Formación Profesional Básica

Matemáticas II

Capítulo 7: Estadística

LibrosMareaVerde.tk

www.apuntesmareaverde.org.es

Este capítulo ha sido realizado por **David Miranda Suárez** para el alumnado que cursa Matemáticas II de Formación Profesional Básica en el Centro Salesianos Loyola - Naranjoven, en Fuenlabrada (Madrid) en los perfiles de Electricidad y Electrónica, y en el perfil de Peluquería y Estética, basándose en el currículo de la **Comunidad de Madrid** (BOCM Real Decreto 127/2014, de 28 de febrero del BOE).

El autor ha utilizado los textos de Matemáticas de **Marea Verde**. Para la elaboración de este capítulo se han utilizado partes del siguiente capítulo de los textos elaborados por el equipo de Matemáticas de **Marea Verde** **(www.apuntesmareaverde.org.es):**

Capítulo 11: **Estadística y probabilidad** de **3ª ESO A** de autor: **Fernando Blasco** y revisor: David Hierro

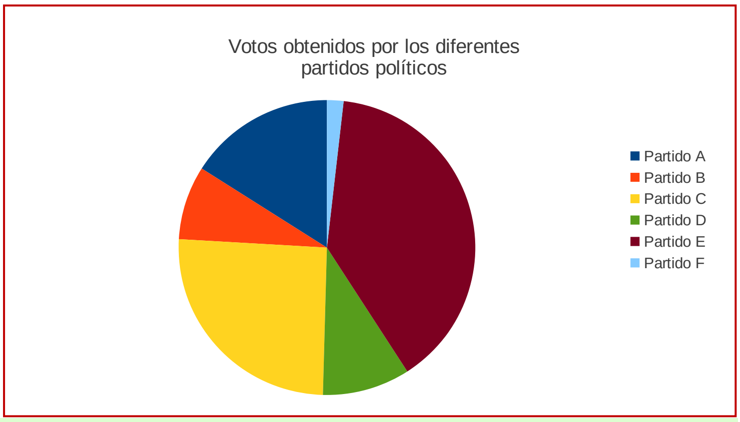
Capítulo 6: **Estadística** de Bachillerato: Matemáticas aplicadas a las **CCSS I** de autor: **Ignasi Clausell** y revisora: Raquel Caro

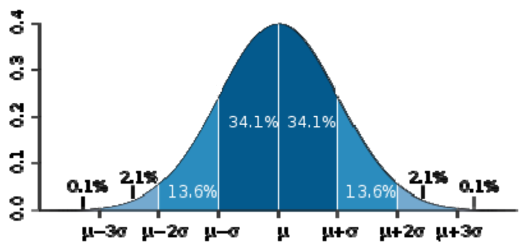
ÍNDICE

Resumen

La Estadística es una Ciencia que surgió para llevar la contabilidad del Estado. De ahí viene su nombre. En el siglo XX se desarrollaron sus técnicas y se separó de las Matemáticas, pasando a ser una ciencia con entidad propia.

En los medios de comunicación encontramos frecuentes estadísticas. En medicina se necesitan métodos estadísticos para probar nuevos medicamentos. En todo experimento científico, tras la recogida de datos, se necesita utilizar pruebas estadísticas que permitan sacar información de esos datos.





Media y desviación típica. Imagen de wikipedia

Vamos a estudiar conceptos de estadística unidimensional, como las tablas de frecuencias y los gráficos estadísticos, calcular las medidas de centralización, media, mediana y moda y las medidas de dispersión, varianza y desviación típica.

Aunque el nombre de Estadística proviene del s. XIX, sin embargo ya se utilizaban representaciones gráficas y otras medidas en pieles, rocas, palos de madera y paredes de cuevas para controlar el número de personas, animales o ciertas mercancías desde la Prehistoria. Los babilonios usaban ya envases de arcilla para recopilar datos sobre la producción agrícola. Los egipcios analizaban los datos de la población y la renta del país mucho antes de construir las pirámides. Los antiguos griegos realizaban censos cuya información se utilizaba hacia 600 a C.

1. LA TOMA DE DATOS

1.1. Un ejemplo para realizar un análisis

Ejemplo:

La Casa de la Moneda quiere estudiar cuántas monedas debe emitir, teniendo en cuenta las que están en circulación y las que se quedan atesoradas (bien en casas particulares, o en máquinas de refrescos, o depositadas en un banco). Se ha hecho una encuesta a pie de calle a 60 personas y se ha apuntado cuántas monedas llevaba cada una de ellas en el bolsillo. Hemos obtenido estos datos:

12	7	11	8	8	9	6	12	7	7	13	0	10	9	13	18	7	6	11	12	16	0	10	10	8	8	9	11	10	8
16	8	5	2	12	8	14	14	16	6	2	0	18	10	10	12	14	6	7	3	12	11	10	18	9	7	12	1	15	8

El primer paso consiste en hacer un esquema para el recuento: usaremos una tabla y marcaremos palotes cada vez que aparezca ese número.

0	///
1	/
2	//
3	/
4	
5	/
6	////

7	///// /
8	///// ///
9	////
10	///// //
11	////
12	///// //
13	//

14	///
15	/
16	///
17	
18	///
19	
20	

Pasar de ese recuento a una tabla de **frecuencias absolutas** es muy sencillo: solo hay que sustituir los palotes por el número que representan.

0	3

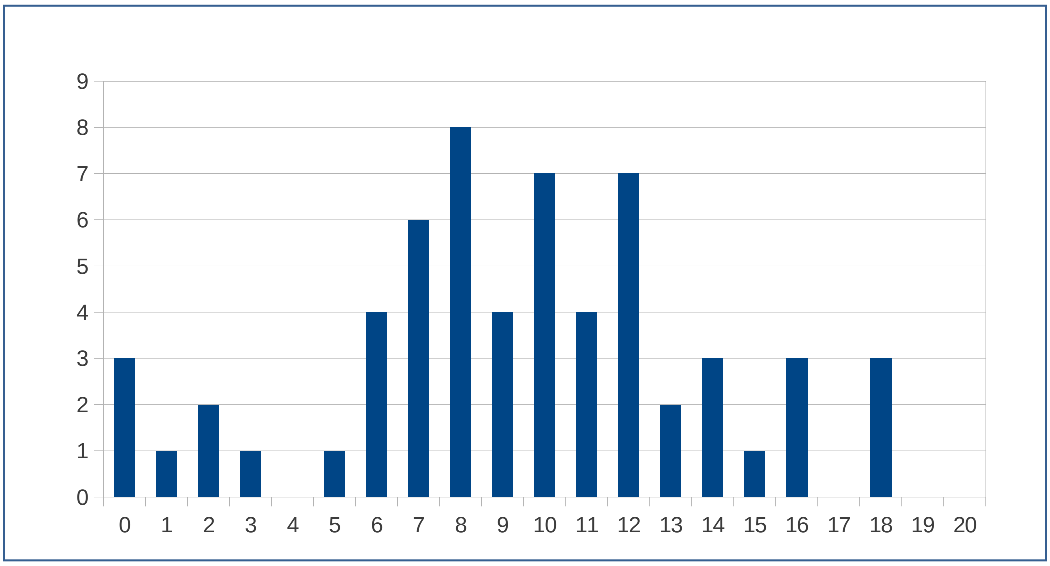
7	6

14	3

Autores: Fernando Blasco, Ignasi Clausell y David Miranda

1	1	**8**	8	**15**	1
2	2	**9**	4	**16**	3
3	1	**10**	7	**17**	0
4	0	**11**	4	**18**	3
5	1	**12**	7	**19**	0
6	4	**13**	2	**20**	0

Es mucho mejor analizar los datos de modo visual. Estamos más acostumbrados a trabajar de esa manera. Podemos representar los datos de la tabla de frecuencias en un **diagrama de barras**, donde la altura de cada barra representa la frecuencia de aparición.



El procesamiento de datos estadísticos se utiliza mucho. Obviamente no se hacen las operaciones a mano, sino que se utilizan calculadoras u hojas de cálculo. Disponer de esos medios tecnológicos será un buen complemento para el capítulo, aunque recordamos que lo más importante es comprender qué se hace en cada momento.

Comenzaremos introduciendo algo de **nomenclatura**. Casi todos estos nombres los has escuchado puesto que los medios de comunicación los utilizan muchísimo

Población es el colectivo sobre el que se quiere hacer el estudio.

Muestra es un subconjunto de la población de modo que a partir de su estudio se pueden obtener características de la población completa.

Individuo es cada uno de los elementos de la población o la muestra.

Ejemplo:

Se quiere hacer un estudio sobre hábitos alimenticios de los estudiantes de 3º de ESO de todo Madrid. Pero como es muy costoso entrevistar a todos los estudiantes se decide tomar un IES por cada distrito y entrevistar a los alumnos de 3º de ESO de esos colegios elegidos.

La **población** objeto del estudio serán todos los estudiantes madrileños matriculados en 3º de ESO.

La **muestra** son los estudiantes de 3º de ESO matriculados en los institutos elegidos.

Cada uno de los estudiantes de 3º de ESO es un **individuo** para este estudio estadístico.

Actividades propuestas

1. Queremos hacer un estudio de la cantidad de monedas que llevan en el bolsillo los estudiantes de tu clase. Pero para no preguntar a todos elige 10 compañeros al azar y anota en tu cuaderno cuántas monedas lleva cada uno.

a) ¿Cuál es la población objeto del estudio?

b) ¿Cuál es la muestra elegida?

c) Especifica 5 individuos que pertenezcan a la población y no a la muestra.

Autores: Fernando Blasco, Ignasi Clausell y David Miranda
Ilustraciones: Banco de Imágenes de INTEF

1.2. Variables estadísticas

Ejemplo:

En un estudio estadístico se puede preguntar cosas tan variopintas como

- ¿Qué frutas comes a lo largo de una semana?
- ¿Cuántas piezas de fruta comes al día?
- ¿Cuántas monedas llevas en el bolsillo?
- ¿Cuál es tu altura?
- ¿Cuántas marcas de chocolate recuerdas?
- ¿Cuáles son las marcas de chocolate que recuerdas?
- ¿Cuántos hermanos tienes?
- ¿Cuál es tu color favorito para un coche?
- ¿Cuánto tiempo pasas al día viendo la televisión?
- ¿Cuántos seguidores tienes en twitter?

