# Matemáticas Generales
# 1º Bachillerato
# Capítulo 11: Estadística



**www.apuntesmareaverde.org.es**

**Autor: Ignasi Clausell**

**Revisores: Raquel Caro y Luis Carlos Vidal**

**Ilustraciones: Del autor, de Wikipedia y del Banco de Imágenes de INTEF**

## Índice

1. ESTADÍSTICA DESCRIPTIVA UNIDIMENSIONAL
- 1.1. MÉTODO ESTADÍSTICO
- 1.2. TIPOS DE VARIABLES
- 1.3. DISTRIBUCIONES DE FRECUENCIAS
- 1.4. GRÁFICOS
- 1.5. PARÁMETROS ESTADÍSTICOS

2. ESTADÍSTICA DESCRIPTIVA BIDIMENSIONAL
- 2.1. INTRODUCCIÓN. TABLAS DE CONTINGENCIA
- 2.2. DISTRIBUCIÓN DE FRECUENCIAS CONJUNTAS
- 2.3. DISTRIBUCIÓN DE FRECUENCIAS MARGINALES
- 2.4. DISTRIBUCIÓN DE FRECUENCIAS CONDICIONADAS
- 2.5. INDEPENDENCIA ESTADÍSTICA
- 2.6. DIAGRAMA DE DISPERSIÓN. NUBE DE PUNTOS

3. COVARIANZA
- 3.1. IDEA CORRELACIÓN. COVARIANZA
- 3.2. COEFICIENTE CORRELACIÓN LINEAL
- 3.3. RECTA REGRESIÓN LINEAL. REGRESIÓN CUADRÁTICA
- 3.4. PREDICCIÓN Y CAUSALIDAD
- 3.5. COEFICIENTE DE DETERMINACIÓN

BOE: Interpretación y análisis de información estadística en diversos contextos. Organización de los datos procedente de variables bidimensionales: distribución conjunta y distribuciones marginales y condicionadas. Análisis de la dependencia estadística. Estudio de la relación entre dos variables mediante la regresión lineal o cuadrática: valoración gráfica de la pertinencia del ajuste. Coeficientes de correlación lineal y de determinación: cuantificación de la relación lineal, predicción y valoración de su fiabilidad en contextos científicos, económicos, sociales, etc. Calculadora, hoja de cálculo o software específico en el análisis de datos estadísticos.

## Resumen

Vamos a repasar los conceptos de estadística unidimensional aprendidos en cursos anteriores, revisando las tablas de frecuencias, calculando las medidas de centralización, media, mediana y moda y las medidas de dispersión, varianza y desviación típica.

El estudio unidimensional lo ampliaremos al análisis conjunto de dos variables, estudio bidimensional, utilizando las tablas de doble entrada para estudiar la relación entre ellas y analizando cada una de las variables por separado desde las tablas, obteniendo así las distribuciones que ahora llamaremos marginales.

Hay parejas de variables que, aunque no puedan relacionarse por medio de una fórmula, sí que hay entre ellas una determinada relación estadística. La visualización por medio de las nubes de puntos nos permitirá hacernos una idea razonable sobre esta relación entre las variables.

Una buena forma de marcar las tendencias de las nubes de puntos es haciendo uso de unas rectas que llamaremos rectas de regresión, o ajustando otras funciones, como las cuadráticas o de segundo grado.

Cuando la correlación es fuerte, los puntos están muy próximos a la recta. En estos casos la recta de regresión resultará muy útil para hacer previsiones, conociendo un valor de una variable podremos calcular el de la otra con razonable seguridad.

# 1. ESTADÍSTICA DESCRIPTIVA UNIDIMENSIONAL

Ya conoces mucho sobre Estadística, recuento de datos, tablas y gráficas, parámetros como media, mediana, moda…. Ahora vamos a revisar estos conocimientos.

## 1.1. 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 $X$, $Y$ .....

## 1.2. 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.

## 1.3. 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 columna 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.

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 $\boldsymbol{l_i}$, $\boldsymbol{L_i}$ respectivamente.
- *Marca de clase* ($\boldsymbol{c_i}$) 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* ($\boldsymbol{a_i}$) es la diferencia entre el extremo superior e inferior: $\boldsymbol{a_i = L_i - l_i}$.
- Al número de observaciones de una clase se le llama *frecuencia de clase* ($\boldsymbol{n_i}$). Si dividimos esta frecuencia por el número total de observaciones, se obtiene la *frecuencia relativa de clase* ($\boldsymbol{f_i}$), y del mismo modo que lo hacíamos para datos sin agrupar definimos ($\boldsymbol{N_i}$) y ($\boldsymbol{F_i}$).

## 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 $\boldsymbol{a_i} = \boldsymbol{a} =$ 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).

***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

**1.** 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 | | | | |

**2.** 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

**Tipos de gráficos estadísticos. Tipos de gráficos estadísticos, veremos los gráficos estadísticos más usados y una breve explicación de sus características y cuándo se deben utilizar: Diagrama de barras, histograma, polígono de frecuencias y gráfico circular. Profe Alex.**

https://www.youtube.com/watch?v=9G4HPNVA5w4&list=PLeySRPnY35dHPj-BOGu_fPKa30L61ITHF

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:

$n$ _____ 360° o 1 _____ 360°

$n_i$ _____ $ángulo_i$ $f_i$ _____ $ángulo_i$

*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.

| $x_i$ | $f_i$ |
|---|---|
| A favor | 0.5 |
| En contra | 0.3 |
| Abstención | 0.2 |

