CAPÍTULO 13: ESTADÍSTICA

1 IDENTIFICACIÓN DE LAS FASES Y TAREAS DE UN ESTUDIO ESTADÍSTICO

Nos enfrentamos a diario a la necesidad de recoger, organizar e interpretar datos y esta necesidad aumentará en el futuro, debido al desarrollo de los sistemas de comunicación y las bases de datos. Es notable el aumento del uso de las redes sociales tales como *Youtube* o *Facebook*, donde las personas tienen oportunidad de presentar información sobre ellos mismos, y de páginas web donde se pueden encontrar y descargar gran variedad de datos estadísticos sobre diversos temas de actualidad: resultados deportivos de sus equipos favoritos, temperatura máxima y mínima a lo largo de un mes, ventas de turrón la pasada navidad, etc. En otras ocasiones los datos son recogidos por el investigador mediante la realización de una encuesta o a través de un experimento. La encuesta requerirá la elaboración de un cuestionario, fijando los objetivos del mismo, eligiendo las variables explicativas y redactando las preguntas que permitan obtener la información deseada de una forma clara y concisa.

En este sentido, la estadística ha jugado un papel primordial en este desarrollo tecnológico que nos está tocando vivir, al proporcionar herramientas metodológicas generales para analizar la variabilidad, determinar relaciones entre variables, diseñar de forma óptima experimentos, mejorar las predicciones y la toma de decisiones en situaciones de incertidumbre.

El tratamiento estadístico de un problema comienza siempre con la presentación de la magnitud que se quiere analizar de una determinada población y la selección de la muestra pertinente para pasar a la recogida de datos. Una vez obtenidos los datos se ordenan y presentan en tablas o gráficas, de forma que sea posible observar las particularidades que señalan.

De aquí se puede considerar que un estudio estadístico consta de una serie de fases y tareas bien diferenciadas:

1. Definición de la población y característica a estudiar.
   Tareas: Identificación de las características cuantitativas y cualitativas; fijación de la población; especificación de la forma de recogida de datos (entrevistas, teléfono, correo electrónico, etc.).
2. Selección de la muestra.
   Tareas: Identificación del tamaño de la muestra y presupuesto necesario.
3. Recogida de datos.
   Tareas: Diseño del cuestionario; diseño muestral.
4. Organización y representación gráfica.
   Tareas: Tablas y gráficas que ayuden a una más fácil interpretación de los datos; esto consiste en un estudio de cada variable, la tabulación y representación (es) gráfica (s) más apropiada (s).
5. Análisis de datos.
   Tareas: Tratamiento de los datos. Esto consistirá en un análisis descriptivo de los datos y/o un análisis multivariante de los datos, dependiendo del tipo de estudio a realizar y costes del mismo.
6. Obtención de conclusiones.
   Tareas: recomendaciones y toma de decisiones a partir de las conclusiones.

Ejemplo:

- Una lista de puntos a tener en cuenta al plantear las preguntas de investigación es la siguiente:
  - ¿Qué quieres probar? ¿Qué tienes que medir /observar /preguntar?
  - ¿Qué datos necesitas? ¿Cómo encontrarás tus datos? ¿Qué harás con ellos?
  - ¿Crees que puedes hacerlo? ¿Encontrarás problemas? ¿Cuáles?
  - ¿Para qué te servirán los resultados?

De esta manera se preparará una lista de las características que queremos incluir en el estudio, analizando las diferentes formas con las que podrían obtenerse los datos. Por simple observación: como el sexo, color de pelo y ojos, si el alumno usa o no gafas; Si se requiere una medición: como el peso, talla, perímetro de cintura; si habría que preguntar, es decir, si se debe realizar una encuesta: cuánto deporte practica, número del calzado, cuantas horas duerme, cuantas horas estudia al día o a la semana, etc.

Por tanto, es importante considerar la naturaleza de las escalas de medida y tipo de variable estadística, puesto que de ellas depende el método de análisis de datos que se puede aplicar. La elección del conjunto de datos es crítica, pues dependiendo del tipo de datos la gama de técnicas estadísticas será más o menos amplia, ya que no todas las técnicas son aplicables a cualquier tipo de dato.

Autora: Raquel Caro
Revisoras: María Molero y Nieves Zuasti
Ilustraciones: Banco de Imágenes de INTEF

2. POBLACIÓN Y MUESTRA. VARIABLES ESTADÍSTICAS

2.1. Población

Población estadística, colectivo o universo es el conjunto de todos los individuos (personas, objetos, animales, etc.) que contengan información sobre el fenómeno que se estudia.

Ejemplo:

- Si estudiamos el precio de la vivienda en una ciudad, la población será el total de las viviendas de dicha ciudad.

Actividades resueltas

- Se va a realizar un estudio estadístico sobre el porcentaje de personas casadas en la península. Para ello no es factible estudiar a todos y cada uno de los habitantes por razones de coste y de rapidez en la obtención de la información. Por lo tanto, es necesario acudir a examinar sólo una parte de esta población. Esa parte es la muestra elegida.

2.2 Muestra

Muestra es un subconjunto representativo que se selecciona de la población y sobre el que se va a realizar el análisis estadístico. El tamaño de la muestra es el número de sus elementos. Cuando la muestra comprende a todos los elementos de la población, se denomina censo.

Ejemplo:

- Si se estudia el precio de la vivienda de una ciudad, lo normal será no recoger información sobre todas las viviendas de la ciudad (ya que sería una labor muy compleja y costosa), sino que se suele seleccionar un subgrupo (muestra) que se entienda que es suficientemente representativo.

Actividades propuestas

1. Señalar en qué caso es más conveniente estudiar la población o una muestra:
   a) El diámetro de los tornillos que fabrica una máquina diariamente.
   b) La altura de un grupo de seis amigos.
2. Se puede leer el siguiente titular en el periódico que publica tu instituto: "*La nota media de los alumnos de 4º ESO de la Comunidad de Madrid es de* 7,9". ¿Cómo se ha llegado a esta conclusión? ¿Se ha estudiado a toda la población? Si hubieran seleccionado para su cálculo solo a las mujeres, ¿sería representativo su valor?

2.3. Individuo o unidad estadística

Individuo o unidad estadística es cualquier elemento que contenga información sobre el fenómeno que se estudia.

Ejemplo:

- Si estudiamos las notas de los alumnos de una clase, cada alumno es un individuo; si estudiamos el precio de la vivienda, cada vivienda es una unidad estadística.

2.4. Variable estadística

En general, supondremos que se está analizando una determinada población, de la que nos interesa cierta característica, representada por una variable observable o estadística X. Las variables que están bajo estudio se pueden clasificar en dos categorías:

Variables cualitativas o atributos (datos no métricos), que no se pueden medir numéricamente. Las escalas de medida no métricas se clasifican en nominales (o categóricas) y ordinales.

Variables cuantitativas, que tienen un valor numérico. Este tipo de variables son las que aparecen con más frecuencia y permiten un análisis más detallado que las cualitativas. Dentro de las variables cuantitativas, se pueden distinguir las variables discretas y las variables continuas. Las variables discretas toman valores aislados, mientras que las variables continuas pueden tomar cualquier valor dentro de un intervalo.

Ejemplo:

- Ejemplos de variables cualitativas son la nacionalidad o la raza de un conjunto de personas.
- Ejemplos de variables cuantitativas son las notas obtenidas en una asignatura, el peso o altura de un conjunto de personas.
- Ejemplos de variables discretas son el número de alumnos que aprueban una asignatura, o el número de componentes defectuosos que se producen al día en una fábrica.
- Ejemplos de variables continuas son el tiempo que tardamos en llegar al instituto desde nuestra casa o la velocidad de un vehículo.

Autora: Raquel Caro
Revisoras: María Molero y Nieves Zuasti
Ilustraciones: Banco de Imágenes de INTEF

Actividades resueltas

- Se va a realizar un estudio estadístico sobre el porcentaje de personas con hijos en una localidad madrileña de 134.678 habitantes. Para ello se eligen 2.346 habitantes y se extienden las conclusiones a toda la población. Identificar variable estadística, población, muestra, tamaño muestral e individuo.
  - *Variable estadística*: si una persona tiene hijos o no.
  - *Población: Los 134.678 habitantes de la localidad.*
  - *Muestra: Los 2.346 habitantes elegidos.*
  - *Tamaño muestral: 2.346 personas.*
  - *Individuo: Cada persona que se le pregunte.*

Actividades propuestas

3. Indicar el tipo de variable estadística que estudiamos y razona, en cada caso, si sería mejor analizar una muestra o la población:
   a) El sexo de los habitantes de un país.
   b) El dinero gastado a la semana por tu hermano.
   c) El color de pelo de tus compañeros de clase.
   d) La temperatura de tu provincia.
   e) La talla de pie de los alumnos del instituto.
4. Para realizar un estudio hacemos una encuesta entre los jóvenes de un barrio y les preguntamos por el número de veces que van al cine al mes. Indica qué características debería tener la muestra elegida y si deberían ser todos los jóvenes de la muestra de la misma edad.

3. TABLAS DE FRECUENCIAS

3.1. Frecuencia absoluta

Cuando se analiza una *variable discreta*, la información resultante de la muestra se encuentra resumida habitualmente en una tabla o distribución de frecuencias. Supongamos que se ha tomado una muestra de tamaño N en la que se han identificado k valores (o modalidades) distintos $x_1, x_2, \ldots, x_k$. Cada uno de ellos se produce con una frecuencia absoluta n_i, es decir, el número de veces que aparece en la muestra.

La información obtenida se puede resumir en una tabla de frecuencias.

Las tablas de frecuencia también se utilizan para representar información de una *variable continua* procedente de una muestra en la que se agrupan las observaciones en intervalos, que se denominan intervalos de clase L_i o celdas.

Aunque este procedimiento supone, de hecho, una pérdida de información, esta pérdida no es de magnitud importante y se ve compensada con la agrupación de la información y la facilidad de interpretación que proporciona una tabla de frecuencias.

En este caso, los valores *xi* se corresponden con el punto medio del intervalo y se denominan marcas de clase.

Ejemplo:

- Cuando realizamos un estudio sobre el ocio y encuestamos a 40 jóvenes de una localidad sobre el número de veces que van al cine los resultados de dicha encuesta los podemos recoger en una tabla para resumir dicha información.

Actividades resueltas

- Se está realizando un control del peso de un grupo de niños. Para ello, se contabilizan el número de veces que comen al día una chocolatina 13 niños durante un mes, obteniendo los siguientes números: 2, 5, 3, 2, 0, 4, 1, 7, 4, 2, 1, 0, 2.

La información obtenida se puede resumir en una tabla de frecuencias absolutas:

Valores	0	1	2	3	4	5	6	7
Frecuencia absoluta	2	2	4	1	2	1	0	1

- En una fábrica se realiza un estudio sobre el espesor, en mm, de un cierto tipo de latas de refresco. Con este fin, selecciona una muestra de tamaño $N = 25$, obteniendo los siguientes valores: 7.8, 8.2, 7.6, 10.5, 7.4, 8.3, 9.2, 11.3, 7.1, 8.5, 10.2, 9.3, 9.9, 8.7, 8.6, 7.2, 9.9, 8.6, 10.9, 7.9, 11.1, 8.8, 9.2, 8.1, 10.5.

Esta información se puede resumir en la siguiente tabla de frecuencias, con 5 intervalos: (7, 8], (8, 9], (9, 10], (10, 11], (11, 12], siendo las marcas de clase los puntos medios de cada intervalo: 7,5; 8,5; 9,5; 10,5; 11,5. Comprueba que las frecuencias absolutas son las indicadas en la tabla:

Intervalos de clase	(7,8]	(8,9]	(9,10]	(10,11]	(11,12]
Marcas de clase	7.5	8.5	9.5	10.5	11.5
Frecuencia absoluta	6	8	5	4	2

Autora: Raquel Caro
Revisoras: María Molero y Nieves Zuasti
Ilustraciones: Banco de Imágenes de INTEF

Actividades propuestas

5. Obtener la tabla de frecuencias absolutas de las notas en inglés de 24 alumnos:

6	6	7	8	4	9	8	7	6	5	3	5
7	6	6	6	5	4	3	9	8	8	4	5

3.2. Frecuencia relativa

Se denomina **frecuencia relativa** (f_i) de un valor de la variable al cociente entre la frecuencia absoluta y el número total de observaciones N. Se escribe:

$$f_i = \frac{n_i}{N} \leq 1$$

Ejemplo:

- De la misma manera podemos recoger la información obtenida a partir de una encuesta a 40 jóvenes de una localidad sobre el número de veces que van al cine mediante porcentaje del número de veces que se repite un valor de la variable sobre el total.

Actividades resueltas

- Se está realizando un control del peso de un grupo de niños. Para ello, se contabilizan el número de veces que comen al día una chocolatina 13 niños durante un mes, obteniendo los siguientes números: 2, 5, 3, 2, 0, 4, 1, 7, 4, 2, 1, 0, 2.

