{n} = \frac{160}{660} = 0{,}24$$

La media total es:

$$\bar{x} = \bar{x}_1 \cdot f_1 + \bar{x}_2 \cdot f_2 = 2 \cdot 0{,}76 + 5 \cdot 0{,}24 = 2{,}72$$

La varianza total se calcula como:

$$S^2 = f_1 \cdot S_1^2 + f_2 \cdot S_2^2 + f_1(\bar{x}_1 - \bar{x})^2 + f_2(\bar{x}_2 - \bar{x})^2$$
$$= 0{,}76 \cdot 4 + 0{,}24 \cdot 9 + 0{,}76(2 - 2{,}72)^2 + 0{,}24(5 - 2{,}72)^2 = 6{,}84$$

Ejemplo 20

A continuación se muestran las edades de los miembros de un curso:

22	18	17	21	21	29	36	16	23	16
25	25	29	23	16	23	52	33	17	23
35	16	32	31	40	23	27	26	22	17

A. Calcular la media, la mediana y la moda.
B. ¿Cuál es la edad máxima del 75% de los miembros?

Aplicando las fórmulas necesarias, el valor de la media es 25,13 y la mediana y la moda salen 23.

Para calcular la edad máxima hay que calcular el percentil 75 o el tercer cuartil, Q_3. Aplicando la fórmula necesaria se obtiene que el el 75% de los miembros tiene 29 o menos años.

Ejemplo 21

Una compañía telefónica A saca un anuncio que dice: «Si nuestros precios promedio no son iguales o menores que los de cualquier otra compañía, le sale la primera factura grati». Un cliente fue una tienda y dijo que había visto 4 tarifas, en función de los megas, con precios menores que los que esta en cuestión ofrecía y que iba a denunciar. Tales tarifas eran:

11,9	16,9	21,9	26,9

Los precios de la compañía A para esas mismas tarifas eran:

9,9	16,9	22,9	28,9

El trabajador le dice al cliente que el anuncio hace referencia a un promedio ponderado de las tarifas y que el promedio es menor porque las ventas de esas tarifas han sido:

25	20	17	12

¿Puede denunciar el cliente o queda resulto el asunto con la justificacion de las medias ponderadas?

Sin hacer medias ponderadas resulta que:

$$Media\ de\ la\ compañía\ A = \frac{9{,}9 + 16{,}9 + 22{,}9 + 28{,}9}{4} = 19{,}65$$

$$Media\ de\ la\ otra\ compañía = \frac{11{,}9 + 16{,}9 + 19{,}9 + 26{,}9}{4} = 19{,}4$$

Haciendo medias ponderadas resulta que:

$$Medida\ ponderada\ de\ la\ compañía\ A = \frac{9{,}9 \cdot 25 + 16{,}9 \cdot 20 + 22{,}9 \cdot 17 + 28{,}9 \cdot 12}{25 + 20 + 17 + 12}$$

$$= 17{,}86$$

$$Medida\ ponderada\ de\ la\ compañía\ A = \frac{11{,}9 \cdot 25 + 16{,}9 \cdot 20 + 19{,}9 \cdot 17 + 26{,}9 \cdot 12}{25 + 20 + 17 + 12}$$
$$= 17{,}98$$

Tal y como se puede observar el cliente tiene razón ya que el promedio sale mayor para la compañía A, pero con las medias ponderadas sale menor por lo que la empresa se cura en salud.

6.5. Gráficos y su interpretación

En este apartado se explicarán los tipos de diagramas más empleados cuando se trata de variables cuantitativas.

Historiogramas

Los histogramas son un tipo de diagramas de barras para variables cuantitativas. El término histograma se emplea cuando se trata de datos agrupados en intervalos. En el caso de datos no agrupados se llaman diagramas de barras. Tanto un diagrama de barras como un histograma de basa en una representación de las frecuencias, ya sean absolutas o relativas, frente a los datos. En el caso de los histogramas, si se considera igual amplitud de intervalos, la anchura de los rectángulos ha de ser la misma y las alturas coincidirán con las frecuencias de cada clase.