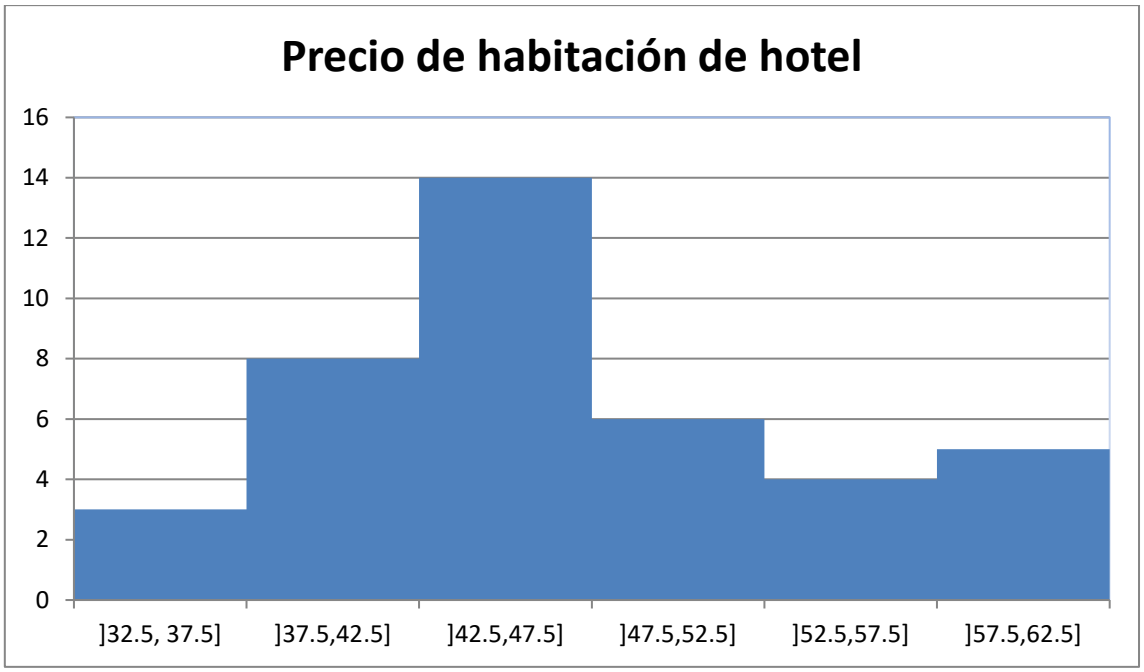

**Crear GRÁFICOS estadísticos. Introducir datos en Excel y crear gráficas. Aprende cómo hacer gráficos en Excel a partir de tus datos. Solo necesitas organizar tus datos en una tabla para después crear un gráfico de una o varias series. Saber Programas**

https://www.youtube.com/watch?v=04pGYGNxRZY

## 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.

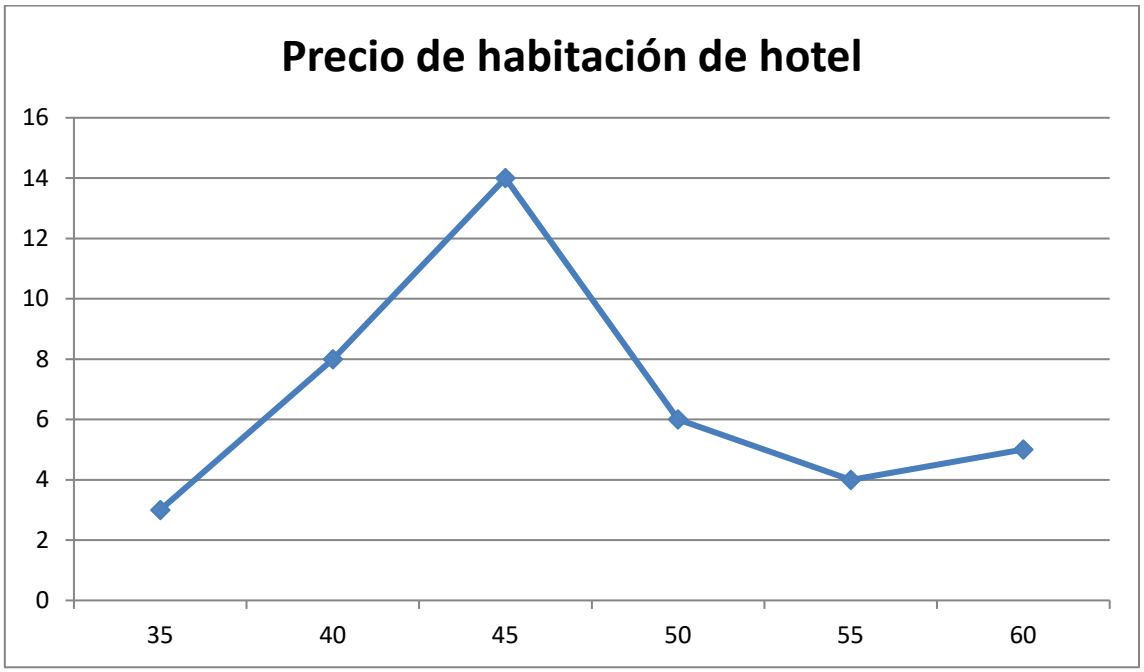
En cualquier caso, observa que, en un histograma el área de los rectángulos es proporcional a la frecuencia representada.




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).

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

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.

$$\bar{x} = \frac{0 \cdot 2 + 1 \cdot 4 + 2 \cdot 21 + 3 \cdot 15 + 4 \cdot 6 + 5 \cdot 1 + 6 \cdot 1}{50} = \frac{126}{50} = 2.52 \text{ hijos.}$$

Utilizando los datos de las frecuencias relativas.

$$\bar{x} = 0 \cdot 0.04 + 1 \cdot 0.08 + 2 \cdot 0.42 + 3 \cdot 0.043 + 4 \cdot 0.12 + 5 \cdot 0.02 + 6 \cdot 0.02 = 2.52 \text{ hijos.}$$

***Ejemplo 2.***

- Precio medio.

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

$$\bar{x} = \frac{35 \cdot 3 + 40 \cdot 8 + 45 \cdot 14 + 50 \cdot 6 + 55 \cdot 4 + 60 \cdot 5}{40} = \frac{1\,875}{40} = 46.875 \text{ €}$$

O equivalentemente:

$$\bar{x} = 35 \cdot 0.075 + 40 \cdot 0.2 + 45 \cdot 0.35 + 50 \cdot 0.15 + 55 \cdot 0.1 + 60 \cdot 0.125 = 46.875 \text{ €.}$$

***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 $\bar{y} = a + b\bar{x}$
4. La suma de todos los valores de la variable restándoles la media es cero.

## 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 4:***

- Si tenemos los datos de 30 valores sobre el peso de los estudiantes de 1º de bachillerato 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 y 68.34

$$\text{Mediana} = Me = (68.26 + 68.34)/2 = 68.3 \text{ kg}.$$

***Ejemplo 5:***

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

$$0\ 1\ 2\ 2\ 2\ 2\ 2\ 2\ 3\ 3\ 3\ 3.$$

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 5:***

- Para la variable consumo de chocolatinas del ejemplo 5 la moda es $Mo = 2$

***Ejemplo 2:***

- Para los datos del ejemplo 2 es el intervalo [42.5, 47.5).

## 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 5, 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 sería el dato 52.35. También es posible aproximarlo mejor mediante una interpolación lineal.

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

## Cuartiles

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

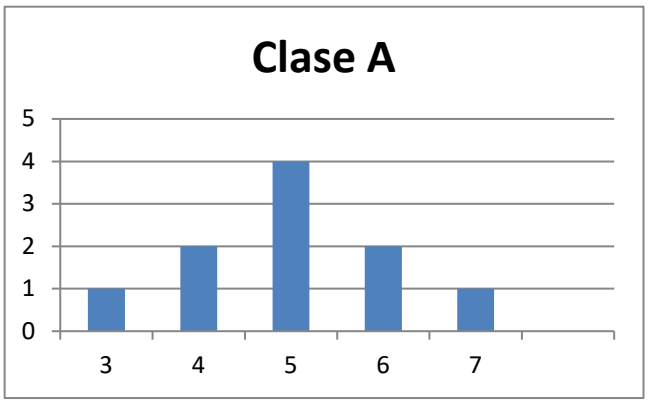
Las medidas de posición estudiadas en el apartado anterior nos dan una información incompleta, por parcial, acerca de los datos.