La información obtenida se puede resumir en una tabla de frecuencias relativas:

Valores	0	1	2	3	4	5	6	7
Frecuencia relativa	0.154	0.154	0.307	0.077	0.154	0.077	0	0.077

Actividades propuestas

6. Construir una tabla de frecuencias relativas con el color de pelo de 24 personas elegidas al azar:

M=moreno; R=rubio; P=pelirrojo

M	R	P	R	R	R	R	P	P	M	M	M	M	R	R
	R	R	R	M	M	M	M	M	P					

3.3. Frecuencia absoluta acumulada

Se denomina **frecuencia absoluta acumulada** de un valor de la variable N_i a la suma de todas las frecuencias absolutas de los valores menores o iguales que él. Se calcula como:

$$N_i = \sum_{j=1}^{i} \left[n_j \right]$$

Se verifica la siguiente relación entre los valores de N_i:

$$N_1 \leq N_2 \leq \cdots \leq N_k = N$$

Ejemplo:

- De la misma manera podemos recoger la información obtenida a partir de una encuesta a 40 jóvenes de una localidad sobre el número de veces que van al cine mediante el número acumulado de veces que se repite un valor de la variable sobre el total.

Actividades resueltas

- Se está realizando un control del peso de un grupo de niños. Para ello, se contabilizan el número de veces que comen al día una chocolatina 13 niños durante un mes, obteniendo los siguientes números: 2, 5, 3, 2, 0, 4, 1, 7, 4, 2, 1, 0, 2.

La información obtenida se puede resumir en una tabla de frecuencias absolutas:

Valores	0	1	2	3	4	5	6	7
Frecuencia absoluta	2	2	4	1	2	1	0	1
Frecuencia absoluta acumulada	2	4	8	9	11	12	12	13

Actividades propuestas

7. El número de horas diarias de estudio de 14 alumnos es el siguiente:

3	4	2	5	3	4	3	2	3	4	5	4	3	2

a) Efectúa un recuento y organiza los resultados obtenidos en una tabla de frecuencias absolutas acumuladas.

b) ¿Qué significan las frecuencias acumuladas que has calculado?

Autora: Raquel Caro
Revisoras: María Molero y Nieves Zuasti
Ilustraciones: Banco de Imágenes de INTEF

3.4. Frecuencia relativa acumulada

Se denomina **frecuencia relativa acumulada** (F_i) de un valor de la variable a la suma de todas las frecuencias relativas de los valores menores o iguales que él. Se calcula como:

$$F_i = \sum_{j=1}^{i} [\, f_j \,]$$

Se verifica la siguiente relación entre los valores de F_i:

$$F_1 \leq F_2 \leq \cdots \leq F_k = 1$$

Ejemplo:

- De la misma manera podemos recoger la información obtenida a partir de una encuesta a 40 jóvenes de una localidad sobre el número de veces que van al cine mediante el porcentaje acumulado del número de veces que se repite un valor de la variable sobre el total.

Actividades resueltas

- Se está realizando un control del peso de un grupo de niños. Para ello, se contabilizan el número de veces que comen al día una chocolatina 13 niños durante un mes, obteniendo los siguientes números: 2, 5, 3, 2, 0, 4, 1, 7, 4, 2, 1, 0, 2.

La información obtenida se puede resumir en una tabla de frecuencias absolutas:

Valores	0	1	2	3	4	5	6	7
Frecuencia relativa	0.154	0.154	0.307	0.077	0.154	0.077	0	0.077
Frecuencia relativa acumulada	0.154	0.308	0.615	0.692	0.846	0.923	0.923	1

- En una fábrica se realiza un estudio sobre el espesor, en mm, de un cierto tipo de latas de refresco. Con este fin, selecciona una muestra de tamaño N = 25, obteniendo los siguientes valores: 7.8, 8.2, 7.6, 10.5, 7.4, 8.3, 9.2, 11.3, 7.1, 8.5, 10.2, 9.3, 9.9, 8.7, 8.6, 7.2, 9.9, 8.6, 10.9, 7.9, 11.1, 8.8, 9.2, 8.1, 10.5.

Esta información se puede resumir en la siguiente tabla de frecuencias, con 5 intervalos:

Intervalos de clase	(7,8]	(8,9]	(9,10]	(10,11]	(11,12]
Marcas de clase	7.5	8.5	9.5	10.5	11.5
Frecuencia absoluta	6	8	5	4	2
Frecuencia relativa	0.24	0.32	0.2	0.16	0.08
Frecuencia relativa acumulada	0.24	0.56	0.76	0.92	1

- Se organiza en una tabla la información recogida de las estaturas, en cm, de un grupo de 20 niñas:

130 127 141 139 138 126 135 138 134 131 143 140
129 128 137 136 142 138 144 136

La estatura es una variable estadística cuantitativa continua. Por tanto, podemos agrupar los valores de la variable en intervalos que llamamos clases o celdas. La amplitud de cada intervalo viene dada por la fórmula:

$$\frac{Máx - Mín}{\sqrt{N}}$$

En nuestro caso concreto tenemos que:

$$\frac{144-126}{\sqrt{20}} = 4.02$$

Aproximando, la amplitud de cada intervalo es de 5 cm.

Estatura en intervalos	**[125-130)**	**[130-135)**	**[135-140)**	**[140-145)**
Frecuencia absoluta	4	3	8	5
Frecuencia relativa	0.2	0.15	0.4	0.25
Frecuencia absoluta acumulada	4	7	15	20
Frecuencia relativa acumulada	0.2	0.35	0.75	1

Actividades propuestas

8. En una evaluación, de los 30 alumnos de una clase, el 30 % aprobó todo, el 10 % suspendió una asignatura, el 40 % suspendió dos asignaturas y el resto más de dos asignaturas.
   a) Realiza la tabla de frecuencias completa correspondiente (frecuencias absolutas, frecuencias relativas, frecuencias absolutas acumuladas y frecuencias relativas acumuladas).
   b) ¿Hay algún tipo de frecuencia que corresponda a la pregunta de cuantos alumnos suspendieron menos de dos asignaturas? Razona la respuesta.

Autora: Raquel Caro
Revisoras: María Molero y Nieves Zuasti
Ilustraciones: Banco de Imágenes de INTEF

4. GRÁFICOS ESTADÍSTICOS

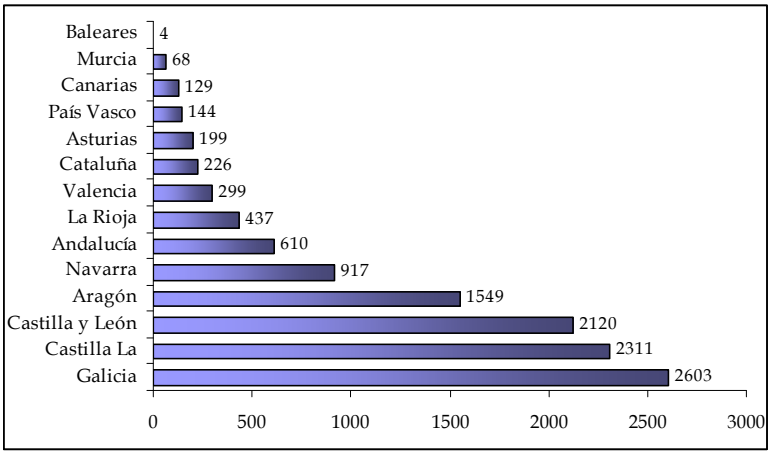
4.1. Diagrama de barras

Existen numerosas maneras de representar gráficamente la información que se ha obtenido de una muestra, dependiendo del tipo de variable que se esté analizando y del fin que se persiga con la representación.

Cuando se quiere representar gráficamente una variable cualitativa (atributo) o una variable cuantitativa discreta se puede utilizar *los diagramas de barras o rectángulos*. Se colocan los valores de la variable (las modalidades del atributo o valores de la variable discreta) en el eje de abscisas y, en el eje de ordenadas las frecuencias (absolutas o relativas). Sobre cada valor se levanta una barra o rectángulo cuya altura es igual a la frecuencia. Por comodidad, a veces también se suelen intercambiar los ejes.

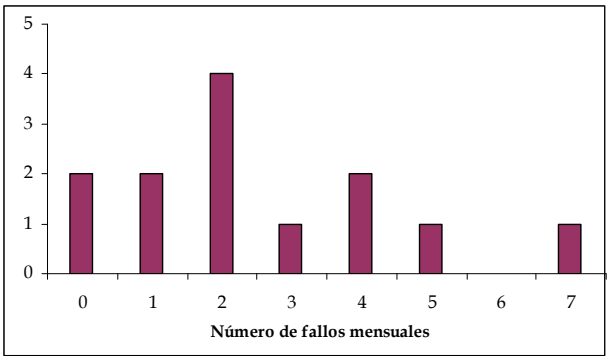
Ejemplo:

- Se ha representado gráficamente la potencia eólica (fuente de energía eléctrica renovable) instalada en España por Comunidad Autónoma en Enero de 2014 (en Megavatios)



Ejemplo:

- Se ha representado gráficamente el número de fallos mensuales de una máquina de helados



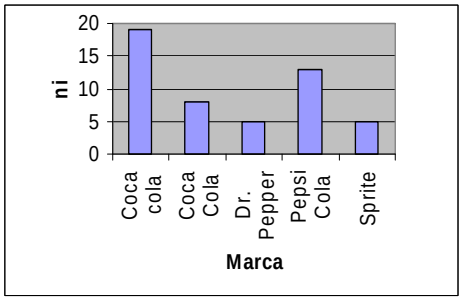
Actividades resueltas

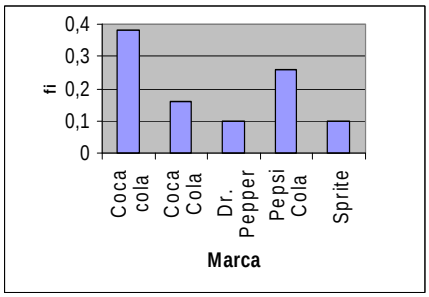
- Dada la siguiente información correspondiente a las preferencias de 50 adolescentes americanos respecto a la marca de refresco que consumen, construye la tabla asociada a estos datos y represéntalos gráficamente en un diagrama de barras de frecuencias absolutas y otro de frecuencias relativas.

COCA-COLA=CC; COCA-COLA LIGHT=CCL; DR.PEPPER=A; PEPSI-COLA=PC, SPRITE=S

CCL	CC	S	A	CC	CC	A	CC	P	CC
S	CCL	P	CCL	CC	CC	CCL	P	P	A
S	S	CC	CC	CC	A	P	CC	CCL	CC
CCL	CC	P	P	P	CCL	P	S	P	CC
CC	P	CCL	CC	CC	P	CC	P	CC	A

Marca	n_i	f_i
Coca Cola	19	0,38
Coca Cola Light	8	0,16
Dr. Pepper	5	0,10
Pepsi Cola	13	0,26
Sprite	5	0,10
	50	1

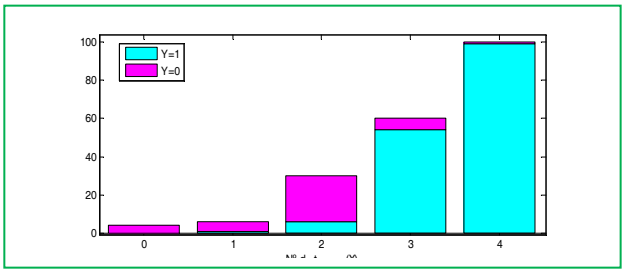




Autora: Raquel Caro
Revisoras: María Molero y Nieves Zuasti
Ilustraciones: Banco de Imágenes de INTEF

Actividades propuestas

9. Si queremos representar conjuntamente valores de la variable correspondientes a diferentes períodos de tiempo, o a distintas cualidades, para comparar situaciones podemos construir un diagrama de barras apiladas. ¿Podrías interpretar este gráfico correspondiente al número de temas que los alumnos de una asignatura de 4º ESO llevan estudiados? Se toma información en dos clases de un instituto (azul y rosa).



10. El sexo de 18 bebés nacidos en un hospital de Madrid ha sido:

H	M	H	H	M	H
H	M	M	H	M	H
M	M	H	H	M	H

Construye la tabla asociada a estos datos y represéntalos.

11. Representa los valores de la variable de la tabla adjunta con el gráfico adecuado correspondientes a una encuesta realizada sobre el sector al que pertenecen un grupo de trabajadores madrileños.

SECTOR	INDUSTRIAL	AGRARIO	SERVICIOS	OTROS
% TRABAJADORES	20	16	45	19

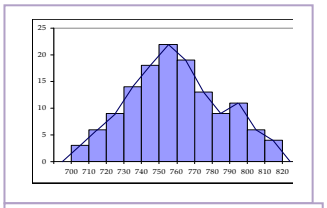
4.2. Histogramas

La representación más utilizada en variables cuantitativas continuas es el **histograma**. En el eje de abscisas se colocan los diferentes intervalos en los que se agrupan las observaciones de la variable. Sobre estos intervalos, se levantan rectángulos cuya área es proporcional a la frecuencia observada en cada uno de ellos. En el caso que todos los intervalos tengan la misma amplitud basta con que los rectángulos sean proporcionales a la frecuencia.