Esas preguntas pueden corresponder a estudios de salud, económicos, publicitarios o socioeconómicos. Algunas se responden con un número y otras se responden con un nombre o un adjetivo. Incluso hay diferencias entre las que se responden con números: el número de monedas que llevas o el número de seguidores de *twitter* se contestan con números enteros, mientras que para hallar tu altura o las horas que pasas delante del televisor necesitamos utilizar números reales (normalmente con representación decimal).

Una variable se dice **cuantitativa** si sus valores se expresan con números.

Las variables cuantitativas pueden ser

- **discretas** si solo admiten valores aislados
- **continuas** si entre dos valores pueden darse también todos los intermedios

Una variable estadística es **cualitativa** cuando sus valores no se expresan mediante un número, sino con una cualidad.

Actividades propuestas

2. Clasifica en variables cualitativas y cuantitativas las que aparecen en el primer ejemplo de esta sección. Para las cuantitativas indica si son continuas o discretas.

1.3. Las fases de un estudio estadístico

En un estudio estadístico hay 6 fases fundamentales:

1. Determinación del objeto del estudio. Esto es, saber qué queremos estudiar.
2. Selección de las variables que se van a estudiar.
3. Recogida de los datos.
4. Organización de los datos.
5. Representación y tratamiento de los datos.
6. Interpretación y análisis.

Autores: Fernando Blasco, Ignasi Clausell y David Miranda
Ilustraciones: Banco de Imágenes de INTEF

2. ESTADÍSTICA DESCRIPTIVA UNIDIMENSIONAL

2.1. Método estadístico

La **Estadística** es la Ciencia que se encarga de la recopilación, representación y el uso de los datos sobre una o varias características de interés para, a partir de ellos, tomar decisiones o extraer conclusiones generales.

Ejemplo 1:

- *El gobierno desea averiguar si el número de hijos por familia ha descendido respecto a la década anterior. Para ello ha entrevistado a 50 familias y les ha preguntado por el número de hijos obteniendo los siguientes datos:*

2 2 2 4 1 2 4 2 3 0   2 2 2 3 2 6 2 3 2 2   3 2 3 3 4 3 3 4 5 2   1 3 2 0 3 2 1 2 3 2   2 3 1 4 2 3 2 4 3 3.

Ejemplo 2:

Un nuevo hotel va a abrir sus puertas en nuestra ciudad. Antes de decidir el precio de sus habitaciones, el gerente investiga los precios por habitación de los 40 hoteles de la misma categoría que hay cerca de nuestra ciudad. Los datos obtenidos son:

53 39 43 50 60 47 51 50 44 57 33 39 43 50 60 47 51 42 44 58 33 43 41 58 44 38 61 43 53 45 40 54 39 47 33 45 47 42 45 48.

La Estadística descriptiva es la parte de la estadística que se encarga de organizar, resumir y dar una primera descripción (sin conclusiones generales) de los datos.

En Estadística se sigue un **método estadístico** que está formado por distintas fases según se trata la información recibida.

0. Planteamiento del problema en términos precisos: ámbito de aplicación (población) y características a estudio (variables).
1. Recogida de datos de la población de interés: *Muestreo*.
2. Organización, presentación y resumen de los datos (o de la muestra): *Estadística descriptiva*.
3. Modelos matemáticos: *Teoría probabilidad*.
4. Obtener conclusiones generales o verificar hipótesis.

Población. Es el conjunto de individuos o entes sujetos a estudio.

Ejemplo 1:

Conjunto de todas las familias españolas.

Ejemplo 2:

- Todos los hoteles de esta categoría de las cercanías.

Algunas poblaciones son finitas y pueden conocerse en su totalidad, otras en cambio pueden ser infinitas y abstractas.

Muestra: Es el número de datos que tomamos de la población para realizar nuestro estudio.

Ejemplo 1:

- Las 50 familias a las que se ha preguntado por el número de hijos.

Ejemplo 2:

- Los 40 hoteles.

Tamaño muestral: Número de observaciones en la muestra.

Habitualmente se denotará por n.

Ejemplo 1:

- $n = 50$.

Ejemplo 2:

- $n = 40$.

Dato: Cada valor observado de la variable.

Ejemplo 1:

- 2 2 2 4 1 2 4 2 3 0 2 2 2 3 2 6 2 3 2 2 3 2 3 3 4 3 3 4 5 2 1 3 2 0 3 2 1 2 3 2 2 3 1 4 2 3 2 4 3 3.

Ejemplo 2:

- 53 39 43 50 60 47 51 50 44 57 33 39 43 50 60 47 51 42 44 58 33 43 41 58 44 38 61 43 53 45 40 54 39 47 33 45 47 42 45 48.

Variable: Característica que estamos midiendo.

Ejemplo 1:

Número de hijos.

Ejemplo 2:

- Precio de la habitación.

Las variables suelen denotarse por las letras mayúsculas X, Y …..

2.2. Tipos de variables

Cualitativas o categóricas: Aquellas que no son medibles, es decir aquellas cuyas observaciones no tienen carácter numérico. Expresan cualidades o categorías.

Ejemplos:

- Sexo, profesión, estado civil…

Cuantitativas: Aquellas que son medibles, es decir, sus observaciones tienen carácter numérico. Estas se dividen en:

Discretas: Toman valores numéricos fijos.

Ejemplos:

- Número de habitaciones, número de hijos de una familia, número de trabajadores de una fábrica…

Continuas: Toman valores en intervalos de números

Ejemplos:

- Peso, estatura,… cuando se organizan los datos en intervalos.

2.3. Distribuciones de frecuencias

Observando los datos de los ejemplos es fácil adivinar cuál será el primer paso. Consistirá en agrupar los datos que se repiten varias veces.

Tenemos las siguientes definiciones:

Frecuencia absoluta (n_i): Es el número de veces que se repite en la muestra un determinado valor (x_i) de la variable.

Ejemplo:

En el ejemplo 1 de número de hijos, para el dato $x_1 = 0$, $n_1 = 2$; para el dato $x_4 = 3$, $n_4 = 15$.

Propiedad:

La suma de todas las frecuencias absolutas es igual al tamaño muestral.

$$\sum n_i = n$$

Frecuencias relativas (f_i): Es igual a la frecuencia absoluta dividida por el número total de datos, es decir por el tamaño muestral.

$$f_i = \frac{n_i}{n}$$

Ejemplo:

$$f_1 = \frac{2}{50} = 0'04 \qquad f_4 = \frac{15}{50} = 0'3$$

Propiedad:

La suma de todas las frecuencias relativas es igual a 1.

Frecuencias acumuladas (N_i): Nos dice el número de datos que hay igual o inferiores a uno determinado.

Se calcula sumando el número de frecuencias absolutas que hay anteriores a llegar a la que queremos calcular.

Ejemplo:

$N_1 = 2$ $N_4 = 42$.

Propiedad:

La última frecuencia acumulada es igual al tamaño muestral, al número total de datos.

Frecuencia relativa acumulada (F_i): Es el resultado de dividir cada frecuencia acumulada por el número total de datos.

$$F_i = \frac{N_i}{n}$$

Ejemplo:

$F_1 = 0'04$ $F_4 = \frac{42}{50} = 0'84$

Propiedad:

La última frecuencia relativa acumulada es siempre 1.

Tabla o distribución de frecuencias de una variable

Llamamos así a una tabla conteniendo el conjunto de diferentes valores que ha tomado una variable (los datos sin repetir) ordenados de menor a mayor con sus correspondientes frecuencias.

Actividades resueltas

La tabla de valores del ejemplo 1 del número de hijos

x_i	n_i	f_i	N_i	F_i
0	2	0’04	2	0’04
1	4	0’08	6	0’12
2	21	0’42	27	0’54
3	15	0’3	42	0’84
4	6	0’12	48	0’96
5	1	0’02	49	0’98
6	1	0’02	50	1

¿Cuál es el número de familias que tiene como máximo dos hijos?

Miramos la columna segunda n_i: 2 + 4 + 21 = 27 o miramos la columna cuarta, tercera fila: N_i: nos da 27

¿Cuántas familias tienen más de un hijo pero como máximo 3?

Miramos la columna segunda: 21 + 15 = 36 o miramos la columna cuarta y restamos las filas cuarta menos segunda 42 – 6 = 36.

¿Qué porcentaje de familias tiene más de 3 hijos?

Miramos en la columna tercera: 0'12 + 0'02 + 0'02 = 0'16 $\rightarrow$ 16 % o en la columnas quinta restando a la última fila la cuarta fila, es decir, 1 – 0'84 = 0'16 $\rightarrow$ 16 %.

Distribuciones de frecuencias agrupadas

Ahora vamos a trabajar con una distribución de frecuencias agrupadas con el ejemplo del precio de una habitación de hotel.

Ejemplo 2:

x_i	n_i	f_i	N_i	F_i
36	0	0	0	0
37	0	0	0	0
38	1	0,025	1	0,025
39	3	0,075	4	0,1
40	1	0,025	5	0,125
41	1	0,025	6	0,15
42	2	0,05	8	0,2
43	4	0,1	12	0,3
44	3	0,075	15	0,375
45	3	0,075	18	0,45
47	0	0	18	0,45
48	4	0,1	22	0,55
49	1	0,025	23	0,575
50	0	0	23	0,575
51	3	0,075	26	0,65
53	2	0,05	28	0,7
54	0	0	28	0,7
56	2	0,05	30	0,75
…	…	…	…	…
…	…	…	…	…

Esta tabla es demasiado grande y muy poco operativa.

Autores: Fernando Blasco, Ignasi Clausell y David Miranda
Ilustraciones: Banco de Imágenes de INTEF

Cuando la variable toma muchos valores, la tabla que se obtiene es demasiado grande y por tanto poco práctica. Esto nos va a ocurrir frecuentemente en el caso en que la variable a estudiar sea continua. La solución a este problema está en agrupar los diferentes valores de la variable en intervalos o **intervalos de clase**. Teniendo en cuenta que lo que ganamos en manejabilidad lo perdemos en información, es decir, los resultados serán aproximados.

Obtener intervalos de clase consiste en agrupar los datos en números relativamente pequeño de intervalos que cumplan:

- No se superpongan entre sí, de forma que no exista ambigüedad con respecto a la clase a que pertenece una observación particular.
- Cubran todo el rango de valores que tenemos en la muestra.