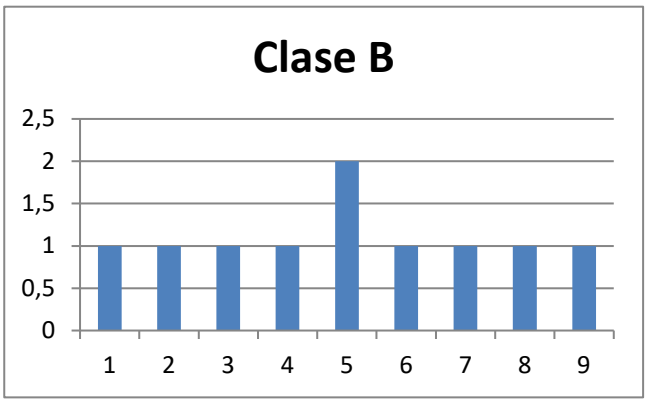
***Veamos un ejemplo:***

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.

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 respecto de 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 | $x_i - \bar{x}$ | $d_i^2$ |
| 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 | $x_i - \bar{x}$ | $d_i^2$ |
| 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 |

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,

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

***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.

## Actividades propuestas

**3.** 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.

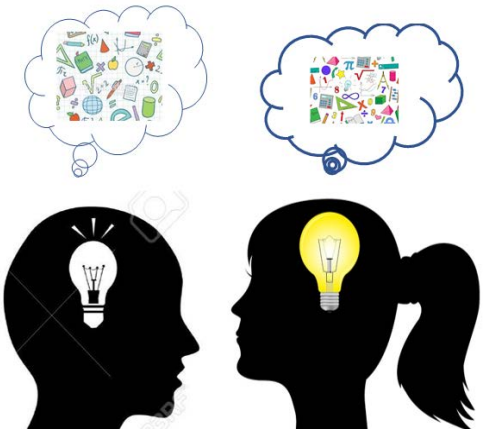
**4.** Muchas personas que invierten en bolsa lo hacen para conseguir beneficios rápidos, por ello el tiempo que mantienen las acciones es relativamente breve. Preguntada una muestra de 40 inversores habituales sobre el tiempo en meses que han mantenido sus últimas inversiones se recogieron los siguientes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

**5.** 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).

**6.** 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.

**7.** 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.

**8.** 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

**9.** 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.

**10.** Dada la distribución de datos 38 432 384 343 38 436 38 438 38 440 con frecuencias 4, 8, 4, 3, 8, halla la media de la distribución.

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

a) El salario medio por trabajador (marca de clase del último intervalo 20 000)

b) El salario más frecuente.

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

| $[l_i, L_i[$ | $n_i$ |
|---|---|
| [0,1 500[ | 2 145 |
| [1 500, 2 000[ | 1 520 |
| [2 000, 2 500[ | 840 |
| [2 500, 3 000[ | 955 |
| [3 000, 3 500[ | 1 110 |
| [3 500, 4 000[ | 2 342 |
| [4 000, 5 000[ | 610 |
| [5 000, 10 000[ | 328 |
| $\geq$ 10 000 | 150 |

**12.** 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 |

**13.** 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.

**14.** 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.

**15.** 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.

**16.** 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.

**17.** 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.

# 2. ESTADÍSTICA BIDIMENSIONAL

## 2.1. Introducción. Tablas de contingencia

***Ejemplo 1:***

- Con el fin de hacer un estudio de aceptación sobre dos modelos de impresoras 3D de reciente fabricación, se consideraron el número de ventas efectuado por un determinado distribuidor durante 25 días.

  Modelo A: 0 2 2 2 1 3 3 3 3 4 4 2 3 3 3 3 2 3 2 4 2 2 3 3 3

  Modelo B: 2 1 2 2 3 1 1 1 2 0 1 1 1 1 1 2 2 1 1 1 2 2 2 2 1

En muchos procesos de la vida se hace necesario estudiar simultáneamente dos características, dos variables. Su estudio conjunto permite determinar las relaciones entre ellas. Supondremos inicialmente que estamos observando dos variables, aunque el tratamiento que se presenta se generaliza sin dificultad a cualquier número de variables.

***Notación.***

Continuando con el ejemplo vamos a llamar:

- $X$ número de impresoras del modelo A vendidas en un día.
- $Y$ número de impresoras del modelo B vendidas en un día.
- $n$ número de pares de observaciones.
- $x_i$ cada dato diferente observado en la muestra de $X$.
- $K$ número de valores distintos de $X$.
- $y_j$ cada dato diferente observado en la muestra de $Y$.
- $h$ número de valores distintos de $Y$.

***Ejemplo 2: Tabla de contingencia***

- Se estudiaron 600 enfermos con lesiones de hígado mediante un procedimiento gráfico, y se comprobaron los resultados mediante un procedimiento histológico para conocer si había sido el diagnóstico correcto. Los datos que se obtuvieron fueron:

| | **Diagnóstico correcto ($C$)** | **Diagnóstico incorrecto ($I$)** | **Total** |
|---|---|---|---|
| **Lesión maligna ($M$)** | 210 | 20 | 230 |
| **Lesión benigna ($B$)** | 320 | 50 | 370 |
| **Total** | 530 | 70 | 600 |

Ahora los datos aparecen en una tabla de contingencia. En este ejemplo los pares observados aparecen ya agrupados en la tabla, siendo $n$ = 600. Las variables no son cuantitativas sino cualitativas, siendo $X$ el tipo de lesión e $Y$ el tipo de diagnóstico.

## 2.2. Distribución de frecuencias conjuntas

Cuando queremos describir conjuntamente dos variables, el primer paso al igual que en el caso univariante, será la representación de los datos en una tabla de frecuencias.

**Frecuencia absoluta conjunta ($n_{ij}$)**

Número de veces que se presenta en la muestra el valor $x_i$ de la variable $X$ con el valor $y_j$ de la variable $Y$.

***Ejemplo 1:***

- Para el par de valores $x_1 = 0$, $y_3 = 2$, $n_{13} = 1$

***Ejemplo 2:***

- Para lesión benigna ($B$), diagnóstico correcto ($C$), $n_{BC} = 320$.

***Propiedad:***

La suma de las frecuencias absolutas es igual a $n$.

**Frecuencia relativa conjunta**

$$f_{ij} = \frac{n_{ij}}{n}$$

***Ejemplo 1:***

$$f_{13} = \frac{1}{25} = 0.04$$

***Ejemplo 2:***

- Para lesión benigna ($B$), diagnóstico correcto ($C$), $f_{BC} = 320/600$.

***Propiedad***

La suma de las frecuencias relativas es igual a la unidad.

**Tabla de frecuencias conjunta**

Llamamos así a una tabla de doble entrada donde se representan en la primera columna los diferentes valores observados para la variable $X$ ordenados de menor a mayor y en la primera fila los diferentes valores observados para la variable $Y$, y en el centro de la tabla sus correspondientes frecuencias conjuntas, tanto absolutas como relativas.