Dependiendo de las frecuencias que se utilicen, se tratará de un histograma de frecuencias relativas, o bien de un histograma de frecuencias absolutas.

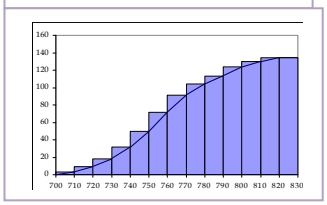
En ocasiones, se unen los puntos medios de los segmentos superiores de los rectángulos, obteniéndose de este modo los **polígonos de frecuencias**, ya sean absolutas o relativas. Estos polígonos se construyen utilizando un intervalo anterior al primero (de la misma longitud que éste) y otro posterior al último (de su misma longitud). De esta manera, los polígonos delimitan un área cerrada.

En ambos casos, también se pueden utilizar las frecuencias acumuladas para construir los respectivos histogramas. Estos histogramas también llevan asociados los correspondientes polígonos de frecuencias, que en este caso se construyen uniendo los vértices superiores derechos de cada uno de los intervalos.



Ejemplo:

- Se ha representado gráficamente la información obtenida a partir de las emisiones específicas de CO_2 de una central de carbón (kg/megavatio-hora) a partir de un histograma y un polígono de frecuencias absolutas.



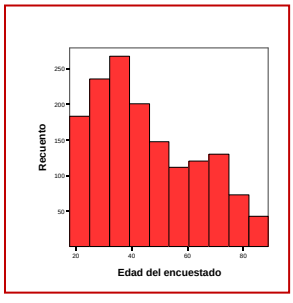
Ejemplo:

- Se ha representado gráficamente la información obtenida a partir de las emisiones específicas de CO_2 de una central de carbón (kg/megavatio-hora) a partir de un histograma y un polígono de frecuencias acumuladas absolutas.

Actividades propuestas

12. Completa la tabla de frecuencias para poder representar la información mediante el histograma de frecuencias acumuladas:

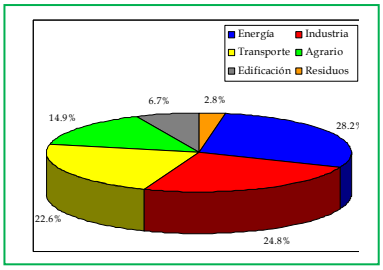
EDAD	[15, 25)	[25, 35)	[35, 45)	[45, 55)
NÚMERO DE PERSONAS	25	45	55	65



13. ¿A qué representación gráficas corresponden el siguiente gráfico correspondiente a la información recogida sobre la edad de 100 personas? ¿Por qué crees que se ha utilizado este y no otro?

4.3. Diagrama de sectores

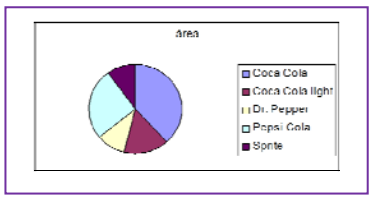
En el **diagrama de sectores** se colocan las modalidades del atributo (variable cualitativa) o valores de una variable cuantitativa discreta en un círculo, asignando a cada uno un sector del círculo de **ángulo proporcional a su frecuencia**. No resuelta muy operativo cuando la variable tiene demasiadas categorías.



Autora: Raquel Caro
Revisoras: María Molero y Nieves Zuasti
Ilustraciones: Banco de Imágenes de INTEF

Ejemplo:

- De la misma manera podemos recoger la información obtenida de emisiones de gases de efecto invernadero en España en el periodo 1999-2012 (%)



Actividades resueltas

- Dada la información correspondiente a las preferencias de 50 adolescentes americanos respecto a la marca de refresco que consumen de la actividad resuelta del apartado 3.1. realizar el gráfico de sectores.

Actividades propuestas

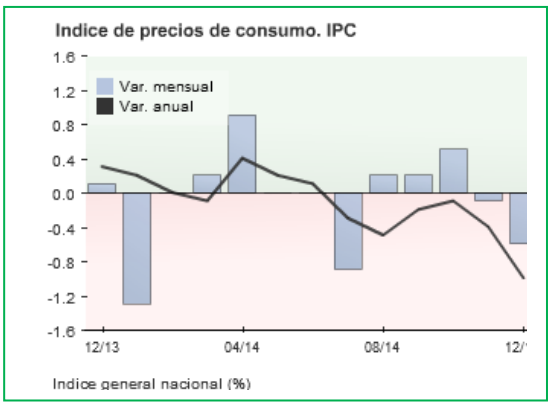
14. De los 100 asistentes a una boda, el 34 % comió ternera de segundo plato, 25 % pato, 24 % cordero y el resto pescado.
    a) Organiza la información anterior en una tabla de frecuencias y representa los datos en un gráfico de sectores.
    b) Realiza un diagrama de barras y explica cómo lo haces. ¿Cuál de los dos gráficos prefieres? ¿Por qué?
15. Se ha recogido información sobre el contenido de sales minerales de 24 botellas de agua de un grupo de escolares en una excursión tal que:

45	45	65	56	33	65	23	23
34	23	43	67	22	43	34	23
12	34	45	34	19	34	23	43

    a) Clasifica la variable estadística estudiada
    b) ¿Sería conveniente tomar o no intervalos al hacer una tabla de frecuencias?
    c) Realiza el gráfico que consideres más oportuno.

4.3. Análisis crítico de tablas y gráficas estadísticas en los medios de comunicación. Detección de falacias

Los medios de comunicación recurren con frecuencia a tablas y gráficas que ayuden a una más fácil interpretación de los datos por parte del público en general. Un caso puede ser el siguiente grafico que presenta el Instituto Nacional de Estadística (INE), que representa el índice de los precios al consumo.



No obstante, no es raro observar cómo se utilizan unos mismos datos estadísticos para obtener conclusiones distintas.

- Una subida de precios o del índice de paro puede parecer más o menos acentuada según quién presente la información
- Un índice de audiencia o el colesterol de un determinado alimento pueden parecer más o menos altos según con qué se lo compare
- Las llamadas telefónicas parecen ser más baratas en una compañía que en otra.

La lista de ejemplos es interminable.

De este modo, la Estadística, además del papel instrumental que hemos presentado hasta ahora, tiene un importante papel en el desarrollo del pensamiento crítico que nos mantendrá atentos a estos excesos.

Los errores más frecuentes, aunque a veces no se trata de errores, sino de manipulaciones tendenciosas, son los siguientes:

- Errores en la obtención de datos
- Limitaciones humanas o de los instrumentos: es imposible, por ejemplo, medir el peso o la estatura de una persona con infinita precisión. Pero incluso en estudios exhaustivos, como los censos, se estiman los errores de muestreo.
- Cuestionarios mal planteados: si no se recogen todas las posibles respuestas, si la pregunta influye en la respuesta, si las preguntas contienen juicios de valor o si las diferentes opciones de respuesta no son equilibradas (por ejemplo: sí, a veces, no). El conjunto de respuestas posibles puede hacer que haya duplicaciones u omisiones. Incurrir en este error, deliberadamente o no, deja a individuos de la población sin representación entre las respuestas y, por lo tanto, los resultados que salgan del estudio estarán sesgados. Las modalidades de la variable deben ser incompatibles y exhaustivas (por ejemplo: si preguntamos por el color favorito y ofrecemos como posibles respuestas "Rojo", "Azul" o "Amarillo", dejamos sin poder responder a quienes quieren escoger otro color; si no estamos interesados en otros colores, podemos incluir un apartado llamado "Otro").
- Delimitación imprecisa de la población: Por ejemplo, si se desea estudiar si los niños madrileños ven demasiado la televisión, habrá que dejar claro qué edades en concreto se considerarán, si entendemos por madrileño a cualquier residente o sólo a los nacidos en Madrid, etc.
- Selección de la muestra inapropiada o no representativa: la muestra no representa a la población. La elección de los individuos concretos que forman parte de la muestra debe hacerse de forma aleatoria. Por ejemplo: si estudiamos los gustos televisivos de los adolescentes de un instituto y pensamos que estos gustos pueden variar en función de la edad, en la selección de la muestra deben escogerse edades variadas, a poder ser, en la misma proporción en la

que se presentan en el instituto.

- Errores en las tablas: los datos no están ordenados, evitar ambigüedades en los extremos de los intervalos para variables continuas, etc.
- Errores en las gráficas: en los diagramas de barras falta el origen, están truncados o en la escala en los ejes, etc. Hay que dejar claras las variables que se miden.
- Errores en los parámetros de medida: por ejemplo la media no es representativa (poblaciones heterogéneas) o se ve afectada por valores muy grandes; confusión entre media y mediana.
- Errores en los pictogramas con superficies donde se inscriben proporcionales al cuadrado de las frecuencias.

5. MEDIDAS DE TENDENCIA CENTRAL

5.1. Medidas de tamaño

Las medidas de tendencia central o de centralización son las que, intuitivamente, aparecen en primer lugar al intentar describir una población o muestra.

Se pueden dividir en tres clases: **medidas de tamaño, de frecuencia y de posición**.

En lo que sigue, supondremos que estamos analizando una población de la que se toma una muestra de tamaño N, es decir, que está compuesta por N individuos (u observaciones), de los cuales se desea estudiar la variable X, lo que da lugar a la obtención de N valores que se representan por $x_1, x_2, \ldots, x_N$. Estos valores no se suponen ordenados, sino que el subíndice indica el orden en el que han sido seleccionados.

Las medidas de tamaño se definen a partir de los valores de la muestra, así como de su frecuencia.

Definimos así la **media aritmética** o **promedio** o, simplemente, **media** como:

$$\overline{x} = \frac{\sum_{i=1}^{N} [x_i]}{N}$$

Se puede interpretar como el centro de masas de las observaciones de la muestra. Dentro de sus ventajas se pueden destacar que utiliza todas las observaciones, que son fácilmente calculables, tienen una interpretación sencilla y buenas propiedades matemáticas. Su inconveniente es que se puede ver afectada por los valores anormalmente pequeños o grandes que existan en la población o muestra (denominados *outliers*).

En el caso que tengamos una variable cuantitativa agrupada en intervalos el valor de la variable X que representa al intervalo para poder calcular la media aritmética es la **marca de clase** y se calcula como la semisuma de los valores extremos del intervalo.

Ejemplo:

- Se recoge la información referida al número de horas de vuelo diarias de 20 azafatas. Si la media es igual a 4,1, esto indica que, por término medio, el número de horas de vuelo es 4,1.

Ejemplo:

- De la misma manera si recogemos la información sobre la edad media de tu clase obtendremos un valor entre 15 y 16 años. La edad media será por ejemplo 15,4, valor teórico, que puede no coincidir con alguno de los valores reales.

Actividades resueltas

- Un fabricante de helados está realizando un control de calidad sobre ciertas máquinas respecto a su capacidad de regular la temperatura de refrigeración. Para ello, selecciona una muestra de N = 16 máquinas de la fábrica y mide con precisión el valor de su capacidad (en la unidad de medida μF), obteniendo los siguientes resultados: 20.5, 19.8, 19.6, 19.2, 23.5, 28.9, 19.9, 19.2, 20.1, 18.8, 19.5, 20.2, 18.6, 19.7, 22.1, 19.3. Utilizando estos valores de capacidad, obtener la media aritmética.

$$\overline{x} = \frac{\sum_{i=1}^{N} [x_i]}{N} = \frac{20.5 + 19.8 + 19.6 + 19.2 + 23.5 + 28.9 + 19.9 + 19.2 + 20.1 + 18.8 + 19.5 + 20.2 + 18.6 + 19.7 + 22.1 + 19.3}{16} = \mathbf{20.56} \quad \mu\mathbf{F}$$

Actividades propuestas

16. Una persona ingresa 10.000 euros en un fondo de inversión el 1 de enero de 2009. Las rentabilidades anuales del fondo durante los años siguientes fueron las siguientes:

Año	2009	2010	2011	2012
Rentabilidades (%)	5	3	–1	4

Si no ha retirado el capital, ¿cuál ha sido la rentabilidad media de dicho fondo durante estos años?

Autora: Raquel Caro
Revisoras: María Molero y Nieves Zuasti
Ilustraciones: Banco de Imágenes de INTEF

17. Interpreta los valores de la variable de esta tabla que representa el peso de 100.000 bombonas de butano de una fábrica, en kilogramos. ¿Qué grafico utilizarías? Calcula la media e interprétala.

Peso [)	fi %	n_i	N_i
14,5-15	0,3	300	300
15-15,5	1,6	1600	1900
15,5-16	7,4	7400	9300
16-16,5	21,5	21500	30800
16,5-17	30,5	30500	61300
17-17,5	24,5	24500	85800
17,5-18	10,7	10700	96500
18-18,5	21,5	21500	30800

5.1. Medidas de frecuencia

Se definen teniendo en cuenta únicamente la frecuencia de los valores de la variable de la muestra.

La moda (*Mo*) se define como el valor de la variable que se ha obtenido con mayor frecuencia. Puede haber más de una moda.