Llamamos:

- ✓ A las fronteras del intervalo, *límites inferior y superior* de clase y los denotaremos por l_i, L_i respectivamente.
- ✓ *Marca de clase* (c_i) al punto medio del intervalo, es decir, al promedio aritmético entre el límite inferior y el superior: $c_i = \frac{L_i + l_i}{2}$. Es el valor que tomaremos como representativo del intervalo o clase.
- ✓ *Amplitud* (a_i) es la diferencia entre el extremo superior e inferior: $a_i = L_i - l_i$.
- ✓ Al número de observaciones de una clase se le llama *frecuencia de clase* (n_i). Si dividimos esta frecuencia por el número total de observaciones, se obtiene la *frecuencia relativa de clase* (f_i), y del mismo modo que lo hacíamos para datos sin agrupar definimos (N_i) y (F_i).

Cómo construir una distribución de frecuencias agrupada en intervalos

1. Empezamos determinando el recorrido de la variable (Re) o *rango* de valores que tenemos en la muestra. Se define como la diferencia entre el mayor y el menor valor de la variable.
2. *Número de clases*. Depende del tamaño de la muestra. Para muestras de tamaño moderado n menor que 50, se suele elegir un número de clases o intervalos igual a $\sqrt{n}$. Para muestras mayores se utiliza la fórmula de *Sturges* $\frac{log(n)}{log(2)} + 1$, en general el número de intervalos no debe sobrepasar de 15 o 20, en casos de muestras muy grandes.
3. Determinamos la *amplitud de los intervalos*. Es más cómodo que la amplitud de todas las clases sea la misma (siempre que sea posible y excepto el primero y el último), si es así $a_i = a =$ Re/nº intervalos.
4. Tomaremos como regla general, a no ser que se indique lo contrario, hacer que el intervalo esté cerrado por la izquierda y abierto por la derecha (excepto el último intervalo).

Autores: Fernando Blasco, Ignasi Clausell y David Miranda
Ilustraciones: Banco de Imágenes de INTEF

Ejemplo:

Representa la distribución de frecuencias agrupadas para los datos del ejemplo del precio de las habitaciones de un hotel.

Recorrido: El menor valor es 33 y el mayor es 61, la diferencia es 28 y por tanto el recorrido es: Re = 28.

Número de clases: N = 40, hacemos que la tabla tenga 6 clases, pues $\sqrt{40} \approx 6$.

Amplitud: a = 28/6 = 4'67

Como la amplitud nos sale un número con decimales los intervalos nos van a quedar raros por tanto hacemos el arreglo siguiente:

Para que los intervalos nos queden con amplitud 5 tomamos como primer valor el 32'5 en lugar del 33 y como último el 62'5 en lugar del 61.

Amplitud: a = 5.

Así pues la tabla queda:

$[l_i, L_i[$	c_i	n_i	f_i	N_i	F_i
[32'5, 37'5[	35	3	0'075	3	0'075
[37'5, 42'5[	40	8	0'2	11	0'275
[42'5, 47'5[	45	14	0'35	25	0'625
[47'5, 52'5[	50	6	0'15	31	0'775
[52'5, 57'5[	55	4	0'1	35	0'875
[57'5, 62'5]	60	5	0'125	40	1

¿Cuántos hoteles tienen un precio entre 32'5 y 37'5 euros?

3

¿Cuántos hoteles tienen un precio superior a 47'5 €?

6 + 4 + 5 = 15

¿Qué porcentaje de hoteles cuestan como mucho 42'5 €?

27'5 %.

Actividades propuestas

3. Completa los datos que faltan en la tabla.

x_i	n_i	f_i	N_i	F_i
10	2	0'05	2	0'05
13	4	0'1	6	0'15
16			16	0'4
19	15			
22	6	0'15	37	0'925
25				

4. Completa los datos que faltan en la tabla.

$[l_i, L_i[$	n_i	f_i	N_i
[0, 10[	60		60
[10, 20[		0'4	
[20, 30[	30		170
[30, 40[		0'1	
[40, 50]			200

1.4. Gráficos

La forma de la distribución de frecuencias se percibe más rápidamente y quizás se retiene durante más tiempo en la memoria si la representamos gráficamente.

Diagrama de barras

Es la representación gráfica usual para las variables cuantitativas sin agrupar o para variables cualitativas. En el eje de abscisas representamos los diferentes valores de la variable x_i. Sobre cada valor levantamos una barra de altura igual a la frecuencia (absoluta o relativa).

Diagrama de sectores o pastel

Es el más usual en variables cualitativas. Se representan mediante círculos. A cada valor de la variable se le asocia el sector circular proporcional a su frecuencia.

Para hallar el ángulo usamos una regla de tres:

n ______________ 360º o 1
360º

$\boldsymbol{n_i}$ ______________ $ángulo_i$ $\boldsymbol{f_i}$
$ángulo_i$

Ejemplo 3:

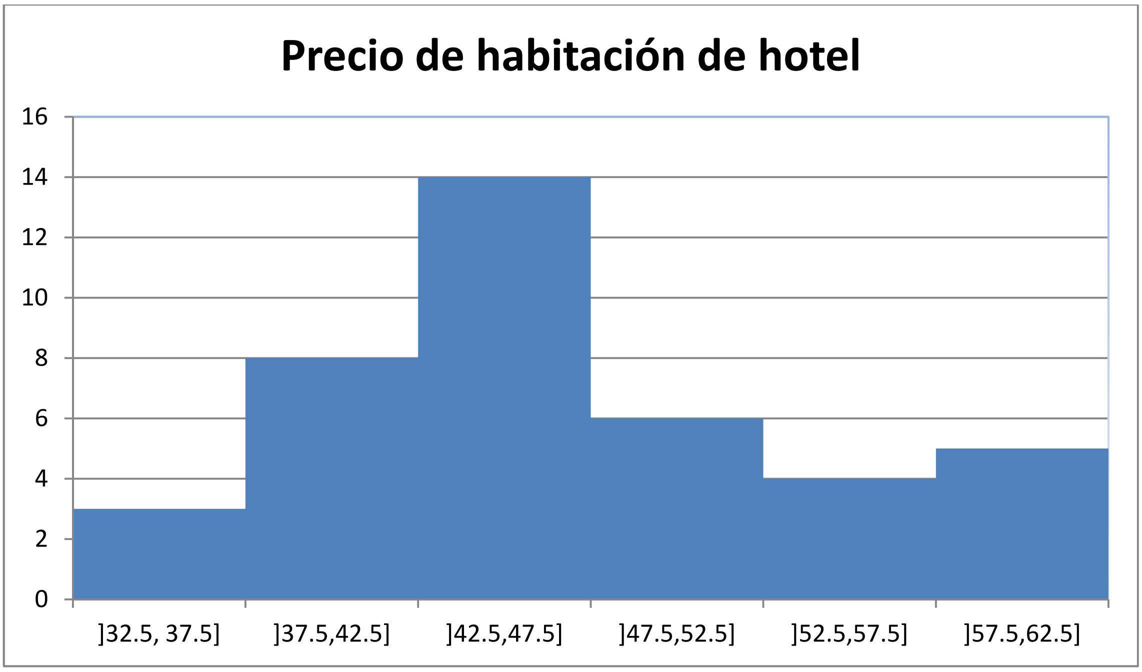
En unas votaciones de una comunidad de vecinos para decidir si cambia la antena de televisión de la comunidad, de 50 vecinos 25 votan a favor, 15 en contra y 10 se abstienen. Representa los datos mediante un diagrama de sectores.

$\boldsymbol{x_i}$	$\boldsymbol{f_i}$
A favor	0'5
En contra	0'3
Abstención	0'2

Histogramas

Es la representación gráfica equivalente al diagrama de barras para datos agrupados. En el eje de ordenadas representamos las clases y levantamos sobre cada clase rectángulos unidos entre sí de altura igual a la frecuencia de la clase (absolutas o relativas) si todas las clases tienen la misma amplitud y $\frac{n_i}{a_i}$ o $\frac{f_i}{a_i}$ si tienen distintas amplitudes.

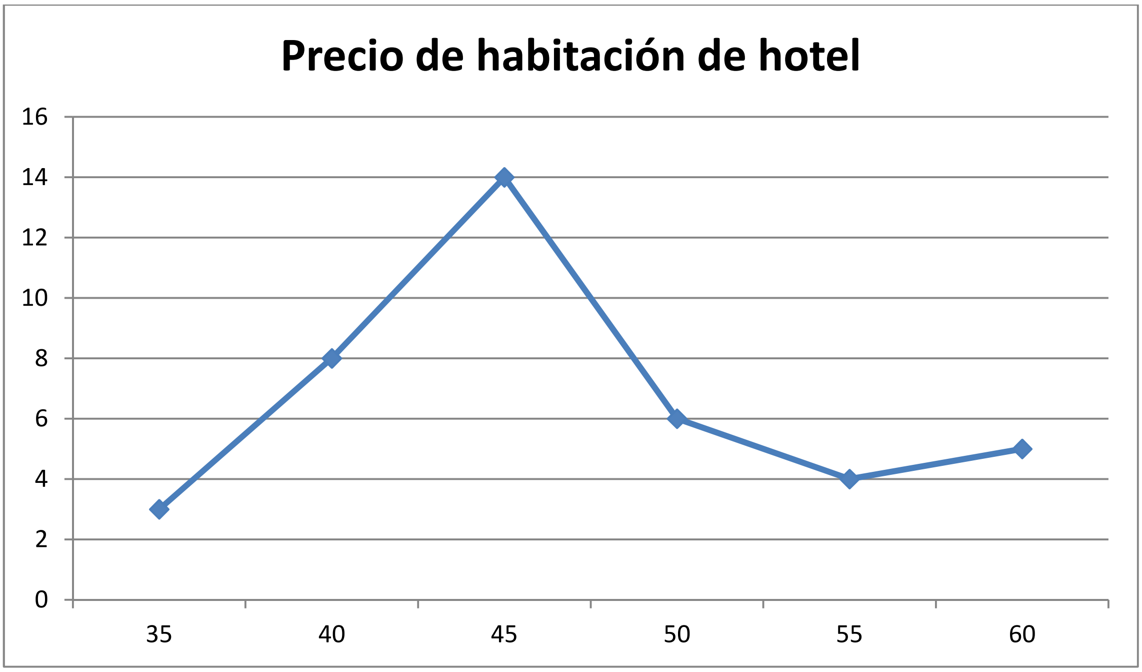
En cualquier caso, observa que, en un histograma el área de los rectángulos es proporcional a la frecuencia representada.