*Ejemplo 1:*

| $x_i$ / $y_j$ | 0 | 1 | 2 | 3 | $n_i$ | $f_i$ |
|---|---|---|---|---|---|---|
| 0 | 0/0 | 0/0 | 1/0.04 | 0/0 | 1 | 0.04 |
| 1 | 0/0 | 0/0 | 0/0 | 1/0.04 | 1 | 0.04 |
| 2 | 0/0 | 3/0.12 | 5/0.20 | 0/0 | 8 | 0.32 |
| 3 | 0/0 | 8/0.32 | 4/0.16 | 0/0 | 12 | 0.48 |
| 4 | 1/0.04 | 2/0.08 | 0/0 | 0/0 | 3 | 0.12 |
| $n_j$ | 1 | 13 | 10 | 1 | 25 | |
| $f_j$ | 0.04 | 0.52 | 0.04 | 0.04 | | 1 |

- ¿Qué porcentaje de días venderemos una impresora del modelo A y 3 del modelo B?: 4 %
- ¿Qué porcentaje de días venderemos más impresoras del modelo B que del modelo A?

8 %; 0.04 + 0.04

*NOTA:*

En el caso en que las variables sean cualitativas la tabla de distribución conjunta también recibe el nombre de tabla de contingencia.

*Más ejemplos de tablas de contingencia.*

**1.-** Se quiere estudiar el efecto de tres fármacos en el tratamiento de una enfermedad infecciosa. Para ello se dispone de un grupo de pacientes infectados, distribuyéndose al azar en tres grupos de tratamiento.

| | Tratamiento A | Tratamiento B | Tratamiento C | Total |
|---|---|---|---|---|
| **Si mejora** | 23 | 33 | 35 | 91 |
| **No mejora** | 12 | 7 | 12 | 31 |
| **Total** | 35 | 40 | 47 | 122 |

**2.-** En un estudio se ha aplicado durante un año una terapia basada en la ejercitación mental para frenar el deterioro cognitivo observado en 3 enfermedades degenerativas, en la tercera edad. Para evaluar el grado en que la terapia es efectiva, se han registrado los resultados observados al cabo de un año de tratamiento en cada tipo de enfermedad, teniendo en cuenta que la evolución natural al cabo de un año, de estas enfermedades, es el empeoramiento.

| | Empeora | Estable | Mejora | Total |
|---|---|---|---|---|
| **Parkinson senil** | 34 | 25 | 17 | 76 |
| **Alzheimer** | 47 | 18 | 6 | 71 |
| **Demencia vascular** | 50 | 23 | 2 | 75 |
| **Total** | 131 | 66 | 25 | 222 |

## 2.3. Distribución de frecuencias marginales

Para distinguir las frecuencias de cada variable al estudiarlas aisladamente llamaremos frecuencias marginales a las de cada variable por separado. De esta forma tendríamos dos distribuciones unidimensionales a partir de las conjuntas.

### Frecuencia absoluta marginal

Para la $X$ ($x_i$) sería el número de veces que se repite el valor $x_i$ sin tener en cuenta los valores de $Y$, la representamos por $n_i$.

Para la $Y$ ($y_j$) sería el número de veces que se repite el valor $y_j$ sin tener en cuenta los valores de la $X$, la representamos por $n_j$.

***Nota:***

**1.-**Con las definiciones de media, desviación típica y varianza del apartado de distribuciones unidimensionales, utilizando para la $X$ los valores $x_i$ y el número de veces que se repite $n_i$ y $N$ el número total de pares observados, y para la $Y$ los valores $y_j$ y el número de veces que se repite $n_j$ y $N$ el número total de pares observados, calcularemos las medias marginales, desviaciones típicas marginales y varianzas marginales.

**2.-** Si nos fijamos bien podemos relacionar el nombre de frecuencias marginales con el hecho de que tanto los valores de las variables, $x_i$ e $y_j$ como las veces que aparece cada uno de estos datos, $n_i$ y $n_j$ los encontramos en los márgenes de la tabla de distribución conjunta.

### Frecuencias relativas marginales

A partir de las anteriores, y del mismo modo, se construirán estas frecuencias $f_i$ y $f_j$.

La distribución de frecuencias marginales puede colocarse en una tabla separadamente. Pero si deseamos tener toda la información en una misma tabla lo que se suele hacer es colocar:

- En la última columna de la tabla conjunta, las frecuencias marginales de $X$, es decir, $n_i$, añadiendo tantas columnas como otros tipos de frecuencias marginales se desee añadir.

- En la última fila de la tabla conjunta, las frecuencias marginales de $Y$, es decir, $n_j$ añadiendo tantas filas como otros tipos de frecuencias marginales se desee añadir.

## 2.4. Distribución de frecuencias condicionadas

A partir de la distribución de frecuencias conjuntas podemos definir otro tipo de distribuciones unidimensionales, tanto para $X$ como para $Y$. Estas distribuciones se obtendrán al fijar el valor de la otra variable y reciben el nombre de distribuciones condicionadas.

**Frecuencia absoluta condicionada** para $X$ ($x_i$) dado que $Y$ ($y_j$) es el número de veces que se repite el valor $x_i$ teniendo en cuenta solo aquellos valores en que $Y$ ($y_j$); así es $n_{i(j)} = n_{ij}$ para todo $i = 1, 2, \ldots, k$.

**Frecuencia absoluta condicionada** para $Y$ ($y_j$) dado que $X$ ($x_i$) es el número de veces que se repite el valor $y_j$ teniendo en cuenta solo aquellos valores en que $X$ ($x_i$); así es $n_{(i)j} = n_{ij}$ para todo $j = 1, 2, \ldots, h$.

En las distribuciones condicionadas no se suelen utilizar las distribuciones absolutas, puesto que como sabemos, estas dependen del número de datos y el número de datos será diferente para cada distribución, pues dependerá de la frecuencia del valor que fijamos de la otra variable. Son mucho más útiles las frecuencias condicionadas que se definen:

**Frecuencia relativa condicionada** para $X$ dado que $Y = y_j$ es

$$f_{i(j)} = \frac{n_{ij}}{n_j}$$

**Frecuencia relativa condicionada** para $Y$ dado que $X = x_i$ es

$$f_{(i)j} = \frac{n_{ij}}{n_i}$$

***Ejemplo:***

- Distribución de frecuencias de $X$ condicionada a $Y = 1$

| $x_i$ | $n_{i(2)}$ | $f_{i(2)}$ |
|---|---|---|
| 0 | 0 | 0 |
| 1 | 0 | 0 |
| 2 | 3 | 0.23 |
| 3 | 8 | 0.62 |
| 4 | 2 | 0.15 |

***Nota:***

Si la tabla resulta muy grande deberemos agrupar una o las dos variables en intervalos de clase del mismo modo que lo hacíamos en el apartado de una variable. En este caso todas las definiciones se aplican tal como las hemos visto en dicho apartado.