Ejemplo 22. Construcción de un histograma

Los datos de la siguiente tabla muestran los aciertos que han tenido un grupo de personas en un test. Construir un histograma.

I_i	f_i	h_i
0-10	33	0,236
10-20	25	0,179
20-30	45	0,321
30-40	17	0,121
40-50	5	0,036
50-60	15	0,107
Total	140	1

Tabla 12. Datos de aciertos en un test.

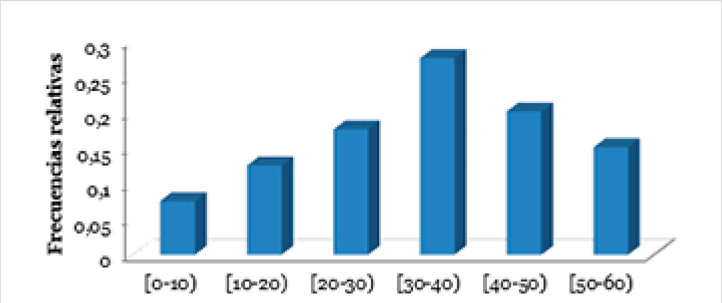


Gráfico 1. Histograma para datos agrupados en intervalos de amplitud constante.

Cuando la amplitud de los intervalos no es constante, la menor amplitud se toma como si fuera la unidad y se van levantando las alturas.

$$Frecuencia\ (área) = amplitud\ del\ intervalo\ \cdot altura$$

2.25.

Ejemplo 23. Construcción de un histograma de frecuencias relativas con datos agrupados en intervalos de diferente amplitud

La tabla a continuación muestra los datos de resultados en un test psicológico realizado a 90 personas. Construir el histograma de frecuencias relativas:

I_i	x_i	f_i	h_i
[0-10)	5	6	2/45
[10-15)	12,5	20	2/9
[15-20)	17,5	25	5/18
[20-30)	25	7	7/90
[30-35)	32,5	23	23/90
[35-50)	42,5	9	1/10
Total		90	1,000

Tabla 19. Datos de un test psicológico.

Se toma amplitud de valor 5 como la unidad. Las alturas son:

$$Frecuencia\ (área) = amplitud\ del\ intervalo\ \cdot altura$$

$$\text{Intervalo } [10-15): \quad \frac{2}{9} = 1 \cdot a_2 \rightarrow a_2 = \frac{2}{9}$$

$$\text{Intervalo } [15-20): \quad \frac{5}{18} = 1 \cdot a_3 \rightarrow a_3 = \frac{5}{18}$$

$$\text{Intervalo } [30-35): \quad \frac{23}{90} = 1 \cdot a_5 \rightarrow a_5 = \frac{23}{90}$$

Para los siguientes intervalos hay que calcular primero cuántas veces es mayor la amplitud del intervalo que el que se ha tomado como unidad y luego aplicar la fórmula. Los resultados son:

$$\text{Intervalo } [0-10): \quad \frac{2}{45} = 2 \cdot a_1 \rightarrow a_1 = \frac{1}{45}$$

$$\text{Intervalo } [20-30): \quad \frac{7}{90} = 2 \cdot a_4 \rightarrow a_4 = \frac{7}{180}$$

$$\text{Intervalo } [35-50): \quad \frac{1}{10} = 3 \cdot a_6 \rightarrow a_6 = \frac{1}{30}$$