Ejemplo:

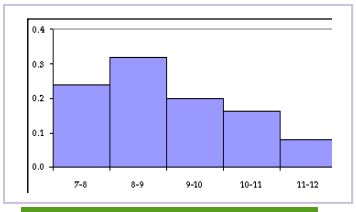
- Se realiza un estudio entre 200 espectadores a un musical en Madrid para determinar el grado de satisfacción, obteniéndose los siguientes resultados:

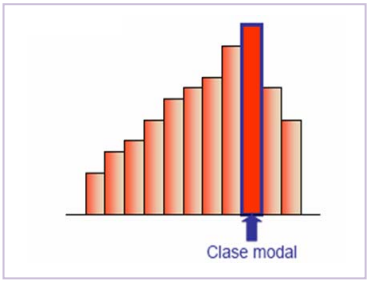
Opinión	Muy bueno	Bueno	Regular	Malo	Muy malo
%	75	25	45	15	40

La modalidad que más se repite es "muy bueno", por lo que la moda es Mo = Muy bueno.

Ejemplo:

- En el caso que la distribución esté agrupada en intervalos habrá que identificar la clase modal, es decir, el intervalo donde hay mayor número de valores de la variable.





Actividades resueltas

- A partir de la tabla de frecuencias del espesor de latas de refresco, podemos dibujar sus histogramas de frecuencias relativas y determinar dónde está su moda. Es decir en el intervalo [8-9). La moda señala que lo más frecuente es tener un espesor entre 8 y 9 mm.

Actividades propuestas

18. Obtener la media y la moda de los siguientes valores de la variable referidos al resultado de lanzar un dado 50 veces.

1	2	3	2	3	4	3	3	3	5
5	5	5	6	5	6	5	6	4	4
3	2	1	2	3	4	5	6	5	4
3	2	3	4	5	6	5	4	3	2
3	4	5	5	5	5	6	6	6	3

19. Realizar la actividad anterior pero agrupando en intervalos de amplitud 2, empezando en 0. ¿Obtienes los mismos resultados? ¿Por qué?

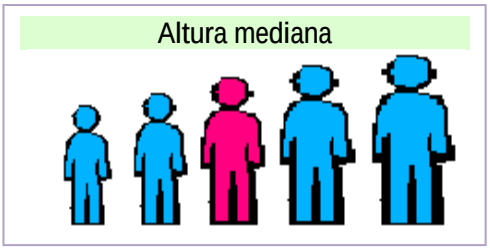
5.3. Medidas de posición

Se definen a partir de la posición de los valores de la muestra.

En general, se conocen con el nombre de centiles o percentiles.

Si reordenamos en orden creciente los valores tomados de la muestra y los denotamos por $x_{\{1\}}$, $x_{\{2\}}$,…, $x_{\{N\}}$ se pueden definir las siguientes medidas de posición:

- La *mediana Me* es un valor tal que el 50 % de las observaciones son inferiores a él. No tiene por qué ser único y puede ser un valor no observado.
- Los *cuartiles* (o cuartilas) Q_1, Q_2 y Q_3 son los valores tales que el 25 %, 50 % y 75 % (respectivamente) de los valores de la variable son inferiores a él.
- Los *deciles* D_1, D_2,…, D_9 son los valores tales que el 10 %, 20 %, …, 90 % (respectivamente) de los valores de la variable son inferiores a él.



Autora: Raquel Caro
Revisoras: María Molero y Nieves Zuasti
Ilustraciones: Banco de Imágenes de INTEF

En general, se define el **percentil** o **centil** del k % (siendo $0 \leq k \leq 100$) como el valor tal que el k % de las observaciones son inferiores a él.

La mediana y el resto de medidas de posición tienen como principal ventaja su fácil interpretación y su robustez (no se ven afectadas por observaciones extremas).

Ejemplo:

- Calcula los cuartiles y el percentil 65 de los siguientes valores de la variable referidos al número de hijos de las familias de un bloque de edificios de la localidad de Madrid:

Número de hijos	f_i	F_i
1	11	11
2	27	38
3	4	42
4	18	60
Total	60	

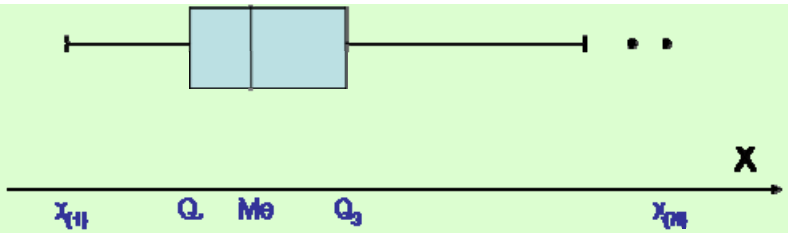
Para hallar el primer cuartil calculamos el 25 % del total muestral N = 60, es decir, 60·0,25 = 15. Así, el primer cuartil tiene 15 valores de la variable menores y el resto mayores. En la columna de frecuencias acumuladas, el primer número mayor o igual que 15 es 38, que corresponde al valor de la variable 2. Por tanto el primer cuartil es 2. De la misma forma el 50 % de 60 es 30, es decir el cuartil 2 (Mediana) sería también 2. El 75 % de 60 sería 45 y de esta forma el cuartil 3 sería 4 puesto que el valor mayor a 45 es 60, que corresponde al valor 4 de la variable objeto de estudio. Por último, el percentil 65 corresponde al valor 3 ya que 65 % de 60 es igual a 39 y el valor mayor que 39 es 42.

Resumen:
25 % de 60 = 15 → 38 > 15 > 11 → $Q_1 = 2$
50 % de 60 = 30 → 38 > 30 > 11 → $Me = Q_2 = 2$
75 % de 60 = 45 → 60 > 45 > 42 → $Q_3 = 4$
65 % de 60 = 39 → 42 > 39 > 38 → $P_{65} = 3$

Las medidas de posición nos permiten realizar otro tipo de gráfico estadístico que se llama el *gráfico de caja*.

Para realizar este gráfico, se construye una *caja* (ya sea horizontal o vertical), cuyos lados coinciden con el *primer y tercer cuartil* Q_1 y Q_3. Por lo tanto, la caja abarca el 50 % de las observaciones realizadas. Dentro de dicha caja, se incluye un segmento (o bien un punto) que corresponde a la *mediana*.

De cada lado de la caja parte un segmento que se extiende hasta los valores correspondientes a las observaciones *mínima* y *máxima* $x_{\{1\}}$ y $x_{\{N\}}$.



Actividades resueltas

- Se está realizando un control de calidad sobre los fallos de unas determinadas máquinas. Para ello, se contabilizan los fallos de $N = 13$ máquinas durante un mes, obteniendo los siguientes números de fallos: 2, 5, 3, 2, 0, 4, 1, 7, 4, 2, 1, 0, 2. Utilizando estos valores obtener las medidas de tendencia central y resumir en una tabla de frecuencias la información obtenida del número de fallos mensuales de las máquinas, obteniendo la media aritmética de otra manera.

$$\bar{x} = \frac{\sum_{i=1}^{N}[x_i]}{N} = \frac{2+5+3+2+0+4+1+7+4+2+1+0+2}{13} = \mathbf{2.54} \quad \textbf{fallos/mes}$$

$$Mo = \mathbf{2} \quad \textbf{fallos/mes}$$

$$Q_1 = x_{\{4\}} = \mathbf{1} \quad \textbf{fallo/mes}$$

$$Q_3 = x_{\{10\}} = \mathbf{4} \quad \textbf{fallos/mes}$$

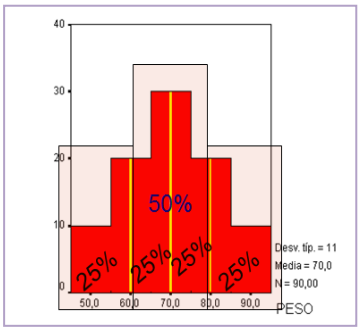
Valores	0	1	2	3	4	5	6	7
Frecuencia absoluta	2	2	4	1	2	1	0	1
Frecuencia relativa	0.154	0.154	0.307	0.077	0.154	0.077	0	0.077
Frecuencia relativa acumulada	0.154	0.308	0.615	0.692	0.846	0.923	0.923	1

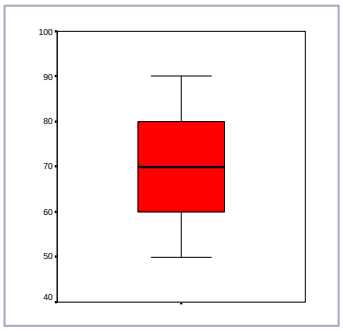
$$\bar{x} = \sum_{i=1}^{k}[f_i \cdot x_i] = 0.154\cdot 0 + 0.154\cdot 1 + 0.307\cdot 2 + 0.077\cdot 3 + 0.154\cdot 4 + 0.077\cdot 5 + 0.077\cdot 7 = \mathbf{2.54} \quad \textbf{fallos/mes}$$

Autora: Raquel Caro
Revisoras: María Molero y Nieves Zuasti
Ilustraciones: Banco de Imágenes de INTEF

- Se recoge información sobre el peso de 90 chicos en una clase de Matemáticas. Determinar los centiles que nos permiten realizar el gráfico de caja.

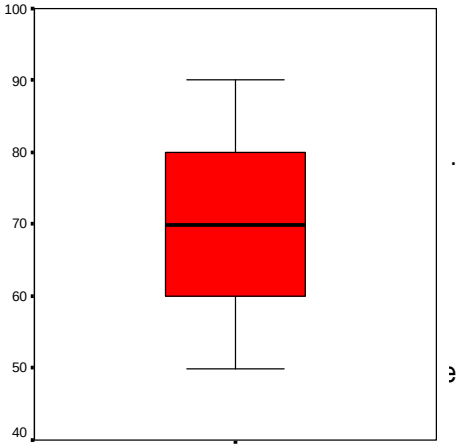
Primer cuartil = percentil 25 = 60 Kg.
Tercer cuartil= percentil 75= 80 kg.





Actividades propuestas

20. Dibujar un diagrama de caja conociendo los siguientes datos.
    Mínimo valor = 2; cuartil 1 = 3; mediana = 6; cuartil 3 = 7; máximo valor = 12
21. Un corredor de maratón entrena, de lunes a viernes recorriendo las siguiente
    Si el sábado también entrena:
    a) ¿Cuántos kilómetros debe recorrer para que la media sea la misma?
    b) ¿Y para que la mediana no varíe?
    c) ¿Y para que la moda no varíe?



22. EL salario mensual en euros de los 6 trabajadores de una empresa textil es e
    medidas de tendencia central describe mejor los sueldos de la empresa?

1700	1400	1700	1155	1340

23. ¿Qué valor o valores podríamos añadir a este conjunto de valores de la variable para que la mediana siga siendo la misma?

12	19	24	23	23	15	21	32	12	6	32	12	12	21

24. Salen 25 plazas para un puesto de auxiliar de enfermería y se presentan 200 personas con las siguientes notas.

notas	3	4	5	6	7	8	9	10
n_i	6	34	25	56	29	10	30	10

    a) ¿Con qué nota se obtiene una de las plazas mediante el examen?
    b) ¿Qué percentil es la nota 5?

6. MEDIDAS DE DISPERSIÓN

6.1. Medidas de desviaciones

Las medidas de tendencia central resultan insuficientes a la hora de describir una muestra. Además de las tendencias, es necesario disponer de medidas sobre la variabilidad de los datos. Dentro de estas medidas, vamos a estudiar las medidas de desviaciones y los rangos.

Las medidas de desviaciones recogen las desviaciones de los valores de la variable respecto de una medida de tendencia central.

La *varianza* se define como:

$$s^2 = \frac{\sum_{i=1}^{N}\left[\left(x_i - \overline{x}\right)^2\right]}{N} = \frac{\sum_{i=1}^{N} x_i^{\,2}}{N} - \overline{x}^2$$

Sus principales ventajas son su manejabilidad matemática y que utiliza todas las observaciones. Sus principales inconvenientes son que es muy sensible a observaciones extremas y que su unidad es el cuadrado de la unidad original de la muestra.

La desviación típica es la raíz cuadrada de la varianza y tiene la principal ventaja de que utiliza las mismas unidades que los valores de la variable originales.

Observa que la desviación típica es una distancia, la distancia de los valores de la variable a la media. Recuerda que la raíz cuadrada es siempre un número positivo.

Autora: Raquel Caro
Revisoras: María Molero y Nieves Zuasti
Ilustraciones: Banco de Imágenes de INTEF

Asociado a la media y la desviación típica, se define el **coeficiente de variación**, definido en muestras con media distinta de cero como:

$$g = \frac{s}{|\bar{x}|}$$

Este coeficiente es adimensional (no tiene unidades y se suele expresar en porcentaje), lo que resulta una gran ventaja, ya que permite comparar la variabilidad de distintas muestras, independientemente de sus unidades de medida. Algunos autores definen este coeficiente utilizando la media en el denominador, en lugar de su valor absoluto. Valores del coeficiente de variación mayores del 100 % indican que la media no se puede considerar representativa del conjunto de valores de la variable.