## 2.5. Independencia estadística

***Definición 1:***

Dos variables $X$ e $Y$ se dice que son independientes estadísticamente cuando la frecuencia relativa conjunta es igual al producto de las frecuencias relativas marginales, es decir, para todo $i$, $j$:

$$f_{ij} = \frac{n_{ij}}{n} = f_i \cdot f_j = \frac{n_i}{n} \cdot \frac{n_j}{n}$$

***Definición 2:***

Dos variables $X$ e $Y$ se dicen que son independientes estadísticamente cuando todas las frecuencias relativas condicionadas son iguales a sus correspondientes frecuencias marginales, es decir:

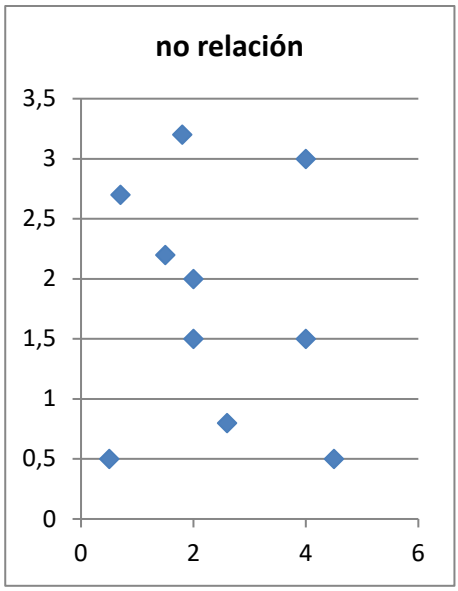
$f_{i(j)} = f_i$ para todo $j$ y $f_{(i)j} = f_j$ para todo $i$.

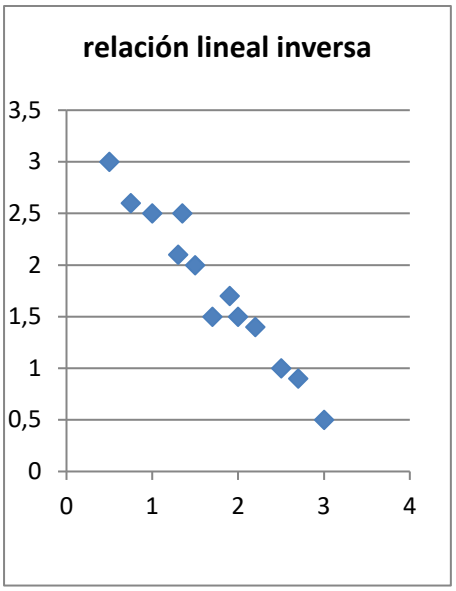
## 2.6. Diagrama de dispersión. Nube de puntos

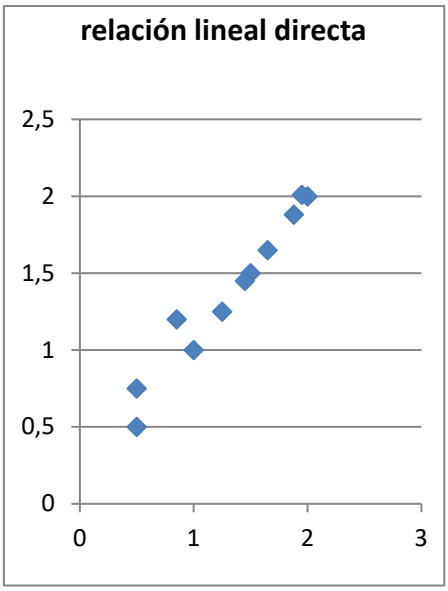

**Nube de puntos. Píldoras Matemáticas**

https://www.youtube.com/watch?v=swqmjtwJr-E

Se obtiene representando cada par observado $(x_i, y_j)$, como un punto del plano cartesiano. Se utiliza con los datos sin agrupar y sobre todo para variables continuas. Si los datos están agrupados se toman las marcas de clase. Es más útil porque nos permite ver visualmente la relación entre las dos variables.



# 3. COVARIANZA

## 3.1. Idea correlación. Covarianza

Al analizar dos variables cuantitativas de forma conjunta, el objetivo que se pretende es, por lo general, determinar si existe o no algún tipo de variación conjunta o covarianza entre ellas: si una variable aumenta, la otra también o lo contrario.

La cantidad se denomina covarianza $S_{xy}$ y tiene la siguiente expresión:

$$S_{xy} = \frac{\sum_i \sum_j (x_i - \bar{x}) \cdot (y_i - \bar{y}) \cdot n_{ij}}{n} = \frac{\sum_i \sum_j x_i \cdot y_i \cdot n_{ij}}{n} - \bar{x} \cdot \bar{y}$$

Ayuda a analizar la covarianza entre dos variables de la forma siguiente:

- Cuando el resultado es positivo, hay una tendencia a que a mayores observaciones de $X$ correspondan mayores observaciones de $Y$.

***Por ejemplo***

- A mayor cantidad de agua de lluvia en un año, suele corresponder una mejor cosecha.
  - Cuando el resultado es negativo, la tendencia resulta contraria; es decir a mayores valores de la variable $X$ solemos encontrar menores valores de la variable $Y$.

***Por ejemplo***

- A mayor renta per cápita en los países suele encontrarse una menor mortalidad infantil.

## 3.2. Coeficiente correlación lineal


**Que es correlación? La correlación sirve para medir la relación que existe entre dos variables. Esta relación puede ser lineal o no lineal. Si es lineal esta podría tener una relación positiva o negativa. Además, la relación positiva o negativa podría tener una relación fuerte, débil o moderada.**

https://www.youtube.com/watch?v=n9odGHfgL8s

El valor de la covarianza dependerá de los valores de las variables, por tanto, de sus unidades. Para poder eliminar las unidades y tener una medida adimensional utilizamos el coeficiente de correlación $r_{xy}$:

$$r_{xy} = \frac{S_{xy}}{s_x \cdot s_y}$$

Siendo también invariante frente a trasformaciones lineales (cambio de origen y escala) de las variables.

Citamos las siguientes propiedades:

- Es un coeficiente adimensional.
- Toma valores entre $-1$ y 1.
- Si hay relación lineal positiva el valor será positivo y próximo a 1.
- Si hay relación lineal negativa el valor será negativo y próximo a $-1$.
- Si no hay relación el valor se aproxima a cero.
- Si $X$ e $Y$ son independiente el valor del coeficiente es cero. Pero no al contrario. Puede ocurrir que el coeficiente de correlación valga cero y las variables sean dependientes.

## 3.3. Recta regresión lineal. Regresión cuadrática

El diagrama de dispersión o nube de puntos nos permitía visualizar la relación entre dos variables $X$ e $Y$. Al representar el diagrama de dispersión podemos encontrar las siguientes situaciones:

- Distribuciones estadísticas para las que la nube de puntos se dispone de tal forma que existe una función matemática cuyos puntos son una parte de su representación gráfica.
- Sin coincidir sus puntos con los de una gráfica de una función matemática, se aproximan a ella con mayor o menor intensidad.
- La nube de puntos presenta un aspecto tal que no existe concentración de puntos hacia ninguna grafica matemática, distribuyéndose de una forma uniforme en una región del plano.