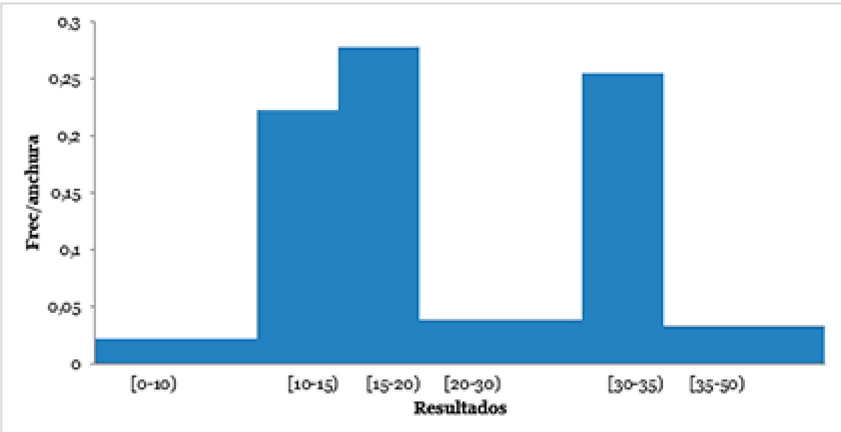


Gráfico 2. Histograma para datos agrupados en intervalos de amplitud variable.

Así mismo, se pueden hacer histogramas con las frecuenicas acumuladas, ya sean absolutas o relativas.

Polígonos de frecuencia

Los polígonos de frecuencias son gráficos que proporcionan la misma información que los histogramas, pero, en vez de ser diagramas de barras, son diagramas de puntos conectados por una línea, de manera que el área que dejan debajo coincide con el área de las barras del histograma y tiene forma poligonal, o curva cuando hay muchos datos. En estos gráficos, se dan dos puntos en los extremos en los que las frecuencias sean cero para cerrar el polígono con el eje horizontal.

Ejemplo 24. Construir un polígono de frecuencias

Construir un polígono de frecuencias con los datos del Ejemplo 22.

I_i	f_i	h_i
0-10	33	0,236
10-20	25	0,179
20-30	45	0,321
30-40	17	0,121
40-50	5	0,036
50-60	15	0,107
Total	140	1

Datos del ejemplo 22.

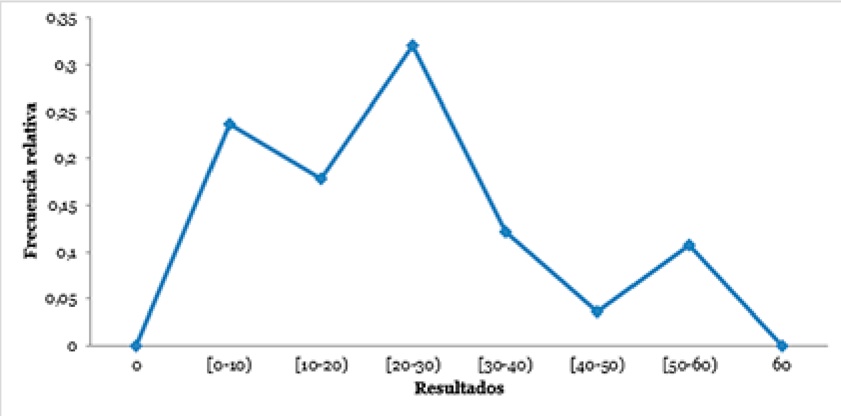
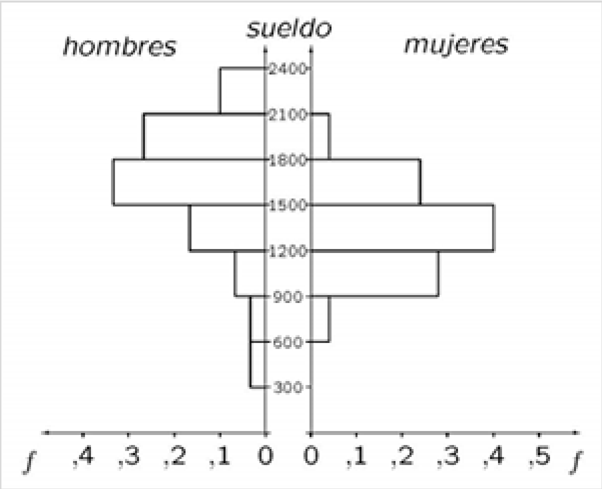


Gráfico 3. Polígono de frecuencias.