Ejemplo:

- La nota media de 6 alumnos de una misma clase de 4º ESO en Matemáticas es de 5. Si la varianza es 0,4, la desviación típica será de 0,632, por tanto la media es bastante homogénea en la distribución. Las notas que se han obtenido están situadas alrededor de la nota media 5.

Actividades resueltas

- El propietario de una instalación mixta solar-eólica está realizando un estudio del volumen de energía que es capaz de producir la instalación. Para ello, mide dicha energía a lo largo de un total de *N* = 16 días que considera suficientemente representativos. La energía (en kilovatio, KWh) producida en dichos días por dos instalaciones se encuentra recogida en la siguiente tabla:

Generación solar	13’1	10’5	4’1	14’8	19’5	11’9	18	8’6
Generación eólica	8’5	14’3	24’7	4	2’3	6’4	3’6	9’2
Generación solar	5’7	15’9	11’2	6’8	14’2	8’2	2’6	9’7
Generación eólica	13’5	1’4	7’6	12’8	10’3	16’5	21’4	10’9

Utilizando estos valores de la variable calcula las medidas de dispersión estudiadas, comparando los resultados en las dos instalaciones

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{13'1 + 10'5 + 4'1 + 14'8 + 19'5 + 11'9 + 18 + 8'6 + 5'7 + 15'9 + 11'2 + 6'8 + 14'2 + 8'2 + 2'6 + 9'7}{16} = 10'925 \text{ Kwh}$$

$$\bar{y} = \frac{\sum_{i=1}^{N} y_i}{N} = \frac{8'5 + 14'3 + 24'7 + 4 + 2'3 + 6'4 + 3'6 + 9'2 + 13'5 + 1'4 + 7'6 + 12'8 + 10'3 + 16'5 + 21'4 + 10'9}{16} = 10'463 \text{ Kwh}$$

$$s_x^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - \bar{x}^2 = \frac{13'1^2 + 10'5^2 + 4'1^2 + 14'8^2 + 19'5^2 + 11'9^2 + 18^2 + 8'6^2 + 5'7^2 + 15'9^2 + 11'2^2 + 6'8^2 + 14'2^2 + 8'2^2 + 2'6^2 + 9'7^2}{16} = \frac{141'5}{16} - 10'9^2 = 22'16$$

$$s_y^2 = \frac{\sum_{i=1}^{N} y_i^2}{N} - \bar{y}^2 = \frac{8'5^2 + 14'3^2 + 24'7^2 + 4^2 + 2'3^2 + 6'4^2 + 3'6^2 + 9'2^2 + 13'5^2 + 1'4^2 + 7'6^2 + 12'8^2 + 10'3^2 + 16'5^2 + 21'4^2 + 10'9^2}{16} = \frac{150'48}{16} - 10'5^2 = 41'01$$

$$g_x = \frac{s_x}{|\bar{x}|} = \frac{\sqrt{22'16}}{10'9} = \frac{4'7}{10'9} = 0'43 \qquad g_y = \frac{s_y}{|\bar{y}|} = \frac{\sqrt{41'01}}{10'5} = \frac{6'4}{10'5} = 0'61$$

La media de la primera instalación es más representativa que la media de la segunda puesto que el coeficiente de variación es menor en la primera. Los datos están menos agrupados en la segunda de las instalaciones. Su desviación típica es mucho mayor.

- Se está realizando un control de calidad sobre los fallos de unas determinadas máquinas. Para ello, se contabilizan los fallos de *N* = 13 máquinas durante un mes, obteniendo los siguientes números de fallos. Utilizando estos valores presentados en la tabla de frecuencias obtener las medidas de dispersión estudiadas.

Valores	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Frecuencia absoluta	2	2	4	1	2	1	0	1
Frecuencia relativa	0’154	0’154	0’307	0’077	0’154	0’077	0	0’077
Frecuencia relativa acumulada	0’154	0’308	0’615	0’692	0’846	0’923	0’923	1

$$\bar{X} = \frac{\sum_{i=1}^{N} [x_i]}{N} = \frac{2 + 5 + 3 + 2 + 0 + 4 + 1 + 7 + 4 + 2 + 1 + 0 + 2}{13} = \mathbf{2.54} \quad \textbf{fallos/mes}$$

Autora: Raquel Caro
Revisoras: María Molero y Nieves Zuasti
Ilustraciones: Banco de Imágenes de INTEF

$$S^2 = \sum_{i=1}^{k}\left[f_i \cdot (x_i - \bar{x})^2 \right] = 0.154 \cdot (-2.54)^2 + 0.154 \cdot (-1.54)^2 + 0.307 \cdot (-0.54)^2 + 0.077 \cdot 0.46^2 + 0.154 \cdot 1.46^2 +$$

$$+ 0.077 \cdot 2.46^2 + 0.077 \cdot 4.46^2 = \mathbf{3.80} \quad (\mathbf{fallos/mes})^2$$

Otra forma de realizar estos mismos cálculos es:

									Suma
Valores	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	
Frecuencia absoluta	2	2	4	1	2	1	0	1	13
x_i^2	0	1	4	9	16	25	36	49	
x_i^2·**Fr. Abs.**	0	2	16	9	32	25	0	49	133

Aplicamos la fórmula: $s^2 = \dfrac{\sum_{i=1}^{N} x_i^{\,2}}{N} - \bar{x}^2$ y obtenemos que

$s^2 = 133/13 - 2'54^2 = 10'23 - 6'45 = 3'80$, por lo que $s = 1'95$.

Actividades propuestas

25. Un grupo de perros pastor alemán tiene una media de 70 kg y desviación típica 2 kg. Un conjunto de perros caniche tiene una media de 15 kg y desviación típica 2 kg. Compara ambos grupos.
26. El tiempo, en minutos, que un conjunto de estudiantes de 4º ESO dedica a preparar un examen de Matemáticas es:

234	345	345	123	234	234	556
234	234	345	223	167	199	490

Las calificaciones de ese conjunto de estudiantes son las siguientes:

4	5	6	7	6	5	8
9	8	7	8	7	6	8

a) ¿Qué tendremos que hacer para comparar su variabilidad? b) ¿En qué conjunto los valores de la variable están más dispersos? c) ¿Es la media siempre mayor que la desviación típica?

6.2. Los rangos

Estas medidas proporcionan información acerca del intervalo total de valores que toma la muestra analizada.

El **rango total** o **recorrido** es la diferencia entre los valores máximos y mínimos que toma la variable en la muestra:

$$R = x_{\{N\}} - x_{\{1\}}$$

El **recorrido intercuartílico** es la diferencia entre el tercer y el primer cuartil: $R_I = Q_3 - Q_1$

Ejemplo:

- Se está realizando un control de calidad sobre los fallos de una determinada máquina. Para ello, se contabilizan los fallos de N = 13 máquinas durante un mes, obteniendo los siguientes números de fallos: 2, 5, 3, 2, 0, 4, 1, 7, 4, 2, 1, 0, 2. Utilizando estos valores obtenemos el rango total igual a 7 y el recorrido intercuartílico igual a 3.

Actividades resueltas

- Salen 25 plazas para un puesto de cajero en un supermercado y se presentan 200 personas. La siguiente información recoge las notas de un test de conocimientos básicos.

notas	2	3	4	5	6	7	8	9	10
n_i	6	4	30	25	56	29	10	30	10

Calcular el rango total de la variable objeto de estudio.

Actividades propuestas

27. Se ha recogido una muestra de 20 recipientes cuyos diámetros son:

0'91 1'04 1'01 1 0'77 0'78 1 1'3 1'02 1
1 0'88 1'26 0'92 0'98 0'78 0'82 1'2 1'16 1'14

a) Calcula todas las medidas de dispersión que conozcas.
b) ¿A partir de qué valor de diámetro de los recipientes se consideran el 20 % con mayor diámetro?

Autora: Raquel Caro
Revisoras: María Molero y Nieves Zuasti
Ilustraciones: Banco de Imágenes de INTEF

7. CONSTRUCCIÓN E INTERPRETACIÓN DE DIAGRAMAS DE DISPERSIÓN. INTRODUCCIÓN A LA CORRELACIÓN

Este apartado se centra en el análisis de datos bidimensional, en el que son dos las variables de interés. De este modo, cuando se está analizando una población y se selecciona una muestra, para cada individuo se toman dos valores, correspondientes a dos características (o variables) distintas. En este sentido, puede ser interesante considerar simultáneamente los dos caracteres a fin de estudiar las posibles relaciones entre ellos.

7.1. Tablas de frecuencia de una variable bidimensional

Cuando se quieren resumir los resultados de una muestra bidimensional utilizando una tabla de frecuencias (ya sea por tratarse de una variable discreta, o porque se deseen agrupar las observaciones de una variable continua), es preciso utilizar lo que se denomina *tabla de doble entrada* (o bidimensional). Sean $x_1, x_2,\ldots, x_k$ las modalidades de la primera variable e $y_1, y_2,\ldots, y_p$ las de la segunda. Estas modalidades pueden corresponder tanto a los valores que se dan en la muestra (si la variable es discreta), como a las marcas de clase de los intervalos utilizados (si la variable es continua). Para construir la tabla de frecuencias, se utilizan las frecuencias absolutas n_{ij} correspondientes a las observaciones que toman simultáneamente valores correspondientes a las clases x_i e y_j. Obviamente, se ha de verificar que:

$$\sum_{i=1}^{k}\sum_{j=1}^{p}\left[n_{ij}\right]=N$$

Con esto, la tabla de frecuencias absolutas se presenta como:

	y_1	y_2	……	y_p	$n_{i\cdot}$
x_1	n_{11}	n_{12}	……	n_{1p}	$n_{1\cdot}$
x_2	n_{21}	n_{22}	……	n_{2p}	$n_{2\cdot}$
……	……	……	……	……	……
x_k	n_{k1}	n_{k2}	……	n_{kp}	$n_{k\cdot}$
$n_{\cdot j}$	$n_{\cdot 1}$	$n_{\cdot 2}$	……	$n_{\cdot p}$	N

Los valores $n_{i\cdot}$ recogen las frecuencias absolutas de la clase x_i, mientras que n_j es la suma de frecuencias absolutas de la clase y_j, con lo que se verifica:

$$n_{i\cdot}=\sum_{j=1}^{p}\left[n_{ij}\right] \qquad n_{\cdot j}=\sum_{i=1}^{k}\left[n_{ij}\right]$$

$$\sum_{i=1}^{k}\left[n_{i\cdot}\right]=N \qquad \sum_{j=1}^{p}\left[n_{\cdot j}\right]=N$$

De la misma manera, se puede realizar una tabla de frecuencias relativas f_{ij}, utilizando los cocientes entre las frecuencias absolutas y el número de observaciones:

$$f_{ij}=\frac{n_{ij}}{N}\leq 1$$

Actividades resueltas

- El propietario de una instalación mixta solar-eólica está realizando un estudio del volumen de energía que es capaz de producir la instalación. Para ello, mide dicha energía a lo largo de un total de $N = 16$ días que considera suficientemente representativos. La energía (en *kWh*) producida en dichos días por las instalaciones solar y eólica se pueden resumir en las siguientes tablas de doble entrada de frecuencias absolutas y de frecuencias relativas:

		Energía eólica				
		[0, 6'5]	(6'5, 13]	(13, 19'5]	(19'5, 26]	$n_{i\cdot}$
Energía solar	[0, 5]	0	0	0	2	2
	(5, 10]	0	3	2	0	5
	(10, 15]	2	3	1	0	6
	(15, 20]	3	0	0	0	3
	$n_{\cdot j}$	5	6	3	2	16

Autora: Raquel Caro
Revisoras: María Molero y Nieves Zuasti
Ilustraciones: Banco de Imágenes de INTEF

		Energía eólica				
		[0, 6'5]	(6'5, 13]	(13, 19'5]	(19'5, 26]	$f_{i\cdot}$
Energía solar	[0,5]	0	0	0	0'125	0'125
	(5, 10]	0	0'1875	0'125	0	0'3125
	(10, 15]	0'125	0'1875	0'0625	0	0'375
	(15, 20]	0'1875	0	0	0	0'1875
	$f_{\cdot j}$	0'3125	0'375	0'1875	0'125	1

7.2. Representación gráfica de una variable bidimensional

Al igual que en el caso de una muestra unidimensional, en numerosas ocasiones resulta interesante realizar una representación gráfica de una muestra bidimensional.

Un modo sencillo de representar una muestra bidimensional es mediante el denominado **diagrama de dispersión** o **nube de puntos**. Esta técnica consiste en representar en el plano (x, y) los valores obtenidos en la muestra.