En el primer caso se dice que existe una **dependencia funcional** o exacta entre las variables $X$ e $Y$, es decir existe una función matemática tal que $y = f(x)$. En el segundo caso se dice que existe una **dependencia estadística** o aproximada entre las dos variables, $Y$ aproxima $f(x)$. Y en el último caso decimos que las variables son **independientes**.

Es el segundo caso del que se ocupa la teoría de regresión.

Las técnicas de regresión tienen por objeto **modelar**, es decir, encontrar una función que aproxime lo máximo posible la relación de dependencia estadística entre variables y predecir los valores de una de ellas: $Y$ (variable dependiente o explicada) a partir de los valores de la otra (u otras): $X$ (variable independiente o explicativa).

Llamamos regresión $Y$ sobre $X$ a la función que explica la variable $Y$ (dependiente) para cada valor de la $X$ (independiente). Los dos tipos de funciones más usados son la **regresión lineal**, si la función de regresión es una recta, y la **regresión cuadrática**, si la función es una parábola.

La recta de regresión que estudiamos es una función lineal por que el modelo de función de regresión seleccionado es una recta.

Recta de regresión $Y$ sobre $X$ es $y = a + bx$ donde $a = \bar{y} - b\bar{x}$ y $b = \frac{S_{xy}}{{s_x}^2}$.

Recta de regresión de $X$ sobre $Y$ es $x = a' + b'y$ donde $a' = \bar{x} - b'\bar{y}$ y $b' = \frac{S_{xy}}{{s_y}^2}$.

Los valores de $b$ y $b'$ son los correspondientes coeficientes de regresión para cada una de las rectas.

Hay que tener en cuenta que la recta de regresión de $x$ sobre $y$ no se obtiene despejando $x$ de la recta de regresión de $y$ sobre $x$.

Si usamos como función para el modelo una parábola tenemos una regresión cuadrática, buscando los valores de los coeficientes de una parábola que mejor se ajusten a unos datos.

En ambos casos se usa el "Método de Mínimos cuadrados".

Pero las herramientas informáticas, como Excel, te pueden ayudar. Basta con que escribas los datos en una hoja de cálculo, le pidas a Excel que dibuje una nube de puntos, y luego, puedes ajustar esa nube con una recta, una parábola….

## Regresión cuadrática

### Actividades resueltas

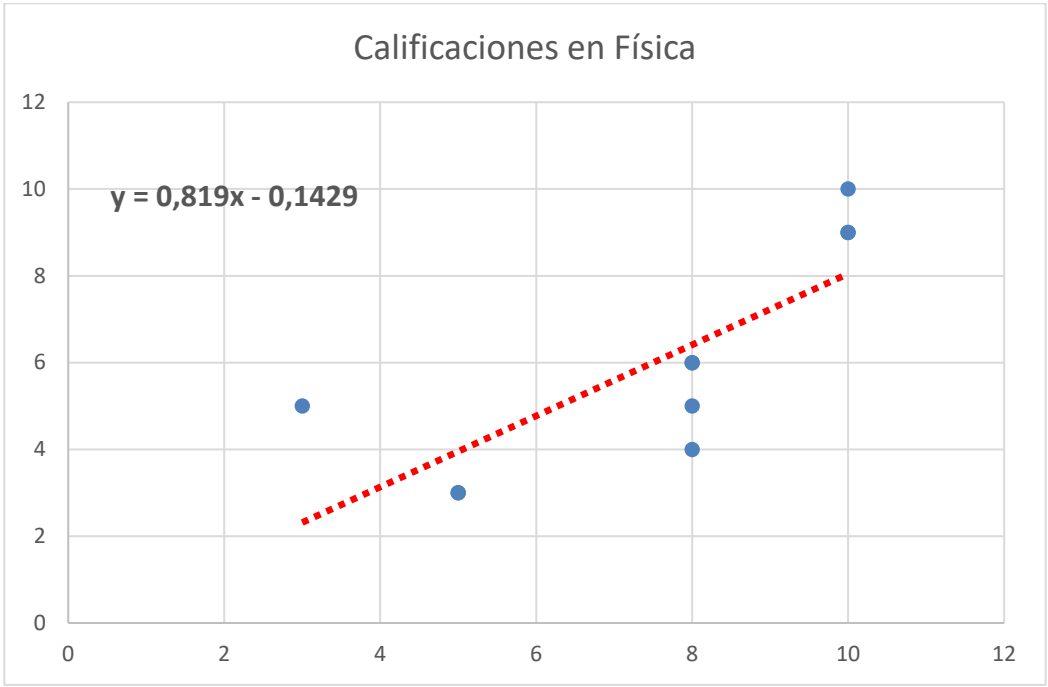
Vamos a trabajar con una hoja de cálculo para ajustar diferentes funciones a una nube de puntos.

- *Preguntamos a 10 estudiantes de 1º de Bachillerato por sus calificaciones en Matemáticas y por sus calificaciones en Física.*

| **Calificaciones de Matemáticas** | 10 | 3 | 8 | 8 | 5 | 10 | 10 | 8 | 5 | 8 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Calificaciones en Física** | 9 | 5 | 6 | 4 | 3 | 10 | 9 | 6 | 3 | 5 |

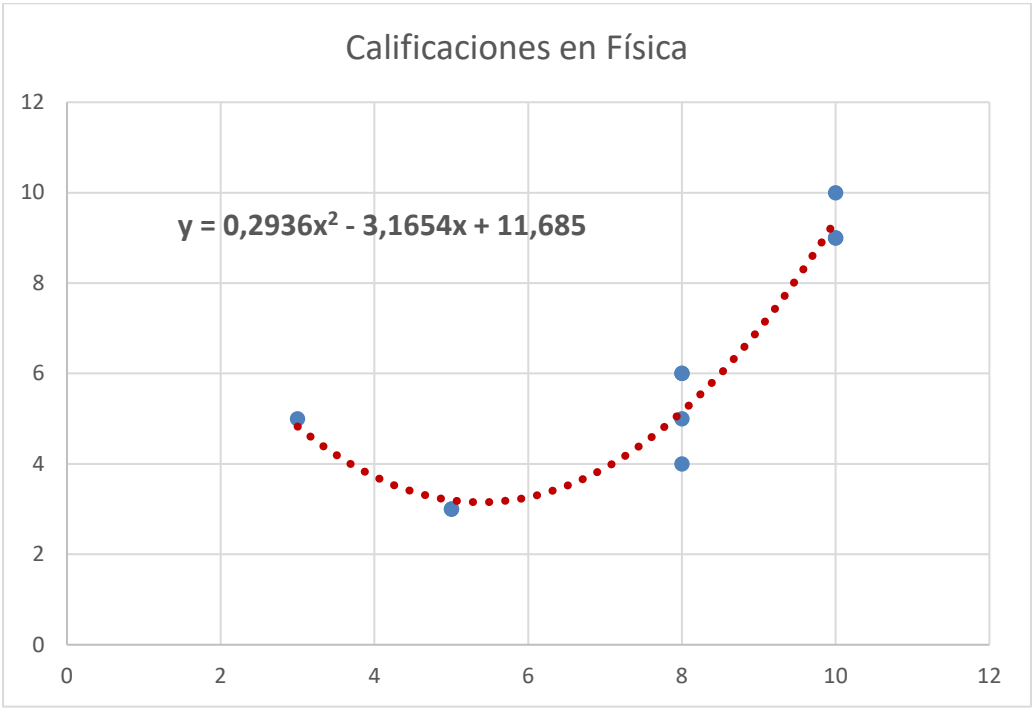
*Queremos estudiar la relación entre las calificaciones de Matemáticas y las de Física.*