A partir del polígono de frecuencias se puede elaborar el correspondiente histograma haciendo líneas paralelas al eje horizontal que pasen por los puntos del polígono, de manera que cada punto caiga en la mitad de la barra.

Ejemplo 25. Interpretación de un histograma

¿Qué se puede decir sobre el salario de los hombres en relación al de las mujeres en función de los siguientes histogramas?



El sueldo medio de los hombres es más alto y se distribuye más asimétricamente.

Ejemplo 26. Interpretación de un diagrama de barras con datos discretos

El siguiente diagrama de barras muestra el número de agricultores jóvenes en unos pueblos de una comunidad autónoma. Responder a la siguientes preguntas:

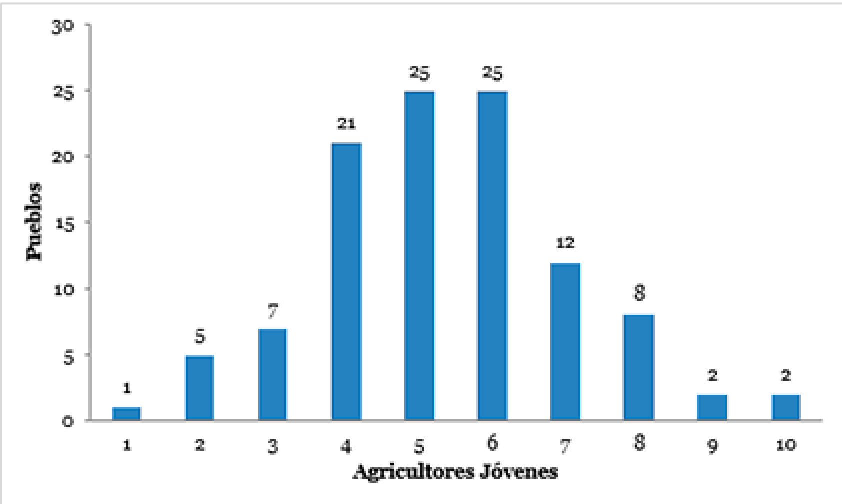


Gráfico 4. Diagrama de barras.

A. ¿Cuántos pueblos se han tomado para el estudio?
108 pueblos.

B. ¿Cuántos pueblos tienen como máximo 7 agricultores jóvenes?
92 pueblos.

C. ¿Cuál es el porcentaje de pueblos con 3 agricultores jóvenes?
6,5% (7/108)

D. ¿Cuál es el número más frecuente de agricultores?

5 y 6 agricultores en 25 empresas.

E. ¿Cuál es el número máximo de agricultores jóvenes en los 25 pueblos más pequeños, es decir, con menos trabajadores?

Se hace la tabla de frecuencias acumuladas:

Agricultores jóvenes	Pueblos	Frecuencia acumulada
1	1	1
2	5	6
3	7	13
4	21	34
5	25	59
6	25	84
7	12	96
8	8	104
9	2	106
10	2	108

En los 25 pueblos con menos agricultores el máximo número de los mismos es 4.

F. Si una empresa quiere suministrar fertilizantes y solo quiere visitar a aquellos pueblos con más de 7 agricultores jóvenes, ¿a qué porcentaje de pueblos irá?

A los que tienen más de 7 agricultores, 12/108=11,1%

G. Si el 50% de pueblos con mayor número de agricultores puede beneficiarse de una subvención, ¿qué mínimo de agricultores ha de haber en los pueblos para conseguir esa subvencion?

El 50% del 108 pueblos es 54. Comparando con las frecuencias acumuladas, el máximo número de agricultores es 5.