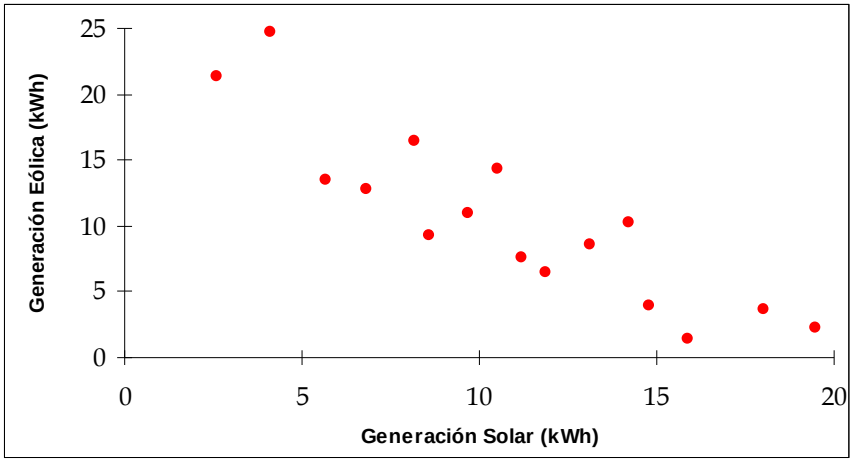


Diagrama de dispersión de la generación solar y eólica (en kWh) de la actividad resuelta

La figura anterior muestra el diagrama de dispersión. Se puede observar la existencia de una dependencia inversa.

7.3. Medidas en una variable bidimensional. Coeficiente de correlación

Cuando se está analizando una muestra bidimensional, se pueden calcular las medidas que caracterizan a cada una de las variables de la muestra por separado, tal y como se ha descrito anteriormente. Pero en este caso se puede dar un paso más y calcular algunas medidas conjuntas, que tienen en cuenta simultáneamente los valores que toman ambas variables en cada individuo.

Al igual que cuando se analiza una única característica, supondremos que se toma una muestra de tamaño N de la población, es decir, que está compuesta por N individuos (u observaciones), de los cuales se desea analizar las características (o variables) X e Y. Esto da lugar a la obtención de N valores para cada una de las dos variables: (x_1, y_1), (x_2, y_2),…, (x_N, y_N). De nuevo, estos valores no se suponen ordenados, sino que el subíndice indica el orden en el que han sido seleccionados.

Siguiendo esta notación se pueden formular los cálculos de los momentos respecto al origen y respecto a la media para una variable bidimensional. Definimos, por tanto:

Momentos respecto al origen de orden (r, s) como: $$a_{r,s} = \frac{\sum_{i=1}^{N}\left[x_i^r \cdot y_i^s\right]}{N}$$

Observa que los momentos respecto al origen de orden (1, 0) y (0, 1) coinciden con las medias de ambas variables:

$$a_{1,0} = \overline{x} \qquad a_{0,1} = \overline{y}$$

También resulta de interés al momento de orden (1, 1): $$a_{1,1} = \frac{\sum_{i=1}^{N}\left[x_i \cdot y_i\right]}{N}$$

Análogamente, se pueden definir los momentos respecto a la media de orden (r, s): $$m_{r,s} = \frac{\sum_{i=1}^{N}\left[\left(x_i - \overline{x}\right)^r \cdot \left(y_i - \overline{y}\right)^s\right]}{N}$$

Los momentos respecto a la media de orden (2, 0) y (0, 2) coinciden con las varianzas de ambas variables:

$$m_{2,0} = {s_X}^2 \qquad m_{0,2} = {s_Y}^2$$

Autora: Raquel Caro
Revisoras: María Molero y Nieves Zuasti
Ilustraciones: Banco de Imágenes de INTEF

El momento respecto a la media de orden (1, 1), que se denomina **covarianza** o momento mixto, es de gran importancia:

$$m_{1,1} = \frac{\sum_{i=1}^{N}\left[(x_i - \bar{x})\cdot(y_i - \bar{y})\right]}{N}$$

Alternativamente a la fórmula anterior, la covarianza se puede calcular a partir de los momentos respecto al origen, según la fórmula:

$$m_{1,1} = a_{1,1} - a_{1,0}\cdot a_{0,1} = \frac{\sum_{i=1}^{N}(x_i \cdot y_i)}{N} - \bar{x}\cdot\bar{y}$$

La covarianza, al igual que la varianza, tiene el inconveniente de que depende de las unidades de la muestra.

Por este motivo, se utiliza el coeficiente de **correlación** lineal de Pearson (que se denota, indistintamente, como ρ o r):

$$\rho = r = \frac{m_{1,1}}{s_x \cdot s_y} = \frac{\frac{\sum_{i=1}^{N}(x_i \cdot y_i)}{N} - \bar{x}\cdot\bar{y}}{s_x \cdot s_y}$$

Este coeficiente tendrá el signo de la covarianza y nos indicará si la dependencia entre las dos variables objeto de estudio son dependientes positiva o negativamente. El coeficiente de correlación (o simplemente correlación) toma un valor comprendido entre –1 y 1. Si la correlación es positiva se dice que existe dependencia directa entre *X* e *Y* (a un aumento de una de las dos variables le corresponde una tendencia al aumento en la otra). En cambio, si la correlación es negativa, se dice que existe una dependencia inversa (a un aumento de una de las dos variables le corresponde una tendencia al decremento en la otra).

Actividades resueltas

- El propietario de una instalación mixta solar-eólica está realizando un estudio del volumen de energía que es capaz de producir la instalación. Para ello, mide dicha energía a lo largo de un total de *N* = 16 días que considera suficientemente representativos. La energía (en kWh) producida en dichos días por las instalaciones solar y eólica se encuentra recogida en la siguiente tabla:

Generación solar (x_i)	13'1	10'5	4'1	14'8	19'5	11'9	18	8'6	5'7	15'9	11'2	6'8	14'2	8'2	2'6	9'7
Generación eólica (y_i)	8'5	14'3	24'7	4	2'3	6'4	3'6	9'2	13'5	1'4	7'6	12'8	10'3	16'5	21'4	10'9

Utilizando estas producciones, vamos a calcular la covarianza y el coeficiente de correlación, denotando a la generación solar como variable *X* y la generación eólica como variable *Y*. Añadimos nuevas filas a nuestra tabla:

Generación solar (x_i)	13'1	10'5	4'1	14'8	19'5	11'9	18	8'6	5'7	15'9	11'2	6'8	14'2	8'2	2'6	9'7
Generación eólica (y_i)	8'5	14'3	24'7	4	2'3	6'4	3'6	9'2	13'5	1'4	7'6	12'8	10'3	16'5	21'4	10'9
x_i^2	171'6	110'3	16'81	219'0	380'3	141'6	324	73'96	32'49	252'8	125'4	46'24	201'6	67'24	6'76	94'09
y_i^2	72'25	204'5	610'1	16	5'29	40'96	12'96	84'64	182'3	1'96	57'76	163'8	106'1	272'3	457'9	118'8
$x_i * y_i$	111'4	150'2	101'3	59'2	44'85	76'16	64'8	79'12	76'95	22'26	85'12	87'04	146'2	135'3	55'64	105'7

Previamente calculamos la media y la desviación típica de cada variable (que ya conocemos de una actividad resuelta anterior). Sumando la primera fila y dividiendo por *N* = 16, obtenemos la media de la Generación Solar en Kwh. Recuerda $\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$; por tanto

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{13'1+10'5+4'1+14'8+19'5+11'9+18+8'6+5'7+15'9+11'2+6'8+14'2+8'2+2'6+9'7}{16} = 10'925 \text{ Kwh}$$

Sumando la segunda fila y dividiendo por *N* = 16 obtenemos la media de la Generación Eólica en Kwh:

$$\bar{y} = \frac{\sum_{i=1}^{N} y_i}{N} = \frac{8'5+14'3+24'7+4+2'3+6'4+3'6+9'2+13'5+1'4+7'6+12'8+10'3+16'5+21'4+10'9}{16} = 10'463 \text{ Kwh}$$

En la tercera fila hemos calculado los cuadrados de los valores de la primera variable y los utilizamos para calcular la varianza: Recuerda $s_x^{\,2} = \frac{\sum_{i=1}^{N} x_i^{\,2}}{N} - \bar{x}^2$; por tanto

Autora: Raquel Caro
Revisoras: María Molero y Nieves Zuasti
Ilustraciones: Banco de Imágenes de INTEF

$$s_x^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - \bar{x}^2 = \frac{13'1^2 + 10'5^2 + 4'1^2 + 14'8^2 + 19'5^2 + 11'9^2 + 18^2 + 8'6^2 + 5'7^2 + 15'9^2 + 11'2^2 + 6'8^2 + 14'2^2 + 8'2^2 + 2'6^2 + 9'7^2}{16} = \frac{1415}{16} - 10'9^2 = 22'16$$

En la cuarta fila los cuadrados de los valores de la segunda variable y calculamos su varianza:

$$s_y^2 = \frac{\sum_{i=1}^{N} y_i^2}{N} - \bar{y}^2 = \frac{8'5^2 + 14'3^2 + 24'7^2 + 4^2 + 2'3^2 + 6'4^2 + 3'6^2 + 9'2^2 + 13'5^2 + 1'4^2 + 7'6^2 + 12'8^2 + 10'3^2 + 16'5^2 + 21'4^2 + 10'9^2}{16} = \frac{1500'48}{16} - 10'5^2 = 41'01$$

La desviación típica es la raíz cuadrada de la varianza, por tanto: $s_x = \sqrt{22'16} = 4'71$ y $s_y = \sqrt{41'01} = 6'4$

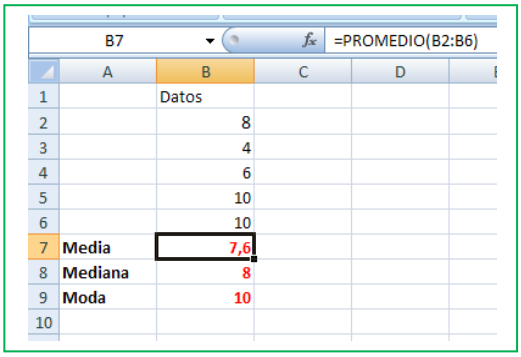
Para calcular el coeficiente de correlación calculamos en la quinta fila los productos de la variable *x* por la variable *y*. Así, 13'1*8'5 = 111'4. Queremos calcula el término: $\frac{\sum_{i=1}^{N}(x_i \cdot y_i)}{N}$. Al sumar obtenemos 1401'2, que dividimos entre 16, le restamos el producto de las medias y dividimos por el producto de las desviaciones típicas:

$$\rho = \frac{\frac{\sum_{i=1}^{N}(x_i \cdot y_i)}{N} - \bar{x} \cdot \bar{y}}{s_x \cdot s_y} = \frac{1401'2 - (10'9 \cdot 10'5)}{4'71 \cdot 6'4} = \frac{-26'728}{4'71 \cdot 6'4} = -0'887$$

Este coeficiente de correlación negativo y cercano a –1 nos indica que la relación entre las dos variables es negativa y bastante importante.

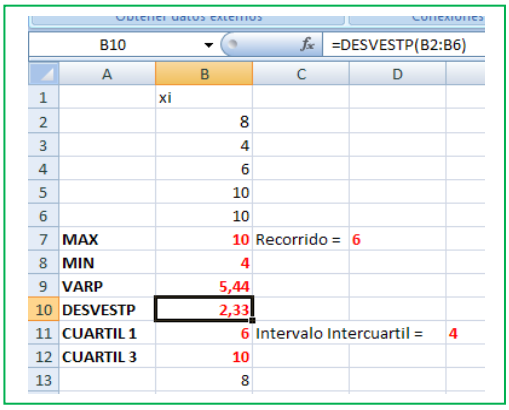
Utiliza el ordenador

- *Nieves ha tenido en Matemáticas las siguientes notas: 8, 4, 6, 10 y 10. Calcula su media, su moda y su mediana.*



Para calcular la media, la mediana y la moda con la hoja de cálculo, copiamos en la casilla B2, B3… los datos: 8, 4, 6, 10 y 10. Escribimos en la casilla A7, Media, y para calcular la media escribimos un signo igual en B7. Buscamos, desplegando las posibles funciones, la función PROMEDIO, y escribimos =PROMEDIO(B2:B6), que significa que calcule la media de los valores que hay en las casillas desde B2 hasta B6.

Del mismo modo calculamos la mediana buscando en las funciones o escribiendo =MEDIANA(B2:B6) y la moda buscando en las funciones o escribiendo =MODA(B2,B6).



Igual que hemos calculado la media, la mediana y la moda, la hoja de cálculo se puede utilizar para obtener:

- El recorrido calculando MAX – MIN → 6.
- La varianza utilizando VARP → 5'44.
- La desviación típica usando DESVESTP →.2'33
- Los cuartiles, (CUARTIL), siendo el cuartil 0 el mínimo; el cuartil 1, Q1; el cuartil 2, la mediana; el cuartil 3, Q3; y el cuartil 4, el máximo.
- Q1 = 6. Q3 = 10. Intervalo intercuartil = 10 – 6 = 4.