- Abrimos una hoja de cálculo y llevamos los datos de Calificaciones en Matemáticas y Calificaciones en Física.
- En INSERTAR seleccionamos insertar un gráfico de "Dispersión" con forma de nube de puntos.
- Pinchando con el ratón en uno de los puntos de la nube, elegimos "Agregar línea de tendencia…"
- Nos aparecen varios opciones: OPCIONES DE LINEA DE TENDENCIA: Exponencial, Lineal, Logarítmica, Polinómica, Potencial…
- Elegimos: Lineal. Y aparece dibujada una recta.
- Hay otras opciones: Interpolar hacia adelante o hacia atrás. Señalar intersección. Presentar ecuación en el gráfico. Presentar el valor de R cuadrado en el gráfico.
- Elegimos "Presentar ecuación en el gráfico" y "Presentar el valor de R cuadrado en el gráfico"



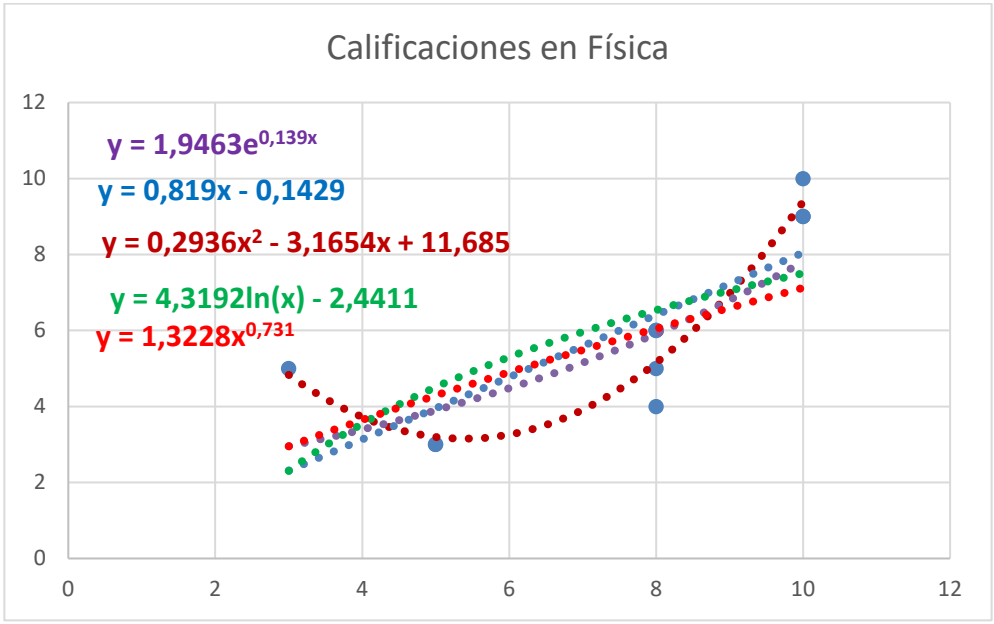

Observamos que la recta de regresión es $y = 0.819x - 0.1429$.

- Volvemos a pinchar sobre un punto, y elegimos ahora: Polinómica y en Orden, seleccionamos 2, para tener una ecuación polinómica de segundo grado, una parábola




El modelo de regresión cuadrática nos dice que se aproxima a la parábola $y = 0.29x^2 - 3.165x + 11.685$.

- Vamos a dibujar en un mismo gráfico todas las líneas de tendencia que nos propone Excel, Cada ecuación y su línea correspondiente tienen el mismo color.




Son: Exponencial, Lineal, Cuadrática, Logarítmica, Potencial.

¿Cuál crees que nos interesa más? La que discrimina más en este caso es la regresión cuadrática.

## 3.4. Predicción y causalidad

El objetivo último de la recta de regresión es la predicción de una variable para un valor determinado de la otra. La predicción de $Y$ para $X = x_0$, será simplemente el valor obtenido en la recta de regresión de $Y$ sobre $X$ al sustituir el valor de $x$ por $x_0$.

Es claro que la fiabilidad de esta predicción será tanto mayor cuanto mayor sea la correlación entre las variables, es decir mayor sea el valor de $r_{xy}$.

## 3.5. Coeficiente de determinación

El **coeficiente de determinación** representa el porcentaje de la variabilidad de la variable $Y$ explicada por la recta de regresión, es decir por su relación con la variable $X$.

El coeficiente determinación es el cuadrado del coeficiente de correlación:

$$R^2 = r_{xy}^2$$

El coeficiente determinación complementa el coeficiente de correlación para evaluar una predicción hecha mediante la recta de regresión. Refleja el porcentaje de variabilidad de los datos que es capaz de explicar la recta de regresión.

$$0 \leq R^2 \leq 1$$

Si $R^2 = 1$ el ajuste es perfecto, si $R^2 = 0$ el ajuste es inadecuado.

***Nota aclaratoria:***

Un coeficiente de correlación $r = 0.75$ podría indicar que existe una dependencia lineal apreciable entre ambas variables. Sin embargo, $R^2 = r^2 = 0.5626$ que refleja que la recta de regresión sólo es capaz de explicar un 56 % de variabilidad de los datos y, por tanto, el ajuste mediante una recta no es bueno.

***Ejemplo:***

- Conocida la recta de regresión del gasto en función de la renta $y = 0.026 + 0.57629x$, determinar el gasto para este año si la renta es de 7.56 millones de euros. Dar una medida de la bondad de la predicción. ¿Cuál es el porcentaje de variabilidad en el gasto atribuible a la renta de los consumidores?

Como la renta esta medida en millones de euros, la predicción del gasto será:

$$y = 0.026 + 0.57629 \cdot 7.56 = 4.3827 \text{ millones de euros.}$$

Una medida de la bondad de la predicción nos vendrá dada por el coeficiente de correlación lineal entre las variables:

$$r = \frac{s_{xy}}{s_x s_y} = \frac{0.193}{\sqrt{0.112} \cdot \sqrt{0.3349}} = 0.99 \text{ predicción muy fiable.}$$

El porcentaje de variabilidad en el gasto atribuible a la renta de los consumidores nos viene dado por el coeficiente determinación:

$$R^2 = r_{xy}^2 = 0.99^2 = 0.98$$

## Actividades propuestas

**18.** Los datos siguientes son las calificaciones obtenidas por los estudiantes de un grupo de 25 de 1º de bachillerato en las asignaturas de Matemáticas y Lengua.