El histograma o diagrama de barras proporcionan mucha información respecto a la estructura de los datos (y si la muestra es representativa de la población, respecto a la estructura de la población): el valor central de la distribución, su dispersión y la forma de la distribución.

Polígono de frecuencias

Es la representación habitual para datos cuantitativos agrupados de las frecuencias (absolutas o relativas, acumuladas absolutas o relativas), mediante puntos se representan las frecuencias en el eje de ordenadas y la marca de clase en el de abscisas. Después se unen estos puntos por segmentos de rectas.



1.5. Parámetros estadísticos

Para datos cualitativos, la distribución de frecuencias proporciona un resumen conciso y completo de la muestra, pero para variables cuantitativas puede complementarse este resumen utilizando medidas descriptivas numéricas extraídas de los datos. Estas medidas son valores numéricos calculados a partir de la muestra y que nos resumen la información contenida en ella.

Parámetros estadísticos de posición

Media aritmética

Es el promedio aritmético de las observaciones, es decir, el cociente entre la suma de todos los datos y el número de ellos. (Teniendo en cuenta que si un valor se repite hay que considerar estas repeticiones).

$$\overline{x} = \frac{\sum_i x_i n_i}{n} = \sum_{i=1}^{k} x_i f_i$$

Si los datos están agrupados en intervalos utilizaremos las marcas de clase, c_i, en vez de x_i.

Es la medida de centralización más importante.

Ejemplo 1.

Número medio de hijos.

$$\overline{x} = \frac{0\cdot 2+1\cdot 4+2\cdot 21+3\cdot 15+4\cdot 6+5\cdot 1+6\cdot 1}{50} = \frac{126}{50} = 2'52 \text{ hijos.}$$

Utilizando los datos de las frecuencias relativas.

$$\overline{x} = 0\cdot 0'04+1\cdot 0'08+2\cdot 0'42+3\cdot 0'043+4\cdot 0'12+5\cdot 0'02+6\cdot 0'02 = 2'52 \text{ hijos.}$$

Ejemplo 2.

Precio medio.

Como tenemos los datos agrupados en intervalos utilizamos las marcas de clase:

$$\overline{x} = \frac{35\cdot 3+40\cdot 8+45\cdot 14+50\cdot 6+55\cdot 4+60\cdot 5}{40} = \frac{1875}{40} = 46'875 \text{ €}$$

O equivalentemente:

$$\overline{x} = 35\cdot 0'075+40\cdot 0'2+45\cdot 0'35+50\cdot 0'15+55\cdot 0'1+60\cdot 0'125 = 46'875 \text{ €.}$$

Propiedades.

1. Si a todos los valores de una variable les sumamos una constante, la media aritmética queda aumentada en esa constante.
2. Si a todos los valores de una variable los multiplicamos por una constante, la media aritmética queda multiplicada por la misma constante.
3. Si consideramos $y_i = a + bx_i$ siendo a y b dos constantes cualesquiera, la nueva media aritmética quedaría $\overline{y} = a + b\overline{x}$
4. La suma de todos los valores de la variable restándoles la media es cero.

Autores: Fernando Blasco, Ignasi Clausell y David Miranda
Ilustraciones: Banco de Imágenes de INTEF

Mediana

Es aquel valor que, al ordenar las observaciones de menor a mayor, ocupa el lugar central, dividiendo al conjunto de observaciones en dos partes iguales. Es decir, que deja a su derecha y a su izquierda el 50 por ciento de las observaciones.

Si el tamaño de la muestra, $\boldsymbol{n}$, es impar, necesariamente existe un dato que ocupa el lugar central, concretamente el dato que al ordenarlos está en la posición $(n+1)/2$; pero si n es par, son dos los datos que encontramos en el lugar central, los que ocupan los lugares $n/2$ y $(n/2)+1$, calculando entonces la mediana como el punto medio entre ambos datos.

Ejemplo:

Si tenemos los datos de 30 valores sobre el peso de los estudiantes de una clase ordenados de menor a mayor.

26'14	28'60	45'41	48'95	52'35	52'44	56'00	56'74	57'29	57'79	58'34	59'44	65'10
65'85	68'26	68'34	68'47	69'24	71'48	74'82	78'37	81'43	81'72	81'84	83'62	86'62
					87'82	91'93	92'78	96'97				

Como n = 30 es par, la mediana será el valor medio de los valores que ocupan las posiciones 15 y 16 en la tabla: 68'26 68'34

Mediana = Me = (68'26 + 68'34)/2 = 68'3 kg.

Ejemplo:

Las 13 primeras observaciones correspondientes al número de chocolatinas consumidas en un día por los estudiantes de una clase son:

0 1 2 2 2 2 2 2 2 3 3 3 3.

El dato que ocupa el valor central, es el que ocupa el lugar séptimo ya que hay 13 valores, ese dato es la mediana, por tanto la mediana es 2.

Me = 2.

Moda

Es aquel valor que tiene mayor frecuencia.

En el caso de las frecuencias agrupadas en intervalos se toma el intervalo que más veces se repite como la moda

Ejemplo:

Para la variable consumo de chocolatinas del ejemplo anterior la moda es Mo = 2

Ejemplo:

Para los datos del ejemplo 2 es el intervalo [42'5, 47'5).

Autores: Fernando Blasco, Ignasi Clausell y David Miranda
Ilustraciones: Banco de Imágenes de INTEF

Percentiles

El percentil p-ésimo es aquel valor que verifica la condición de que el p % de los datos son menores o iguales a él.

Así, el percentil 70 supone que el 70 % de los datos son menores o iguales a él.

Ejemplo:

Queremos calcular el percentil 30 de los datos del ejemplo sobre el peso de estudiantes, tendremos en cuenta que el 30 % de 30 datos que hay es 9, así buscamos el dato que ocupa esa posición en la ordenación del ejemplo 5, que es 57'29.

Si queremos calcular el percentil 15, tenemos en cuenta que el 15 % de 30 es 4'5, pero como este dato no pertenece a ninguna posición tomamos la aproximación por exceso, o sea tomamos el dato que ocupa la posición 5 por tanto el percentil 15 seria el dato 52'35. También es posible aproximarlo mejor mediante una interpolación lineal.

Nota:

Los percentiles 25, 50 y 75 reciben el nombre de **primer cuartil**, segundo cuartil y **tercer cuartil**.

Además el segundo cuartil que es el percentil 50 coincide con la **mediana**.

Si los datos están ordenados en intervalos tomamos el intervalo correspondiente al porcentaje del percentil como valor del percentil correspondiente.

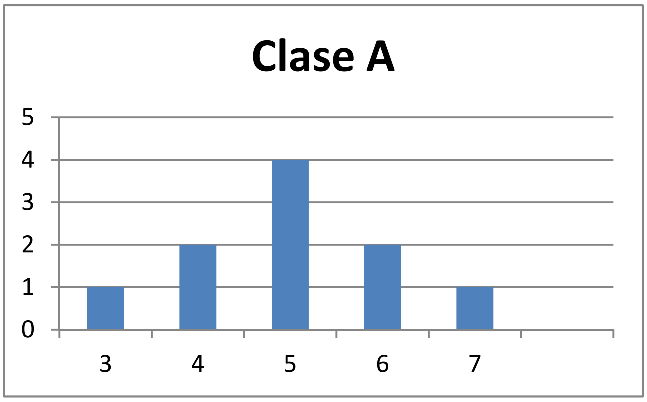
Parámetros estadísticos de dispersión

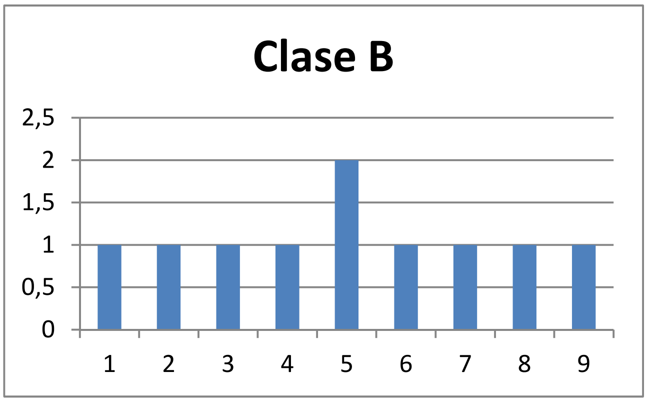
Las medidas de posición estudiadas en el apartado anterior, nos dan una información incompleta, por parcial, acerca de los datos.

Veamos un ejemplo:

Supongamos las notas de matemáticas de los estudiantes pertenecientes a dos clases distintas clase A y clase B, con 10 estudiantes cada una.

Clase A	4, 3, 5, 6, 4, 5, 5, 7, 5, 6
Clase B	1, 4, 3, 5, 6, 8, 2, 7, 5, 9





En los dos casos la media, como podemos calcular es 5, pero sus diagramas de frecuencias son muy distintos.

Autores: Fernando Blasco, Ignasi Clausell y David Miranda
Ilustraciones: Banco de Imágenes de INTEF

Los diagramas de frecuencias anteriores nos muestran que los valores se distribuyen simétricamente respecto a la nota 5, pero en la clase A existe una menor dispersión que en la clase B. ¿Cómo medir la distinta manera en que los valores se agrupan alrededor de la media? Las distintas medidas de dispersión proporcionan esta información. Al igual que ocurre para la posición, existen diversas formas para medir la dispersión, de entre ellas estudiaremos: rango, desviación típica, varianza y rango intercuartílico.

Rango

Es la diferencia entre el dato mayor y el dato menor.

Así por ejemplo

- El rango de las notas de la clase A vale 7 – 3 = 4 y el rango en la clase B vale 9 – 1 = 8, denotando mayor dispersión de la variable en la clase B.

La varianza y la desviación típica

Puesto que se trata de medir cómo se agrupan los datos alrededor de la media, podríamos utilizar como criterio las desviaciones de dichos datos respectos aquella, es decir, las diferencias entre la media y los datos y más concretamente la media de esas diferencias. Aunque a primera vista la sugerencia pueda ser buena, vamos a aplicarla a los valores de las notas de clase para evidenciar el inconveniente insalvable que una medida de este tipo tiene.

En los cuadros aparecen las notas de cada clase y en columnas sucesivas sus desviaciones respecto a la media y el cuadrado de estas desviaciones, al que aludiremos más tarde.