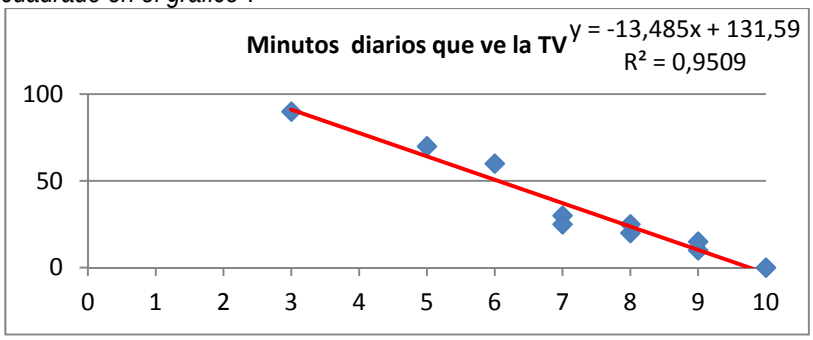
Utiliza el ordenador

- Preguntamos a 10 alumnos de 4º ESO por sus calificaciones en Matemáticas, por el número de minutos diarios que ven la televisión, por el número de horas semanales que dedican al estudio, y por su estatura en centímetros. Los datos se recogen en la tabla adjunta. Queremos dibujar las nubes de puntos que los relacionan con las calificaciones de Matemáticas, el coeficiente de correlación y la recta de regresión.

Calificaciones de Matemáticas	10	3	7	8	5	9	9	8	6	7
Minutos diarios que ve la TV	0	90	30	20	70	10	15	25	60	25
Horas semanales de estudio	15	2	9	12	7	14	13	11	7	8
Estatura (en cm)	177	168	157	159	163	179	180	175	169	170

Para hacerlo, entramos en Excel, y copiamos los datos. Seleccionamos la primera y la segunda fila, luego la primera y la tercera y por último la primera fila y la cuarta.

Con la primera y segunda filas seleccionadas, vamos a *Insertar*, *Dispersión* y elegimos la *nube de puntos*. Podemos conseguir que el eje de abscisas vaya de 0 a 10 en "*Dar formato al eje*". Pinchamos sobre un punto de la nube, y elegimos "*Agregar línea de tendencia*". Para que dibuje el ordenador la recta de regresión la línea de tendencia debe ser *Lineal*. En la pantalla que aparece marcamos la casilla que dice: "*Presentar ecuación en el gráfico*" y la casilla que dice "*Presentar el valor de R cuadrado en el gráfico*".



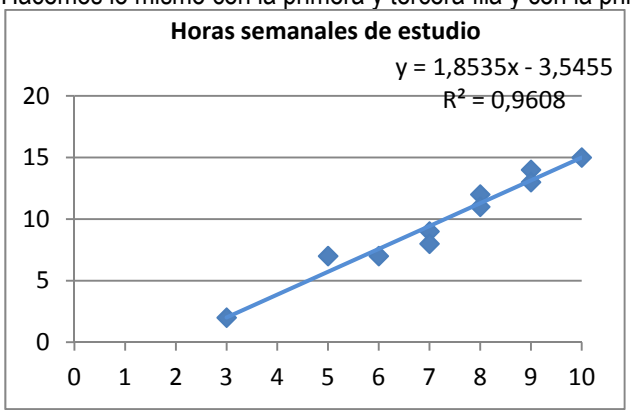
Observa, la recta de regresión, en color rojo, es decreciente y su ecuación es aproximadamente:

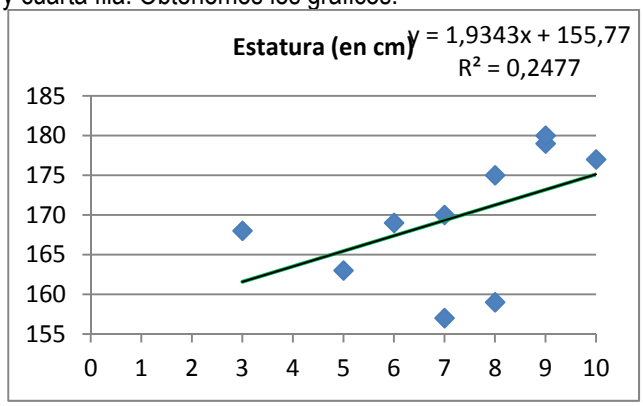
$$y = -13'5\,x + 132.$$

El cuadrado del coeficiente de correlación es $\rho^2 = 0'95$. La correlación es negativa y alta:

$$\rho = \sqrt{0'95} = -0{,}975$$

Hacemos lo mismo con la primera y tercera fila y con la primera y cuarta fila. Obtenemos los gráficos:





Observa que en ambos casos la pendiente de la recta de regresión es positiva pero en el primero el coeficiente de correlación, positivo, es próximo a 1, $\rho = \sqrt{0'96} = 0{,}98$. La correlación es alta y positiva.

En el segundo $\rho = \sqrt{0'25} = 0{,}5$.

Actividades propuestas

28. Se han medido los pesos y alturas de 6 personas, como muestra de las personas que están en una fila o cola de espera, obteniéndose los siguientes resultados:

Pesos (kg)	65	60	65	63	68	68
Alturas (cm)	170	150	168	170	175	180

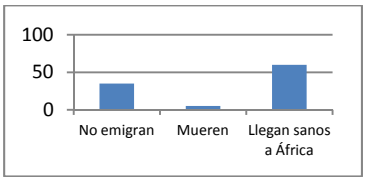
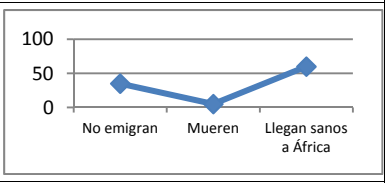
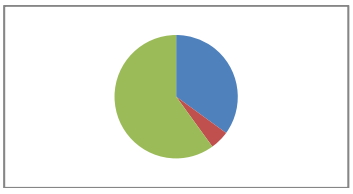
Se pide:

a) Calcular las medias y las varianzas de esos dos conjuntos de datos unidimensionales.
b) ¿Qué medidas están más dispersas, los pesos o las alturas?
c) Representar gráficamente ese conjunto de datos bidimensional. Calcular la covarianza e interpretar su valor.
d) Dar una medida de la correlación entre ambas variables. Interpretar su valor.

RESUMEN

		Ejemplos
Población estadística, colectivo o universo	El conjunto de todos los individuos (personas, objetos, animales, etc.) que contengan información sobre el fenómeno que se estudia.	Número de personas en España entre 16-65 años
Muestra	Es un subconjunto representativo que se selecciona de la población y sobre el que se va a realizar el análisis descriptivo. El tamaño de la muestra es el número de sus elementos. Cuando la muestra comprende a todos los elementos de la población, se denomina censo.	Número de personas en un barrio de Madrid entre 16-65 años.

Autora: Raquel Caro
Revisoras: María Molero y Nieves Zuasti
Ilustraciones: Banco de Imágenes de INTEF

Variable observable o estadística X	En general, supondremos que se está analizando una determinada población, de la que nos interesa cierta característica que viene dada por la variable *X*.	Las variables que están bajo estudio se pueden clasificar en dos categorías: Variables cualitativas o atributos (datos no métricos) Variables cuantitativas, que tienen un valor numérico.
Frecuencia absoluta	Número de veces que se repite un valor de la variable	Si al tirar un dado hemos obtenido 2 veces el 3, 2 es la frecuencia absoluta de 3.
Frecuencia relativa	Frecuencia absoluta dividido por el número de experimentos	Si se realiza un experimento 500 veces y la frecuencia absoluta de un suceso es 107, la frecuencia relativa es 107/500.
Frecuencia acumulada	Se suman las frecuencias anteriores	
Diagrama de rectángulos o barras	Los valores de la variable se representan mediante rectángulos de igual base y de altura proporcional a la frecuencia. Se indica en el eje horizontal la variable y en el vertical las frecuencias.	
Polígono de frecuencias	De unen los puntos medios superiores de un una diagrama de barras	
Diagrama de sectores	En un círculo se dibujan sectores de ángulos proporcionales a las frecuencias	
Media aritmética	Es el cociente entre la suma de todos los valores de la variable y el número total de datos.	En los datos 3, 5, 5, 7, 8, la media es: (3 + 5 + 5 + 7 + 8)/5 = 28/5 = 5,6.
Mediana	Deja por debajo la mitad de los valores y por encima la otra mitad	La mediana es 5
Moda	El valor que más se repite.	La moda es 5.
Varianza	Medida de desviación que recoge las desviaciones de los valores de la variable respecto de la media aritmética.	$s^2 = \frac{\sum_{i=1}^{N}\left[\left(x_i - \overline{x}\right)^2\right]}{N}$
Desviación típica	La desviación típica es la raíz cuadrada de la varianza	
Coeficiente de variación	Permite comparar la variabilidad de distintas muestras, independientemente de sus unidades de medida.	$g = \frac{s}{\lvert\overline{x}\rvert}$
Rango total o recorrido	Diferencia entre los valores máximos y mínimos que toma la variable en la muestra	$R = x_{\{N\}} - x_{\{1\}}$
Recorrido intercuartílico	Diferencia entre el tercer y el primer cuartil	$R_I = Q_3 - Q_1$

Autora: Raquel Caro
Revisoras: María Molero y Nieves Zuasti
Ilustraciones: Banco de Imágenes de INTEF

EJERCICIOS Y PROBLEMAS.

Población y muestra. Variables estadísticas. Tablas de frecuencias

1. Se lanza una moneda 700 veces y se obtiene cara 355 veces. Expresa en una tabla las frecuencias absolutas, relativas y calcula también las frecuencias acumuladas absolutas y acumuladas relativas de caras y cruces en este experimento.
2. Se lanzar un dado 500 veces y se obtienen los siguientes resultados:

Resultado	1	2	3	4	5	6
Número de veces	70	81	92	85		81

a) ¿Cuántas veces ha salido el 5?
b) Construir tabla con las frecuencias absolutas y las frecuencias absolutas acumuladas
c) Construir una tabla con las frecuencias relativas y las frecuencias relativas acumuladas

3. Una urna que contiene 10 bolas numeradas del 0 al 9, sacamos una bola, anotamos el número y devolvemos la bola a la urna. Repetimos el experimento 1000 veces y se han obtenido los resultados indicados en la tabla:

Resultado	0	1	2	3	4	5	6	7	8	9
Frecuencia absoluta	79	102			93	98	104	77		
Frecuencia relativa			0,12	0,13					0,1	
Frecuencia absoluta acumulada	79	181								
Frecuencia relativa acumulada										1

a) ¿Cuál es la frecuencia absoluta de 9?
b) ¿Cuál es la frecuencia absoluta acumulada de 2?
c) ¿Cuál es la frecuencia relativa acumulada de 1?
d) Copia la tabla en tu cuaderno y complétala.

4. Pepa ha tirado un dado 25 veces y ha obtenido los siguientes resultados:

1, 2, 5, 6, 3, 1, 4, 5, 6, 1, 3, 1, 2, 2, 1, 6, 2, 2, 4, 3, 4, 6, 6, 1, 4

a) Construir una tabla de frecuencias absolutas.
b) Construir una tabla de frecuencias relativas.
c) Dibuja un diagrama de barras.
d) Dibuja un polígono de frecuencias y una representación por sectores.

5. En una clase se ha medido el tamaño de las manos de cada uno de los alumnos y alumnas, y el resultado en centímetros ha sido el siguiente:

19, 18, 20, 19, 18, 21, 19, 17, 16, 20,
16, 19, 20, 21, 18, 17, 20, 19, 22, 21,
23, 21, 17, 18, 17, 19, 21, 20, 16, 19

a) ¿Qué tamaño ha sido el valor mínimo? ¿Y el máximo? ¿Cuál es el rango total de la variable?
b) Construir una tabla de frecuencias absolutas y otra de frecuencias relativas.
c) Construir una tabla de frecuencias absolutas acumuladas y otra de frecuencias relativas acumuladas.

6. Calcula la frecuencia absoluta de los datos de una encuesta en la que se ha elegido entre ver la televisión, t, o leer un libro, l:

t, l, t, t, t, l, t, t, l, t, l, t, l, t, t, t, l, l, t, l, t, l, t, l, t.

7. La duración en minutos de unas llamadas telefónicas ha sido:

7, 3, 6, 3, 7, 5, 4, 3, 5, 7, 10, 1, 9, 12, 2

Construir una tabla de frecuencias absolutas y una tabla de frecuencias relativas.

Gráficos estadísticos

8. Se ha preguntado en un pueblo de la provincia de Madrid el número de hermanos que tenían y se ha obtenido la siguiente tabla de frecuencias absolutas sobre el número de hijos de cada familia:

Número de hijos	1	2	3	4	5	6	7	8 o más
Número de familias	46	249	205	106	46	21	15	6

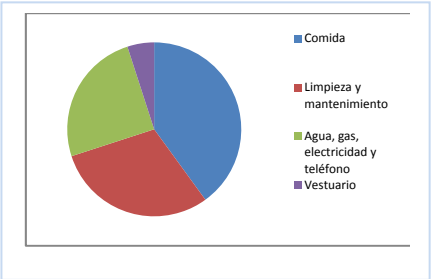
a) Escribe en tu cuaderno una tabla de frecuencias relativas.
b) Haz un diagrama de barras de frecuencias absolutas y otro de frecuencias relativas.
c) Haz un polígono de frecuencias absolutas y otro de frecuencias absolutas acumuladas.