Ojivas

Las ojivas son tipos de diagramas para datos agrupados que dan información de las frecuencias acumuladas. Se muestran los datos que están por debajo o por encima de unos derminados valores en vez de los datos de los intervalos o las marcas de clase. Hay dos tipos de ojivas: ojivas 'menor que' y ojivas 'mayor que'. Las ojivas 'menor que' dan información sobre frecuencias menores que una que se está comparando y se inclinan positivamente, mientras que las ojivas 'mayor que', al revés. En la ojiva 'menor que' el extremo derecho no toca al eje horizontal, mientras que en la ojiva 'mayor que' es el extremo izquierdo. En el eje horizoantal se ponen los límites de los intervalos.

La frecuencia absoluta o relativa acumulada 'menor que' para la ojiva 'menor que' se calcula a partir de la primera frecuencia acumulada del primer intervalo y se acumula hacia abajo, mientras que la frecuencia absoluta o relativa 'mayor que' para la ojiva 'mayor que' se calcula a partir de la última frecuencia acumulada del último intervalo y se acumula hacia arriba.

Ejemplo 27. Construir una ojiva 'menor que' y otra 'mayor que'

A partir de los datos del ejemplo 22 hacer una ojiva 'menor que' y una ojiva 'mayor que'.

Ojiva 'menor que'		Ojiva 'mayor que'	
Clase	F_i (menor que)	Clase	F_i (mayor que)
Menor que 0	0	Mayor que 0	140
Menor que 10	33	Mayor que 10	107
Menor que 20	58	Mayor que 20	82
Menor que 30	103	Mayor que 30	37
Menor que 40	120	Mayor que 40	20
Menor que 50	125	Mayor que 50	15
Menor que 60	140	Mayor que 60	0

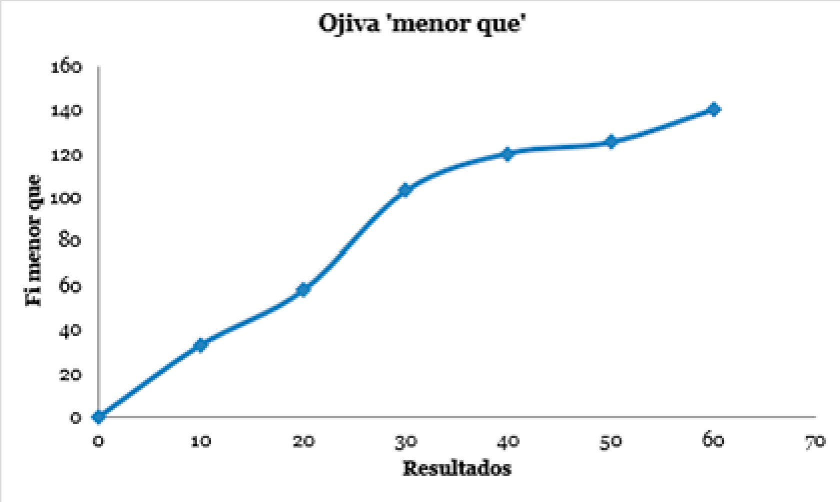


Gráfico 5. Ojiva 'menor que'.