Al tratar de obtener la media de las diferencias, que recordemos es la suma de todas ellas divididas por su número, nos encontramos que dicha media es 0 en ambos casos, porque existiendo desviaciones positivas y negativas, unas anulan los efectos de las otras.

En realidad eso nos ocurrirá con cualquier otro conjunto de datos, porque puede demostrarse que esa es una propiedad que tienen las desviaciones respecto de la media.

Clase A		
Nota	$x_i - \overline{x}$	d_i^2
4	1	1
3	2	4
5	0	0
6	–1	1
4	1	1
5	0	0
5	0	0
7	–2	4
5	0	0
6	–1	1
Suma	0	12

Clase B		
Nota	$x_i - \overline{x}$	d_i^2
1	4	16
4	1	1
3	2	4
5	0	0
6	–1	1
8	–3	9
2	3	9
7	–2	4
5	0	0
9	–4	16
Suma	0	60

Autores: Fernando Blasco, Ignasi Clausell y David Miranda
Ilustraciones: Banco de Imágenes de INTEF

En las tablas aparecen las desviaciones respecto de la media y sus cuadrados para las notas de las dos clases.

Puesto que el uso de las desviaciones respecto de la media parece razonable, ¿cómo resolver el problema de que las sumas den 0? Una sencilla manera de hacerlo es utilizar, no las desviaciones, sino sus cuadrados. Al ser éstos cantidades positivas, su suma nunca podrá ser cero. De acuerdo con esto la varianza se define por la fórmula.

$$Varianza = s^2 = \frac{suma\ del\ cuadrado\ de\ las\ desviaciones}{n} = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 \cdot n_i}{n} = \frac{\sum_{i=1}^{k}(x_i)^2 \cdot n_i}{n} - \bar{x}^2$$

La **desviación típica** se define como la raíz cuadrada de la varianza y la designaremos por s.

$$s = \sqrt{Varianza}$$

Ejemplo:

- Para el ejemplo de las notas de las clases.

Clase A $\quad s^2 = \frac{12}{9} = 1'33 \quad s = \sqrt{1'33} = 1'15$

Clase B $\quad s^2 = \frac{60}{9} = 6'66 \quad s = \sqrt{6'66} = 2'58$

Que ponen de manifiesto la diferente distribución de los valores en un caso y en el otro.

Propiedad de la desviación típica

1. Aproximadamente el 68 % de los datos distan como mucho una desviación típica de la media.
2. Aproximadamente el 95 % de los datos distan como mucho dos desviaciones típicas de la media.
3. Aproximadamente más del 99 % de los datos distan como mucho tres desviaciones típicas de la media.

Rango intercuartílico.

Se define como la diferencia entre el tercer y el primer cuartil. El intervalo intercuartílico es el intervalo definido por los cuartiles primero y tercero, cuya longitud es, el rango intercuartílico. Este intervalo así definido contiene el 50 % de los datos.

Coeficiente variación

Si queremos comparar dos secuencias de datos, y decir en cual hay mayor dispersión, sobre todo en el caso en que sean datos expresados en diferentes unidades, con los parámetros definidos, desviación típica, intervalo intercuartílico, lo tenemos complicado, por eso se hace necesario definir el coeficiente de variación como,

$$CV = \frac{s}{\bar{x}} \cdot 100$$

Autores: Fernando Blasco, Ignasi Clausell y David Miranda
Ilustraciones: Banco de Imágenes de INTEF

Ejemplo:

- En el ejemplo de las calificaciones de dos clases nos permite comparar las dos secuencias de datos.

Clase A $CV = (1'15/5)\cdot 100 = 23$ %.

Clase B $CV = (2'58/5)\cdot 100 = 51'6$ %.

Llegando a la misma conclusión que percibíamos en los histogramas ya que la clase B tiene una mayor dispersión de las notas.

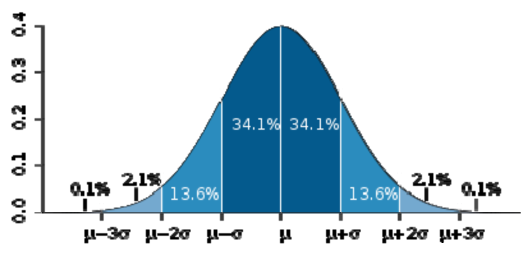
2.6. Interpretación conjunta de la media y la desviación típica

Hemos visto que la desviación típica nos mide la distancia de los datos respecto de la media. Nos da mucha información. Informa sobre cómo se agrupan los datos alrededor de la media.

La media y la desviación típica están relacionadas.

1. Aproximadamente el 68 % de los datos distan como mucho una desviación típica de la media.
2. Aproximadamente el 95 % de los datos distan como mucho dos desviaciones típicas de la media.
3. Aproximadamente más del 99 % de los datos distan como mucho tres desviaciones típicas de la media.

Si los datos que hemos recogido tuvieran una distribución normal (de momento no sabemos lo que esto significa exactamente dentro de la Estadística, pero puedes suponer que significa eso, que son normales, que no les pasa nada raro) resulta que en el intervalo entre la media menos una desviación típica y la media más una desviación típica están más del 68 % de los datos. En el intervalo entre la media menos 2 desviaciones típicas y la media más 2 desviaciones típicas están más del 95 % de los datos, y entre la media menos 3 desviaciones típicas y la media más 3 desviaciones típicas están más del 99'7 % de los datos.



Media y desviación típica. Imagen de wikipedia

Se podría decir que algo, por ejemplo la inteligencia de una persona, la altura de una planta o el peso de un animal... es normal si está dentro de ese intervalo $(\bar{x} - \sigma, \bar{x} + \sigma)$, que es inteligente, alto o pesado si está entre $(\bar{x} + \sigma, \bar{x} + 2\sigma)$, o que es un genio, gigante o muy pesado si está en el intervalo $(\bar{x} + 2\sigma, \bar{x} + 3\sigma)$.

Observa que estamos diciendo que prácticamente todos los datos distan de la media menos de tres desviaciones típicas y que más del 68 % distan menos de una desviación típica. Esto va a ser de gran utilidad pues conecta con otras ramas de la Estadística. Hasta ahora hemos estado describiendo lo que ocurre. Ahora vamos a poder tomar decisiones, inferir o predecir con una cierta probabilidad lo que va a ocurrir. Por eso vamos a estudiar a continuación las probabilidades.

Autores: Fernando Blasco, Ignasi Clausell y David Miranda
Ilustraciones: Banco de Imágenes de INTEF

Actividades propuestas

5. Clasifica las siguientes variables como cualitativas o cuantitativas, y estas últimas como continuas o discretas.

a) Intención de voto de un partido
b) Número de correos electrónicos que recibes en un mes.
c) Número de calzados.
d) Número de kilómetros recorridos en fin de semana.
e) Marcas de cerveza
f) Número de empleados de una empresa
g) Altura
h) Temperatura de un enfermo.

6. Muchas personas que invierten en bolsa lo hacen para conseguir beneficios rápidos, por ello el tiempo que mantienen las acciones es relativamente breve. Preguntada una muestra de 40 inversores habituales sobre el tiempo en meses que han mantenido sus últimas inversiones se recogieron los siguientes datos:

10'5 11'2 9'9 15'0 11'4 12'7 16'5 10'1 12'7 11'4 11'6 6'2 7'9 8'3 10'9
8'1 3'8 10'5 11'7 8'4 12'5 11'2 9'1 10'4 9'1 13'4 12'3 5'9 11'4 8'8
7'4 8'6 13'6 14'7 11'5 11'5 10'9 9'8 12'9 9'9

Construye una tabla de frecuencias que recoja esta información y haz alguna representación gráfica.

7. Investigados los precios por habitación de 50 hoteles de una provincia se han obtenido los siguientes resultados.

70 30 50 40 50 70 40 75 80 50 50 75 30 70 100 150 50 75 120
80 40 50 30 50 100 30 40 50 70 50 30 40 70 40 70 50 40 70 100
75 70 80 75 70 75 80 70 70 120 80.

Determinar:

a) Distribución de frecuencia de los precios, sin agrupar y agrupando en 5 intervalos de la misma amplitud.

b) Porcentaje de hoteles con precio superior a 75.

c) ¿Cuántos hoteles tienen un precio mayor o igual que 50 pero menor o igual a 100?

d) Representa gráficamente las distribuciones del apartado a).

Autores: Fernando Blasco, Ignasi Clausell y David Miranda
Ilustraciones: Banco de Imágenes de INTEF

8. El gobierno desea saber si el número medio de hijos por familia ha descendido respecto a la década anterior. Para ello se ha encuestado a 50 familias respecto al número de hijos y se ha obtenido los datos siguientes.

2 4 2 3 1 2 4 2 3 0 2 2 2 3 2 6 2 3 2 2 3 2 3 3 4 3
3 4 5 2 0 3 2 1 2 3 2 2 3 1 4 2 3 2 4 3 3 2 2 1.

a) Construye la tabla de frecuencias con estos datos.

b) ¿Cuántas familias tienen exactamente 3 hijos?

c) ¿Qué porcentaje de familias tienen exactamente 3 hijos?

d) ¿Qué porcentaje de familias de la muestra tiene más de dos hijos? ¿Y menos de tres?

e) Construye el gráfico que consideres más adecuado con las frecuencias no acumuladas.

f) Construye el gráfico que consideres más adecuado con las frecuencias acumuladas.

9. En un hospital se desea hacer un estudio sobre los pesos de los recién nacidos. Para ello se recogen los datos de los 40 bebes y se tiene:

3'2 3'7 4'2 4'6 3'7 3'0 2'9 3'1 3'0 4'5 4'1 3'8 3'9 3'6 3'2 3'5 3'0
2'5 2'7 2'8 3'0 4'0 4'5 3'5 3'5 3'6 2'9 3'2 4'2 4'3 4'1 4'6 4'2 4'5
4'3 3'2 3'7 2'9 3'1 3'5

a) Construye la tabla de frecuencias.

b) Si sabemos que los bebes que pesan menos de 3 kilos lo hacen prematuramente ¿Qué porcentaje de niños prematuros han nacido entre estos 40?

c) Normalmente los niños que nacen prematuros que pesan más de 3 kilos y medio no necesitan estar en incubadora. ¿Puedes decir que porcentaje de niños están en esta situación?

d) Representa gráficamente la información recibida.