9. Haz una encuesta similar con tus compañeros y compañeras de curso preguntando el número de hermanos y confeccionando una tabla sobre el número de hijos y el número de familias.

a) Construye una tabla de frecuencias relativas
b) Haz un diagrama de barras de frecuencias absolutas y relativas. Completa con un polígono de frecuencias
c) Compara la tabla de frecuencias relativas y el diagrama de barras de frecuencias relativas que obtengas con el obtenido en el ejercicio anterior.

Autora: Raquel Caro
Revisoras: María Molero y Nieves Zuasti
Ilustraciones: Banco de Imágenes de INTEF

10. Un batido de frutas contiene 25 % de naranja, 15 % de plátano; 50 % de manzana y, el resto de leche. Representa en un diagrama de sectores la composición del batido.
11. En un campamento de verano se han gastado diez mil euros. El gráfico muestra la distribución del gasto:
    1. Comida: 40 %
    2. Limpieza y mantenimiento: 30 %
    3. Agua, gas, electricidad y teléfono: 25 %
    4. Vestuario: .....



a) ¿Qué porcentaje se gastó en vestuario?
b) ¿Cuántos euros se gastaron en comida?
c) ¿Cuánto mide el ángulo del sector correspondiente a actividades?

12. Busca en revistas o periódicos dos gráficas estadísticas, recórtalas y pégalas en tu cuaderno. En muchas ocasiones estas gráficas tienen errores. Obsérvalas detenidamente y comenta las siguientes cuestiones:

a) ¿Está clara la variable a la que se refiere? ¿Y las frecuencias?
b) ¿Son correctas las unidades? ¿Pueden mejorarse?
c) Comenta las gráficas.

13. Se hace una encuesta sobre el número de veces que van al cine unos jóvenes al mes. Los valores de la variable están en la tabla:

Veces que van al cine	0	1	2	3	4	5
Frecuencia absoluta	1	7	9	5	2	1

a) Representa un diagrama de barras de frecuencias absolutas.
b) Representa un polígono de frecuencias relativas.
c) Representa los valores de la variable en un diagrama de sectores.

14. Se hace un estudio sobre lo que se recicla en una ciudad y se hace una tabla con el peso en porcentaje de los distintos tipos de residuos:

Tipo de residuo	Porcentaje
Orgánico	15
Papel y cartón	1
Vidrio	15
Plástico	1
Pilas	15

a) Construye un diagrama de barras
b) Representa un polígono de frecuencias.
c) Representa los valores de la variable en un diagrama de sectores.

15. En un ejercicio anterior se ha tenido el resultado de medir en una clase el tamaño de las manos de cada uno de los alumnos y alumnas, y el resultado en centímetros ha sido el siguiente:

19, 18, 20, 19, 18, 21, 19, 17, 16, 20,
16, 19, 20, 21, 18, 17, 20, 19, 22, 21,
23, 21, 17, 18, 17, 19, 21, 20, 16, 19

Representa los valores de la variable en un diagrama de barras y en un polígono de frecuencias.

16. El 35 % de las cigüeñas no ha emigrado este año a África y el 6 % murió por el camino. Dibuja un diagrama por sector que describa esta situación.
17. En una clase se ha preguntado por las preferencias deportivas y se ha obtenido:

Futbol	Baloncesto	Natación	Kárate	Ciclismo
8	9	7	6	10

a) Copia la tabla en tu cuaderno y construye una tabla de frecuencias relativas.
b) Representa estos valores de la variable en un diagrama de sectores.

Medidas de centralización y dispersión

18. Pepa ha tirado un dado 25 veces de un ejercicio anterior y ha obtenido los siguientes resultados:

1, 2, 5, 6, 3, 1, 4, 5, 6, 1, 3, 1, 2, 2, 1, 6, 2, 2, 4, 3, 4, 6, 6, 1, 4

a) Calcula la media aritmética
b) Calcula la mediana
c) ¿Cuál es la moda? ¿Es única?
d) Calcula la varianza y desviación típica interpretando su resultado

Autora: Raquel Caro
Revisoras: María Molero y Nieves Zuasti
Ilustraciones: Banco de Imágenes de INTEF

19. Sara ha tenido las siguientes notas en sus exámenes de Matemáticas: 9, 7, 8, 6, 9, 10, 9

a) Calcula la media aritmética
b) Calcula la mediana
c) ¿Cuál es la moda? ¿Es única?
d) Calcula el percentil 45 interpretando su resultado
e) Calcula el percentil 75 interpretando su resultado. ¿qué otro nombre recibe?
f) Calcula la varianza y desviación típica interpretando su resultado
g) Calcula el coeficiente de variación interpretando su resultado

20. En un ejercicio anterior se ha tenido el resultado de medir en una clase el tamaño de las manos de cada uno de los alumnos y alumnas, y el resultado en centímetros ha sido el siguiente:

19, 18, 20, 19, 18, 21, 19, 17, 16, 20, 16, 19, 20, 21, 18, 17, 20, 19, 22, 21, 23, 21, 17, 18, 17, 19, 21, 20, 16, 19

a) Calcula la media aritmética
b) Calcula la mediana
c) ¿Cuál es la moda? ¿Es única?
d) Calcula el percentil 45 interpretando su resultado
e) Calcula el percentil 75 interpretando su resultado. ¿qué otro nombre recibe?
f) Calcula la varianza y desviación típica interpretando su resultado
g) Calcula el coeficiente de variación interpretando su resultado

21. Nos interesa conocer la distribución de notas obtenidas por 40 estudiantes. Las notas son:

4, 1, 7, 10, 3, 2, 8, 9, 0, 0, 5, 8, 2, 7, 1, 2, 8, 10, 2, 10, 3, 4, 8, 9, 3, 6, 3, 7, 2, 4, 9, 4, 9, 5, 1, 3, 3, 9, 7, 8, 10

a) Escribe en tu cuaderno una tabla de frecuencias absolutas.
b) Haz un polígono de frecuencias absolutas.
c) Calcula la media
d) Calcula la mediana
e) Calcula la moda
f) Calcula el percentil 45 interpretando su resultado
g) Calcula el percentil 75 interpretando su resultado. ¿qué otro nombre recibe?
h) Calcula la varianza y desviación típica interpretando su resultado
i) Calcula el coeficiente de variación interpretando su resultado
j) Si las notas de los mismos alumnos respecto a otra asignatura tienen una media de 5,3 y desviación típica de 2, ¿cuál de las dos asignaturas tiene una media más homogénea?

22. Los jugadores de un equipo de balonmano tiene las siguientes edades:

12, 14, 13, 12, 15, 11, 12, 12, 13, 14, 11, 12, 12.

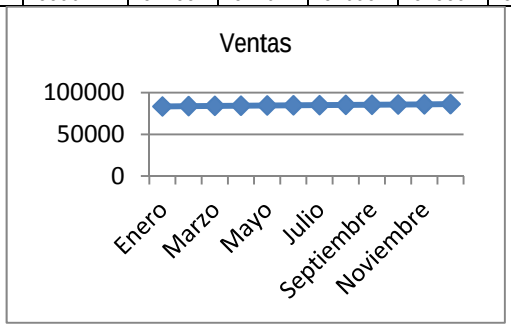
a. Calcula la media
b. Calcula la mediana
c. Calcula la moda
d. Calcula el percentil 45 interpretando su resultado
e. Calcula el percentil 75 interpretando su resultado. ¿qué otro nombre recibe?
f. Calcula la varianza y desviación típica interpretando su resultado
g. Calcula el coeficiente de variación interpretando su resultado

Problemas

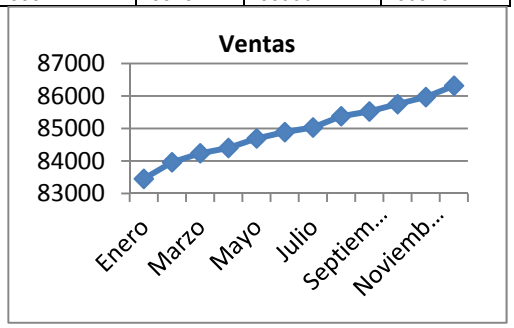
23. El Director Comercial de una empresa va a ser evaluado. Para ello debe dar cuenta de los resultados obtenidos. Quiere quedar bien, pues eso le puede suponer un aumento de sueldo. Se han vendido las siguientes cantidades:

Meses	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre
Ventas	83451	83962	84238	84401	84693	84889	85032	85378	85524	85751	859967	86316

El estadístico de la empresa le ha entregado la siguiente gráfica:



No le ha gustado nada, y para la presentación él se ha confeccionado el siguiente gráfico:



Autora: Raquel Caro
Revisoras: María Molero y Nieves Zuasti
Ilustraciones: Banco de Imágenes de INTEF

Ambos gráficos son correctos. Escribe un informa sobre cómo pueden los distintos gráficos dar impresiones tan diferentes.

24. Tira una moneda 15 veces y anota las veces que cae cara y las que no. Construye luego dos tablas: una de frecuencias absolutas y otra de frecuencias relativas. Representa el resultado en un diagrama de frecuencias y en un polígono de frecuencias.
25. La media de seis números es 5. Se añaden dos números más pero la media sigue siendo 5. ¿Cuánto suman estos dos números?
26. La siguiente tabla expresa las estaturas, en metros, de 1000 soldados:

Talla	1,50 – 1,56	1,56 – 1,62	1,62 – 1,68	1,68 - 1,74	1,74 - 1,80	1,80-1,92
Nº de soldados	20	150	200	330	200	100

Calcula:

a) La media y la desviación típica. b) Los intervalos donde se encuentran la mediana y los cuartiles.

c) El intervalo $(\overline{X} - \sigma, \overline{X} + \sigma)$ y el porcentaje de individuos en dicho intervalo. d) Representa los datos en un histograma.

29. Una compañía aérea sospecha que existe una relación entre las variables *X*, tiempo de un vuelo, en horas; e *Y*, consumo de combustible (gasóleo) para dicho vuelo, en litros. Por esta razón, se han obtenido los siguientes datos, dentro del rango de niveles de interés para X en esta compañía.

X_i	0´4	0´5	0´6	0´65	0´7	0´8	1	1´15	1´2	1´4	1´5	1´6	1´8	2´2	3
Y_i	1.350	2.220	2.900	3.150	3.350	3.550	3.900	4.330	4.500	5.050	5.320	5.650	6.400	7.500	10.250

Se pide:

a) Mediante la representación del diagrama de dispersión razonar el interés de relacionar dichas variables.

b) Obtener la covarianza y el coeficiente de correlación entre ambas variables. Interpretar los resultados.

AUTOEVALUACIÓN

1. Un diagrama de caja informa sobre:
   a) Los cuartiles y curtosis. b) Asimetría y varianza. c) Datos atípicos y simetría.
2. Sea la variable aleatoria número de personas que es capaz de levantar un ascensor. Para calcular el nº de personas a partir del cual se recoge el 30 % de los valores de la variable necesitamos obtener
   a) El percentil 30 b) El percentil 3 c) El percentil 70
3. El 25 % de los madrileños gastan en la factura del móvil por encima de 100 euros, mientras que el 25 % gastan por debajo de 20 euros. Entonces conocemos:
   a) 100 y 20 son valores que corresponden al cuartil 1 y 3, respectivamente.
   b) 100 y 20 son valores que corresponden al cuartil 3 y 1, respectivamente.
   c) 100 y 20 son valores que no corresponden a ningún cuartil.
4. En un diagrama de barras de frecuencias absolutas, la suma de sus alturas es proporcional a:
   a) 100 b) 1 c) Total de valores de la variable d) Suma de sus bases
5. La media de los siguientes valores de la variable 3, 4, 6, 7, 5, 8, es: a) 6 b) 7 c) 4,8 d) 5,5
6. La mediana de los siguientes valores de la variable 3, 4, 6, 7, 8, es: a) 6 b) 7 c) 4 d) 5
7. La moda de los siguientes valores de la variable 3, 4, 6, 7, 5, 8, 7, 7, es: a) 6 b) 7 c) 4 d) 5
8. La media de 7 números es 8. Se añaden dos números más pero la media sigue siendo 8. ¿Cuánto suman estos dos números? a) 10 b) 16 c) 20 d) 14
9. Dos revistas especializadas en empleo, A y B, han publicado una media de ofertas de trabajo, de $m_A = 10$ y $m_B = 20$ con varianzas, respectivamente de $s^2_A = 4$ y $s^2_B = 9$.

a) La revista B presenta mayor coeficiente de variación que la revista A.

b) La revista A presenta mayor coeficiente de variación que la revista B.

c) La revista B presenta igual coeficiente de variación que la A

10. El 70 % de los madrileños gastan en regalos navideños por encima de 100 euros, mientras que el 5 % gastan por encima de 500 euros. Entonces conocemos:

A) El valor correspondiente al percentil 30. B) El valor correspondiente al percentil 70. C) Al percentil 5.

Autora: Raquel Caro
Revisoras: María Molero y Nieves Zuasti
Ilustraciones: Banco de Imágenes de INTEF