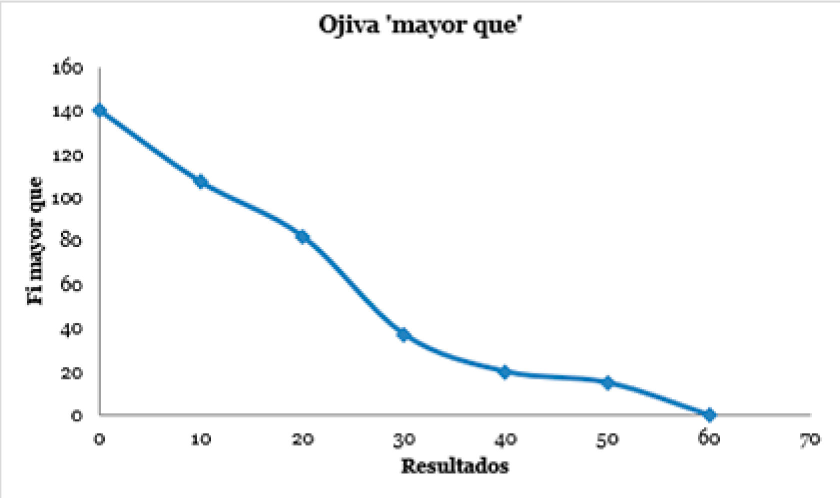


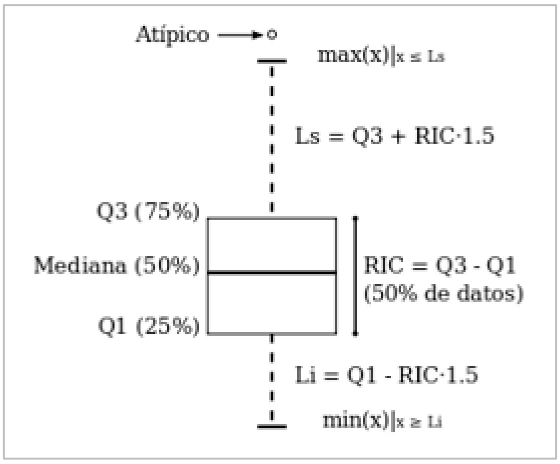
Gráfico 6. Ojiva 'mayor que'.

Si aquellos participantes que hayan tenido menos de 35 aciertos en un test se quedan sin poder asitir a la siguiente fase, ¿qué porcentaje de participantes se quedará sin asistir? Interpolando en la ojiva 'menor que' se obtiene que son 105 participantes.

Así mismo, se puede hacer ojivas con las frecuencias relativas acumuladas.

Diagramas de caja

Un diagrama de caja es un tipo de representación basada en los cuartiles de una distribución. Su representación es una caja, donde quedan los valores de los cuartiles y unos brazos llamados bigotes que van hasta los límites de la distribución. Son ampliamente empleados para detectar la presencia de algún dato anómalo dentro de un conjunto de datos. En la siguiente figura, se puede ver una representación y la información que proporciona un diagrama de cajas.



Fuente: wikipedia.org

Li y *Ls* son los límites inferior y superior y *RIC* el rango intercuartílico.

Para hacer un diagrama de cajas primero hay que ordenar los datos, después calcular los cuartiles y el rango intercuartílico, seguidamente calcular los límites superior e inferior, (tal y como se muestra en la figura) y, por último, dibujar el diagrama para lo cual hay que calcular los anchos de los rectángulos de la caja. Si algún valor cae fuera de los límites superior o inferior permitidos, se toma como dato atípico.

Ejemplo 27. Elaborar un diagrama de cajas

A partir de los datos presentados a continuación elaborar un diagrama de cajas e interpretar los datos.

34	48	49	51	52	54	55	71	76

Calculamos la mediana y los cuartiles 1 y 3.

$$Posición\ de\ la\ Mediana = \frac{n+1}{2} = \frac{9+1}{2} = 5$$

La mediana es el dato 52.

$$Posición\ de\ Q_1 = \frac{n+1}{4} = \frac{9+1}{4} = 2{,}5$$

Q_1 es 48,5.

$$Posición\ de\ Q_3 = \frac{3(n+1)}{4} = \frac{3(9+1)}{4} = 7{,}5$$

Q_3 es 63.

Calculamos el rango intercuartílico:

$$Rango\ intercuartílico = Q_3 - Q_1 = 63 - 48{,}5 = 14{,}5$$

Calculamos los límites inferior y superior:

$$L_i = Q_1 - 1{,}5 \cdot RIC = 48{,}5 - 1{,}5 \cdot 14{,}5 = 26{,}75$$

$$L_s = Q_3 + 1{,}5 \cdot RIC = 63 + 1{,}5 \cdot 14{,}5 = 84{,}75$$

A continuación, se calcula la anchura de cada rectángulo del diagrama:

	Valores	Anchos
Mínimo	34	34
Q_1	48,5	14,5
Mediana	52	3,5
Q_3	63	11
Máximo	76	13

Se dibuja el diagrama de cajas:

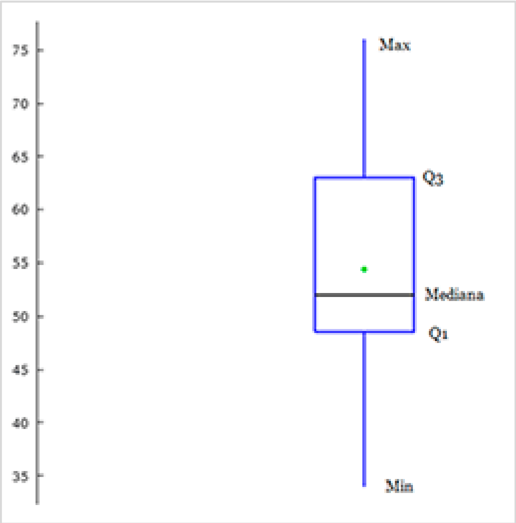


Gráfico 7. Diagrama de cajas.

En el diagrama de cajas se puede ver que no hay ningún valor que exceda los valores de los límites superior ni inferior, por lo que no hay ningún valor atípico.

Lo + recomendado

No dejes de leer...

Estadística y computación en la creación de contenido para el periodismo de datos

Flores, J. (29 de noviembre de 2014). *Estadística y computación en la creación de contenido para el periodismo de datos* [Mensaje en un blog].

Este post contiene información sobre el uso de la estadística y la computación o la informática para obtener contenidos en la rama del periodismo de datos.

Accede al artículo a través del aula virtual o desde la siguiente dirección web:
http://internetmedialab.com/2014/11/29/estadistica-y-computacion-en-la-creacion-de-contenido-para-el-periodismo-de-datos/

No dejes de ver...

Estadística descriptiva en el periodismo

En esta web se recogen varios vídeos sobre los principios básicos de la estadística descriptiva en el periodismo de datos.

Accede al vídeo a través del aula virtual o desde la siguiente dirección web:
https://itunes.apple.com/es/itunes-u/estadistica-umh-1200/id619865416?mt=10

+ Información

A fondo

La estadística de movimientos turísticos en fronteras

Este artículo muestra los aspectos más relevantes de la estadística de movimientos turísticos en fronteras.

Accede al artículo a través del aula virtual o desde la siguiente dirección web:
http://dialnet.unirioja.es/servlet/articulo?codigo=2167502

Bibliografía

Flores, J. M., y Salinas, C. (2013). El periodismo de datos como especialización de las organizaciones de noticias en Internet. *Correspondencias & Análisis*, 3, 15-34.

Test

1. El crecimiento de empleo en los últimos 7 meses ha sido el siguiente: 0,13, 0,10, 0,065, 0,07, 0,077, 0,12 y 0,14. ¿Cuál es el incremento promedio porcentual?

A. 10%.

B. 10,5%.

C. 0,1%.

D. Ninguno es correcto.

2. Se invirtieron 10.000 euros al 2,3%, 3,7%, 4,1% y 4,5% en cuatro años. ¿Cuál fue el factor de crecimiento medio?

A. Ninguno de los anteriores.

B. 3,6%.

C. 1,036.

D. 0,036%.

3. ¿Cuál es la productividad media de una empresa de pinturas si las 5 secciones que tiene fabrican 20, 35, 27, 45 y 62 botes a la hora?

A. 37,8.

B. 32,5.

C. 35.

D. 32.

4. ¿Cuál es la mediana de estos datos?

Intervalo	f_i
[2-9)	15
[9-13)	25
[13-20)	14
[20-24)	10
[24-31)	3

A. 11,9.

B. 34.

C. 12.

D. 40.

5. ¿Cuál es el percentil 75 de los datos de la esta tabla?