10. En una finca de vecinos de Benicasim, se reúnen la comunidad de vecinos para ver si contratan a una persona para que les lleve la contabilidad. El resultado de la votación es el siguiente: 25 vecinos a favor de la contratación, 15 vecinos en contra y 5 vecinos se abstienen. Representa la información mediante un diagrama de sectores

11. Se toman ocho mediciones del diámetro interno de los anillos para los pistones del motor de un automóvil. Los datos en mm son:

74'001 74'003 74'015 74'000 74'005 74'002 74'005 74'004

Calcula la media y la mediana de estos datos. Calcula también la varianza, la desviación típica y el rango de la muestra.

12. Dada la distribución de datos 38432 384343 38436 38438 38440 con frecuencias 4, 8, 4, 3, 8, halla la media de la distribución.

13. La distribución de los salarios en la industria turística española es la que figura en la tabla. Calcula:

a) El salario medio por trabajador (marcas de clase del último intervalo 20000

b) El salario más frecuente.

c) El salario tal que la mitad de los restantes sea inferior a él.

$[l_i, L_i[$	n_i
[0,1500[	2145
[1500, 2000[	1520
[2000, 2500[	840
[2500, 3000[	955
[3000, 3500[	1110
[3500, 4000[	2342
[4000, 5000[	610
[5000, 10000[	328
≥10000	150

14. Calcula la mediana, la moda, primer y tercer cuartil y nonagésimo percentil de la distribución:

x_i	n_i
5	3
10	7
15	5
20	3
25	2

15. Se han diseñado dos unidades gemelas de plantas pilotos y han sido puestas en funcionamiento en un determinado proceso. Los resultados de los diez primeros balances en cada una de las unidades han sido los siguientes:

Unidad A 97'8 98'9 101'2 98'8 102'0 99'0 99'1 100'8 100'9 100'5

Unidad B 97'2 100'5 98'2 98'3 97'5 99'9 97'9 96'8 97'4 97'2

a) Haz una representación gráfica de estas muestras.

b) Determina las medias y las varianzas.

Autores: Fernando Blasco, Ignasi Clausell y David Miranda
Ilustraciones: Banco de Imágenes de INTEF

16. En cierto barrio se ha encontrado que las familias residentes se han distribuido, según su composición de la forma siguiente:

Composición	***Nº de familias***
0-2	110
2-4	200
4-6	90
6-8	75
8-10	25

a) ¿Cuál es el número medio de personas por familia?

b) ¿Cuál es el tamaño de la familia más frecuente?

c) Si solo hubiera plazas de aparcamiento para el 75 % de las familias y estas se atendieran por familias de mayor tamaño a menor, ¿qué componentes tendría que tener una familia para entrar en el cupo?

d) Número de miembros que tienen como máximo el 85 % de las familias.

17. Al lanzar 200 veces un dado se obtuvo la siguiente distribución de frecuencias.

x_i	1	2	3	4	5	6
n_i	a	32	35	33	b	35

Halla la mediana y la moda de la distribución, sabiendo que la media aritmética es 3’6.

18. Los siguientes datos son medidas de la capacidad craneal de un grupo de homínidos:

84, 49,61, 40, 83, 67, 45, 66, 70, 69, 80, 58, 68, 60, 67, 72, 73, 70, 57, 63, 70, 78, 52, 67, 53, 67, 75, 61, 70, 81, 76, 79, 75, 76, 58, 31.

a) Calcula la media y la mediana muestrales.

b) Halla los cuartiles primero y tercero.

c) Halla los percentiles cincuenta y noventa.

d) Calcula el rango muestral.

e) Calcula la varianza muestral y la desviación estándar muestral.

19. Los siguientes datos proceden de un estudio de contaminación del aire.

6’5 2’1 4’4 4’7 5’3 2’6 4’7 3’0 4’9 8’6 5’0 4’9 4’0 3’4 5’6 4’7 2’7
2’4 2’7 2’2 5’2 5’3 4’7 6’8 4’1 5’3 7’6 2’4 2’1 4’6 4’3 3’0 4’1 6’1
4’2

a) Construye un histograma.

b) Determina los cuartiles.

c) Calcula la media y la desviación típica.

Autores: Fernando Blasco, Ignasi Clausell y David Miranda
Ilustraciones: Banco de Imágenes de INTEF

CURIOSIDADES. REVISTA

CONTRA LA SUPERSTICIÓN, ESTADÍSTICA

Vivimos en un mundo dominado por la ciencia y la tecnología, a pesar de ello las supersticiones y las creencias seudocientíficas siguen dominando entre la población general, incluso más que en otras épocas. La Estadística es un arma importante para desenmascarar algunas afirmaciones que circulan impunemente y que mucha gente cree, como las derivadas de la astrología. Existen cientos de estudios que prueban que aunque existan coincidencias entre el signo astrológico de las personas y sus formas de ser, gustos, comportamientos, profesiones, etc. éstas están siempre en torno a la media estadística.

Una creencia muy habitual es que los nacimientos se producen con mayor frecuencia durante los días, y especialmente las noches, de luna llena. Resultaría sencillo coger los registros civiles y comprobar si eso es verdad, pero los que afirman semejante dato nunca se molestan en hacerlo. Recientemente se ha puesto de manifiesto mediante el análisis de los datos de un conjunto de estudios al respecto que las variaciones de nacimientos entre fases lunares son de apenas un 1 %, sin embargo también el mismo estudio ha puesto de manifiesto que el 60 % de los nacimientos se producen entre las 6 de la mañana y las seis de la tarde, mostrando así una diferencia mucho más significativa que suele tener su explicación en la organización de los hospitales.

EL EFECTO PLACEBO Y EL EFECTO NOCEBO

Antes de que un medicamento pueda comercializarse debe superar una serie de estrictas pruebas que arrojen seguridad acerca de su eficacia curativa.

Una de las pruebas más comunes consiste en seleccionar una muestra de enfermos y dividirlos aleatoriamente en dos grupos; un grupo recibe el medicamento, y el otro, sin saberlo, una sustancia en apariencia igual, pero sin ningún poder terapéutico: un placebo.

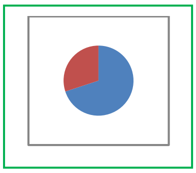
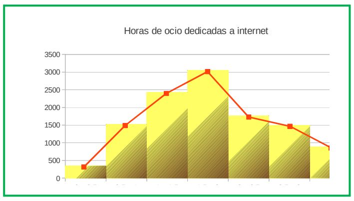
De esta forma, al final del ensayo pueden compararse los resultados entre los dos grupos y determinar la eficacia del medicamento. Para ello se emplean herramientas estadísticas como la correlación.

Sorprendentemente, hay un número significativo de pacientes que, habiendo recibido el placebo, mejoran de forma ostensible. Por ejemplo, esta contrastado que, en muchas enfermedades relacionadas con el dolor, entre el 10 % y el 15 % de los pacientes experimenta un alivio notable habiendo seguido un tratamiento exclusivamente de placebo.

Este fenómeno se conoce como efecto placebo, y se sabe que lo causan las sustancias neurotransmisoras que produce el cerebro ante la expectativa de que la dolencia va a mejorar. Y tiene su contrapartida: el **efecto nocebo**, el empeoramiento que se produce a causa de la creencia de que una medida terapéutica va a resultar perjudicial.

Autores: Fernando Blasco, Ignasi Clausell y David Miranda
Ilustraciones: Banco de Imágenes de INTEF

RESUMEN

Población	Colectivo sobre el que se hace el estudio	Estudiantes de todo Madrid
Muestra	Subconjunto de la población que permita obtener características de la población complete.	Alumnos se 3º de ESO seleccionados
Individuo	Cada uno de los elementos de la población o muestra	Juan Pérez
Variables estadística	Cuantitativa discreta Cuantitativa continua Cualitativa	Número de pie que calza Estatura Deporte que practica
Gráficos estadísticos	Diagrama de barras Histograma de frecuencias Polígono de frecuencias Diagrama de sectores 	
Media	$\bar{x} = \frac{\sum x_i}{n}$ = $(x_1 + x_2 + ... + x_n)/n$ $\bar{x} = \frac{\sum_i x_i n_i}{n} = \sum_{i=1}^{k} x_i f_i$	Con los datos: 8, 2, 5, 10 y 10 *Media* = 35/5 = 7 $\bar{x} = \frac{0\cdot2+1\cdot4+2\cdot21+3\cdot15+4\cdot6+5\cdot1+6\cdot1}{50} = \frac{126}{50} = 2'5$
Moda	Es el valor más frecuente	*Mo* = 10
Mediana	Deja por debajo la mitad	4 < 6 < **8** < 10 = 10. *Me* = 8.
Rango o recorrido	Es la diferencia entre el dato mayor y el dato menor.	10 – 2 = 8
Desviación media	Es la media de las distancias de los datos a la media de los datos de los que dispongamos.	(|8–7|+|2–7|+|5–7|+|10–7|+|10–7|)/5 =(1+5+2+3+3)/5 = 14/5 = DM
Varianza	Es la media de los cuadrados de las distancias de los datos a la media: $s^2 = \frac{\sum_{i=1}^{n}(x_i - m)^2}{n} = \frac{\sum_{i=1}^{n} x_i^2}{n} - m^2$ $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot f_i}{n} = \frac{\sum_{i=1}^{k}(x_i)^2 \cdot f_i}{n} - \bar{x}^2$	V = (1 + 25 + 4 + 9 + 9)/5 = 47/5 = 9,4
Desviación típica	Es la raíz cuadrada de la varianza= $\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n} - m^2}$	σ = $\sqrt{47/5}$ = 3,06

Autores: Fernando Blasco, Ignasi Clausell y David Miranda
Ilustraciones: Banco de Imágenes de INTEF

EJERCICIOS Y PROBLEMAS

1. Se han recogido los datos sobre el número de hijos que tienen 20 matrimonios. ¿Cómo es la variable utilizada? Escribe una tabla de frecuencias de los datos recogidos y representa los datos en un diagrama de sectores:

   3, 1, 1, 2, 0, 2, 3, 1, 1, 1, 1, 0, 3, 2, 1, 2, 1, 2, 2, 3.

2. Con los datos del problema anterior calcula la media, la mediana, la moda y los cuartiles.
3. Con los datos del problema anterior calcula el rango, la desviación media, la varianza, la desviación típica y el intervalo intercuartílico.
4. Representa esos datos en un diagrama de cajas.
5. 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	10	140	210	340	210	90

a) Representa los datos en un histograma.

b) Calcula la media y la desviación típica.

c) Determina el intervalo donde se encuentran la mediana.

6. Se pregunta a un grupo de personas por el número de televisores que hay en su hogar y los resultados son:

Número de televisores	0	1	2	3	4	5
Número de hogares	2	27	15	4	2	1

¿Qué tipo de variables es? Represente los datos en la representación que te parezca más adecuada.

Calcula la media y la desviación típica-

7. Con los datos del problema anterior calcula la mediana y el intervalo intercuartílico.
8. En un centro escolar se ha recogido información sobre el número de ordenadores en las casas de 100 familias y se han obtenido los siguientes resultados:

Número ordenadores	0	1	2	3	4
Número de familias:	24	60	14	1	1

Representa los datos en un diagrama de barras y calcula la media, la mediana y la moda.

9. Con los datos del problema anterior calcula el rango, la desviación media, la varianza y la desviación típica. Haz un diagrama de cajas.
10. Se pregunta a un grupo de personas por el número de veces que han visitado al dentista en el último año. Las respuestas obtenidas se recogen en la siguiente tabla:

Número de visitas:	1	2	3	4	5
Número de personas:	13	18	7	5	7

Representa los datos en un diagrama de sectores y calcula la media, la mediana y la moda.

Autores: Fernando Blasco, Ignasi Clausell y David Miranda
Ilustraciones: Banco de Imágenes de INTEF

11. Se pregunta a un grupo de personas por el número de veces que han visitado al dentista en el último año. Las respuestas obtenidas se recogen en la siguiente tabla:

Número de visitas:	1	2	3	4	5
Número de personas:	13	18	7	5	7

Calcula el rango, la desviación media, la varianza y la desviación típica.

12. En las elecciones de 2014 al Parlamento Europeo se obtuvieron los siguientes escaños por grupo parlamentario (DM: demócrata – cristianos; S: socialistas; L: Liberales; V: verdes; C: conservadores; I: izquierda unitaria; LD: Libertad y democracia; NI: No inscritos; Otros).

Partidos	DM	S	L	V	C	I	LD	NI	Otros	Total
Escaños	213	190	64	52	46	42	38	41	65	751

¿Qué representación de los datos te parece más adecuada? ¿Puedes calcular la media o el rango? ¿Qué tipo de variables es la de la tabla?

13. En las elecciones de 2014 al Parlamento Europeo se obtuvieron los siguientes escaños por alguno de los estados miembro:

Estado	Alemania	España	Francia	Italia	Polonia	Reino Unido	Portugal	Grecia	Otros	Total
Escaños	96	54	74	73	51	73	21	21		751

¿Qué representación de los datos te parece más adecuada? ¿Puedes calcular la media o el rango? ¿Qué tipo de variables es la de la tabla? Determina el número de escaños de los otros países miembros de la Unión Europea.

14. En las elecciones de 2004, 2009, 2014 al Parlamento Europeo se obtuvieron los siguientes porcentajes de votos por algunos de los estados miembros:

Estado	Alemania	España	Francia	Italia	Reino Unido	Portugal	Grecia	Bélgica	% total
2004	43	45’14	42’76	71’72	38’52	38’6	63’22	90’81	45’47
2009	43’27	44’87	40’63	65’05	34’7	36’77	52’61	90’39	43
2014	47’6	45’9	43’5	60	36	34’5	58’2	90	43’09

¿Qué representación de los datos te parece más adecuada? ¿Puedes calcular la media o el rango? ¿Qué tipo de variables es la de la tabla? Ordena a los países de mayor a menos porcentaje de votantes en las elecciones de 2014.

15. Con los datos del problema anterior sobre las elecciones de 2004’ 2009’ 2014 al Parlamento Europeo se obtuvieron los siguientes porcentajes de votos por algunos de los estados miembros:

Estado	Alemania	España	Francia	Italia	Reino Unido	Portugal	Grecia	Bélgica	% total
2004	43	45’14	42’76	71’72	38’52	38’6	63’22	90’81	45’47
2009	43’27	44’87	40’63	65’05	34’7	36’77	52’61	90’39	43
2014	47’6	45’9	43’5	60	36	34’5	58’2	90	43’09

Representa en un polígono de frecuencias los porcentajes de participación del total de los estados miembros.

16. Con los datos del problema anterior sobre las elecciones de 2004, 2009, 2014 al Parlamento Europeo se obtuvieron los siguientes porcentajes de votos por algunos de los estados miembros:

Estado	Alemania	España	Francia	Italia	Reino Unido	Portugal	Grecia	Bélgica	% total
2004	43	45’14	42’76	71’72	38’52	38’6	63’22	90’81	45’47
2009	43’27	44’87	40’63	65’05	34’7	36’77	52’61	90’39	43
2014	47’6	45’9	43’5	60	36	34’5	58’2	90	43’09

Separa los Estados Miembros en dos grupos, los que tuvieron un porcentaje superior al porcentaje medio y los que lo tuvieron menor en 2004. Haz lo mismo para 2014. ¿Son los mismos? Analiza el resultado.

17. Con los datos del problema anterior sobre las elecciones de 2004, 2009, 2014 al Parlamento Europeo se obtuvieron los siguientes porcentajes de votos por algunos de los estados miembros:

Estado	Alemania	España	Francia	Italia	Reino Unido	Portugal	Grecia	Bélgica	% total
2004	43	45’14	42’76	71’72	38’52	38’6	63’22	90’81	45’47
2009	43’27	44’87	40’63	65’05	34’7	36’77	52’61	90’39	43
2014	47’6	45’9	43’5	60	36	34’5	58’2	90	43’09

Calcula el porcentaje de participación medio para Alemania en esas tres convocatorias y la desviación típica. Lo mismo para España, para Bélgica y para Portugal.

18. En las elecciones de 2014 al Parlamento Europeo los resultados de España han sido:

Censo	Total de votantes	Abstención	Votos nulos	Votos en blanco
35.379.097	15.920.815	19.458.282	290.189	357.339

Representa en un diagrama de sectores estos datos. Haz una tabla de porcentajes: el censo es el 100 %. Determina los otros porcentajes. ¿Consideras que ha ganado la abstención?

19. En las elecciones de 2014 al Parlamento Europeo los resultados de España han sido:

PP	PSOE	Izquierda plural	Podemos	UPyD	Otros	Total de votantes
4.074.363	8.001.754	1.562.567	1.245.948	1.015.994		15.920.815

Determina el número de votos de los otros partidos. Representa en un diagrama de barras estos datos. Haz una tabla de porcentajes para cada partido. Tienes que distribuir 54 escaños, ¿cómo los distribuirías por partidos?

20. Se conoce el volumen semanal de residuos sólidos recogidos en m3 durante 10 semanas, en un municipio pequeño:

25'5, 27'1, 31'8, 34'2, 38'9, 21'3, 28'7, 33'2, 36'5, 39'6

Calcula:

a) Las medidas de **centralización**: la media, mediana, moda

b) Las medidas de **dispersión**: desviación típica, varianza, coeficiente de variación, valor mínimo, valor máximo, recorrido, primer cuartil, tercer cuartil e intervalo intercuartílico.

c) Haz una representación gráfica en **serie temporal**, que permita observar tendencias, ciclos y fluctuaciones. Recuerda que en una serie temporal, en el eje de abscisas está el tiempo de observación y en el eje de ordenadas la magnitud de observación.

21. Una compañía de seguros desea establecer una póliza de accidentes. Para ello, selecciona al azar a 100 propietarios y les pregunta cuántos euros han gastado en reparaciones del automóvil. Se han agrupado en intervalos los valores de la variable obtenidos:

Euros	[0, 100)	[100, 200)	[200, 400)	[400, 600)	[600, 800)	[800, 3000)
Número de personas	20	20	10	20	20	10

a) Calcula las marcas de clase y escribe en tu cuaderno una tabla de frecuencias absolutas, frecuencias relativas, frecuencias acumuladas absolutas y frecuencias relativas acumuladas.
b) Representa los datos en un diagrama de barras, otro de líneas y uno de sectores.
c) Representa un histograma de frecuencias relativas. *Cuidado*: Los intervalos no son todos iguales.
d) Calcula la media y la desviación típica.
e) Calcula la mediana y los cuartiles.

22. Se ha preguntado a 40 alumnos por el número de hermanos que tenía, y se ha obtenido

Número de hermanos	**0**	**1**	**2**	**3**	**4**	**5**	**6 o más**
Número de veces	5	15	7	6	4	2	1

a) Representa un diagrama de barras de frecuencias absolutas y un diagrama de líneas de frecuencias relativas.
b) Calcula la media, la mediana y la moda.

23. Se ha preguntado a 50 estudiantes de un curso por el número de hermanos que tenía, y se ha obtenido:

Número de hermanos	**0**	**1**	**2**	**3**	**4**	**5**	**6 o más**
Número de veces	8	19	8	7	5	2	1

a) Representa los datos en un diagrama de barras de frecuencias absolutas, en un diagrama de líneas de frecuencias relativas, y en un diagrama de sectores.
b) Haz un histograma.
c) Calcula la media, la mediana y la moda. Calcula los cuartiles.
d) Calcula la varianza, la desviación típica, el recorrido y el intervalo intercuartílico.

Autores: Fernando Blasco, Ignasi Clausell y David Miranda
Ilustraciones: Banco de Imágenes de INTEF

Utiliza una hoja de cálculo con el ordenador

Se conoce el volumen semanal de residuos sólidos recogidos en m^3 durante las 52 semanas de un año, en un municipio pequeño:

25'5, 27'1, 31'8, 34'2, 38'9, 21'3, 28'7, 33'2, 36'5, 39'6, 25'2, 24'7, 23'2, 23'3, 22'2, 26'4, 26'7, 29'6, 31'3, 30'5, 28'3, 29'1, 26'7, 25'2, 24'5, 23'7, 25'4, 27'2, 31'7, 34'5, 38'4, 21'2, 28'1, 33'7, 36'8, 39'9, 31'7, 34'4, 38'2, 21'9, 28'1, 33'5, 25'2, 24'7, 23'2, 23'3, 22'2, 26'4, 25'9, 24'1, 23'2, 23'6, 26'4.