Intervalo	f_i
[2-9)	15
[9-13)	25
[13-20)	14
[20-24)	10
[24-31)	3

A. 51.
B. 18.
C. 15,8.
D. Ninguna es correcta.

6. ¿Para qué empresa es más representativo el salario medio anual?

Empresa A		Empresa B	
Salario	f_i	Salario	f_i
8000-10000	23	10.000-13.000	36
10.000-13.000	19	13.000-16.000	45
13.000-17.000	12	16.000-30.000	11
17.000-25.000	7	30.000-45.000	3
25.000-40.000	5	45.000-60.000	1

A. La A.
B. La B.
C. Las dos igual.
D. No se sabe.

7. Los salarios de dos categorías de una empresa son:

	Categoría 1	Categoría 2
Trabajadores	10	20
Salario medio	1800	850
Moda	1600	790
Desviación estándar	340	180

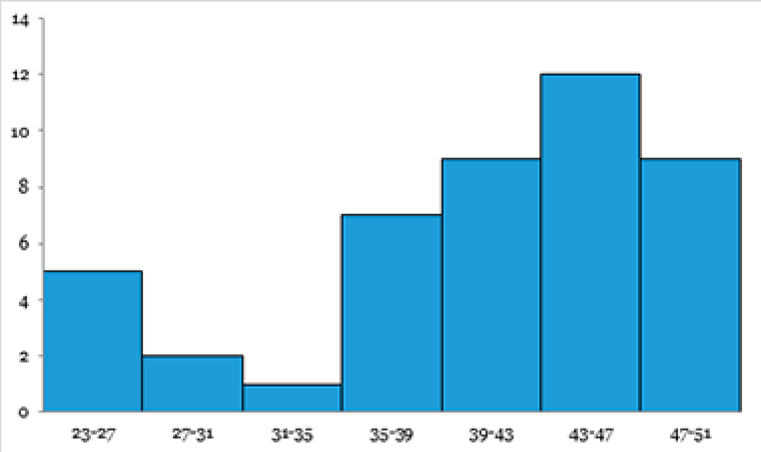
A. 1325 euros.
B. 1195 euros.
C. No se puede saber.
D. 1166 euros.

8. El rango intercuartil de los siguientes datos es:

x_i	n_i	N_i
0	13	13
1	11	24
2	16	40
3	27	67
4	15	82
5	15	97
	97	

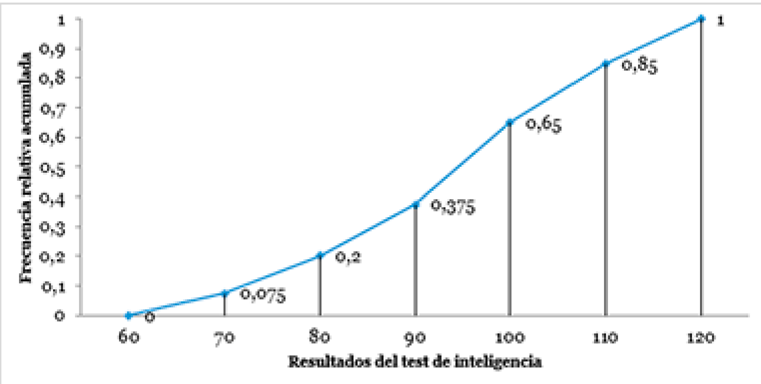
A. 2.
B. 1.
C. 4.
D. Ninguno de los valores.

9. A partir de los datos del siguiente gráfico se puede decir que:



A. La moda es igual que la mediana.
B. La moda es mayor que la mediana y esta que la media.
C. La media y la mediana son iguales.
D. La moda y la media son iguales.

10. La siguiente ojiva muestra los resultados de un test de inteligencia. Sabiendo que a partir de 110 se considera que una persona es superdotada, ¿qué afirmación es cierta?



A. En torno al 15% son superdotados.
B. El 20% ha obtenido 80 en la puntuación.
C. El 65% han obtenido menos de 100 en la puntación.
D. A y C son ciertas.