Calcula, utilizando Excel u otra hoja de cálculo:

24. Parámetros estadísticos

a) Las medidas de centralización: la media, mediana, moda

b) Las medidas de **dispersión**: desviación típica, varianza, coeficiente de variación, valor mínimo, valor máximo, recorrido, primer cuartil, tercer cuartil e intervalo intercuartílico.

c) Otros coeficientes: coeficiente de asimetría y coeficiente de curtosis que encuentres. Investiga las posibilidades del ordenador para obtener parámetros estadísticos.

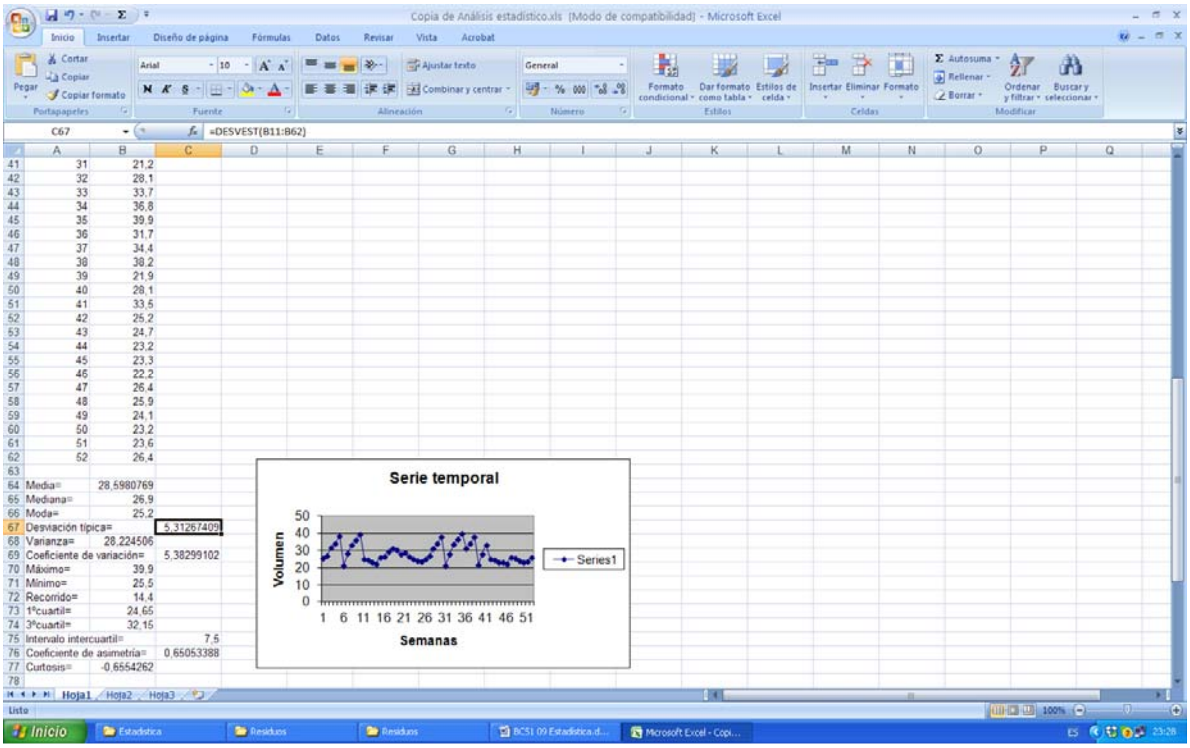
d) Haz una representación gráfica en **serie temporal**, que permita observar tendencias, ciclos y fluctuaciones. Recuerda que en una serie temporal, en el eje de abscisas está el tiempo de observación y en el eje de ordenadas la magnitud de observación.

Para ello, escribe en la casilla A12, 1, en A13, 2, y arrastra para escribir el orden de las semanas, hasta que aparezca el 52. Escribe en la columna B el volumen recogido cada semana.

En la casilla A11 un título, por ejemplo, "Residuos sólidos".

En la casilla C12 escribe Media, y en la casilla D12 calcúlala usando la función PROMEDIO. De igual forma calcula los otros parámetros.

Observa un trozo de pantalla con algunos resultados:

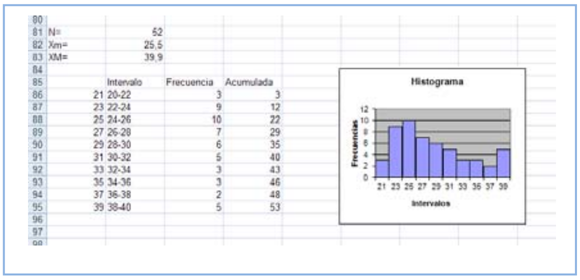


Autores: Fernando Blasco, Ignasi Clausell y David Miranda
Ilustraciones: Banco de Imágenes de INTEF

25. Los datos de la práctica anterior se quieren representar en un **histograma** para mejor determinar su distribución. Para ello:

a) Indica el número total de datos, N, el menor valor: X_m, el mayor valor, X_M, y el recorrido R.

b) La cantidad de barras del histograma, k, se suele tomar, para menos de 50 datos, entre 5 y 7. Para N entre 50 y 100, entre 6 y 10. Para N entre 100 y 250, entre 7 y 12. Y para N mayor de 250, entre 10 y 20. En este caso N es igual a 52, luego el número de barras podría ser entre 6 y 10. Al dividir R entre 10 se obtiene 1,87 que sería el intervalo de clase. Para facilitar la división en clases fijamos el intervalo de clase, h, en 2, y el número de barras, k, en 10. Para no tener valores en los límites de clase tomamos el inicio del primer intervalo en 20. Así, los intervalos son: (20, 22), de valor central: 21; [22, 24), de valor central 23... Ahora ya se puede construir la tabla de frecuencias y dibujar el histograma.



c) Calcula y representa en el histograma los puntos m, $m \pm s$, $m \pm 2s$, $m \pm 3s$, donde m y s son la media y la desviación típica, respectivamente

- Vamos a investigar qué ocurre al hacer un cambio de variables. Dijimos que si consideramos $\boldsymbol{y_i = a + bx_i}$ siendo a y b dos constantes cualesquiera, la nueva media aritmética quedaría $\bar{y} = a + b\bar{x}$.

a) Abre Excel. Introduce los datos: X = 255, 271, 318, 342, 389,... en la columna A, a partir de la fila 11. ¿Qué cambio de variable se ha hecho? Observa: $x = X/10$.

b) En la columna C, a partir de la fila 11 escribe los límites de clase, en la columna D el valor medio, en la columna E vamos a contar las frecuencias absolutas y en la columna F las frecuencias acumuladas. Utiliza la función CONTAR.SI para contar. Por ejemplo, escribe en E11, CONTAR.SI(A11:A63; <220). En F11 escribe =E11. En E12 escribe CONTAR.SI(A11:A63; <240)-F11. Completa la tabla de frecuencias. Escribe títulos en la fila 10.

c) Calcula la media y la desviación típica. Para ello escribe en la fila 3 y 4, columna B, las funciones =PROMEDIO(A11:A63) y =DESVEST(A11:A63). Escribe los resultados con 2 decimales.

d) ¿Cómo obtienes ahora la media y la desviación típica de los datos reales? ¿Cómo deshaces el cambio? Si no lo recuerdas, o no tienes seguridad, investígalo. Calcula la media y la desviación típica, antes y después del cambio. Escribe este resultado, en general, para un cambio de variables lineal $y = ax+b$.

e) Dibuja el histograma. No olvides nunca indicar las unidades en ambos ejes, y toda la información que ayude a comprender el gráfico. Añade siempre el tamaño, N, y los valores de la media y la desviación típica.

f) Discute el resultado. ¿Es grande la dispersión? La distribución, ¿es simétrica?

- Otra investigación: Vamos a investigar la distribución de la media. Para ello vamos a tomar muestras de tamaño 5. Utiliza la columna G. En G11 escribe =PROMEDIO(B11:B15), en G12 la media de B16 a B20, y así hasta el final. Tenemos calculadas las 10 medias de muestras de tamaño 5. Calcula la media y la desviación típica de estas medias. Compara con los resultados anteriores. Escribe en tu cuaderno las conclusiones.

Autores: Fernando Blasco, Ignasi Clausell y David Miranda
Ilustraciones: Banco de Imágenes de INTEF

AUTOEVALUACIÓN

1. Se hace un estudio sobre el color que prefieren los habitantes de un país para un coche. La variable utilizada es:

a) cuantitativa b) cualitativa c) cuantitativa discreta d) cuantitativa continua

2. En un histograma de frecuencias la altura de los rectángulos es:

a) proporcional al área b) igual a la frecuencia absoluta

c) proporcional a la frecuencia relativa d) proporcional a la frecuencia acumulada

3. Ana ha obtenido en Matemáticas las siguientes notas: 7, 8, 5, 10, 8, 10, 9 y 7. Su nota media es de:

a) 7,6 b) 8,2 c) 8 d) 9

4. En las notas anteriores de Ana la mediana es:

a) 9 b) 8 c) 7,5 d) 8,5

5. En las notas anteriores de Ana la moda es:

a) 10 b) 8 c) 7 d) 7, 8 y 10

- Realizamos una prueba a 20 aspirantes a un puesto de grabador consistente en un dictado con cierto tiempo de duración (en minutos) y luego contar el número de errores cometidos al transcribirlo a ordenador. Los resultados fueron.

Tiempo	7	6	5	4	5	8	7	8	9	6	5	8	6	8	7	8	7	6	6	9
Errores	8	7	6	6	7	10	9	9	10	8	6	10	8	9	8	8	7	8	6	8

6. La media de errores es

a) 6’75 b) 7 c) 7’9 d) 6’9

7. La media de tiempos es

a) 6’75 b) 7 c) 7’9 d) 6’9

8. La desviación típica de errores es

a) 1 b) 1’41 c) 1’33 d) 1’2

9. La desviación típica de tiempos es

a) 1 b) 1’41 c) 1’33 d) 1’2

10. El primer cuartil, la mediana y el tercer cuartil de los tiempos valen respectivamente:

a) 7, 8 y 9 b) 5, 6 y 7 c) 5’9, 6’1 y 7’3 d) 6, 7 y 8

11. El primer cuartil, la mediana y el tercer cuartil de los errores valen respectivamente:

a) 7, 8 y 9 b) 5, 6 y 7 c) 6’5, 7’5 y 8’5 d) 6, 7 y 8

Autores: Fernando Blasco, Ignasi Clausell y David Miranda
Ilustraciones: Banco de Imágenes de INTEF