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Matemáticas** | 4 | 5 | 5 | 6 | 7 | 7 | 7 | 7 | 7 | 7 | 8 | 8 |
| **Lengua** | 3 | 5 | 6 | 7 | 7 | 7 | 7 | 8 | 8 | 8 | 7 | 7 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Matemáticas** | 8 | 8 | 8 | 8 | 9 | 9 | 9 | 9 | 9 | 10 | 9 | 8 |
| **Lengua** | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 10 | 10 | 10 | 9 | 9 |

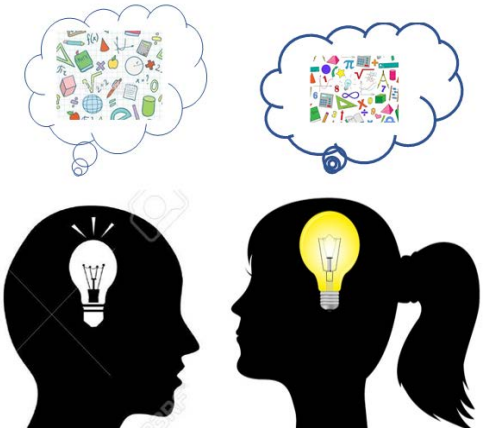
a) Escribe la tabla de frecuencias conjunta.

b) Proporción de estudiantes que obtiene más de un cinco en ambas asignaturas, proporción de estudiantes que obtiene más de un cinco en Matemáticas, proporción estudiantes que obtiene más de un cinco en Lengua.

c) ¿Son independientes las calificaciones de Matemáticas y Lengua?

d) Representa gráficamente.

e) Calcula el coeficiente correlación.



**19.** Para realizar un estudio sobre la utilización de una impresora en un determinado departamento, se midió en un día los minutos transcurridos entre las sucesivas utilizaciones $X$ y el número de páginas impresas $Y$, obteniéndose los siguientes resultados.

| | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $X$ | 9 | 9 | 4 | 6 | 8 | 9 | 7 | 6 | 9 | 9 | 9 | 9 | 9 | 10 | 9 | 15 | 10 | 12 | 12 | 10 | 10 | 12 | 10 | 10 | 12 | 12 |
| $Y$ | 3 | 8 | 3 | 8 | 3 | 8 | 8 | 8 | 3 | 8 | 12 | 12 | 20 | 8 | 20 | 8 | 8 | 20 | 8 | 8 | 12 | 8 | 20 | 20 | 3 | 3 |

a) Escribe la distribución de frecuencias conjunta. Porcentaje de veces que transcurren más de nueve minutos desde la anterior utilización y se imprimen menos de doce páginas. Número de veces que se imprimen menos de doce páginas y transcurren nueve minutos desde la utilización anterior.

b) Frecuencias marginales. Veces que se imprimen como mucho doce páginas. Número de páginas que se imprimen en el 80 % de las ocasiones.

c) Calcula la distribución del número de páginas impresas condicionada a que han transcurrido nueve minutos entre sucesivas utilizaciones.

d) Dibuja el diagrama de dispersión.

**20.** Las estaturas de los 30 niños nacidos en una maternidad durante una semana fueron los siguientes:

| **Estatura** | 50 | 51 | 53 | 50 | 51 | 48 | 50 | 49 | 52 | 52 | 49 | 50 | 52 | 51 | 52 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Peso** | 3.2 | 4.1 | 4.5 | 3.0 | 3.6 | 2.9 | 3.8 | 3.8 | 3.6 | 3.9 | 3.0 | 3.8 | 4.1 | 3.5 | 4.0 |

| 49 | 50 | 51 | 52 | 53 | 52 | 52 | 51 | 50 | 51 | 54 | 50 | 51 | 51 | 51 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3.1 | 3.3 | 3.9 | 3.7 | 4.1 | 4.2 | 3.5 | 3.8 | 3.6 | 3.4 | 4.6 | 3.5 | 3.6 | 3.1 | 4.0 |

a) Construye una tabla de doble entrada, agrupando los pesos en intervalos de 0.5 kg.

b) ¿Es la estatura independiente del peso?

**21.** En el examen de una asignatura que consta de parte teórica y parte práctica, las calificaciones de nueve alumnos fueron:

| **Teoría** | 5 | 7 | 6 | 9 | 3 | 1 | 2 | 4 | 6 |
|---|---|---|---|---|---|---|---|---|---|
| **Práctica** | 6 | 5 | 8 | 6 | 4 | 2 | 1 | 3 | 7 |

Calcula la covarianza y el coeficiente de correlación lineal. Dibuja la nube de puntos. Comenta los resultados.

**22.** Se desea investigar el ganado caprino y el ganado ovino de un país. En la tabla de doble entrada adjunta se presentan los resultados de un estudio de 100 explotaciones ganaderas, seleccionadas aleatoriamente del censo agropecuario. Se proporcionan las frecuencias conjuntas del número de cabezas (en miles) de cabras $X$ y ovejas $Y$ que poseen las explotaciones.

| $X$ / $Y$ | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| 0 | 4 | 6 | 9 | 4 | 1 |
| 1 | 5 | 10 | 7 | 4 | 2 |
| 2 | 7 | 8 | 5 | 3 | 1 |
| 3 | 5 | 5 | 3 | 2 | 1 |
| 4 | 2 | 3 | 2 | 1 | 0 |

a) Halla las medias, varianzas y desviaciones típicas marginales.

b) Halla el número medio de ovejas condicionado a que en la explotación hay 2000 cabras.

c) Halla el número medio de cabras que tienen aquellas explotaciones que sabemos que no tienen ovejas.

d) Halla la covarianza y el coeficiente de correlación entre ambas variables.

**23.** El volumen de ahorro y la renta del sector familias en millones en euros constantes de 2005 para el periodo 2005-2014 fueron.

| **Años** | 05 | 06 | 07 | 08 | 09 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Ahorro** | 1.9 | 1.8 | 2.0 | 2.1 | 1.9 | 2.0 | 2.2 | 2.3 | 2.7 | 3.0 |
| **Renta** | 20.5 | 20.8 | 21.2 | 21.7 | 22.1 | 22.3 | 22.2 | 22.6 | 23.1 | 23.5 |

a) Recta regresión del ahorro sobre la renta.

b) Recta de regresión de la renta sobre el ahorro.

c) Para el año 2015 se supone que la renta era de 24.1 millones de euros. ¿cuál será el ahorro esperado para el año 2015?

d) Estudiar la fiabilidad de la predicción anterior.

**24.** Se midió el tiempo en segundos que tardaron en grabarse los mismos 24 ficheros en un lápiz USB $X$ y en un disco duro exterior $Y$.

| $X$ | 1.2 | 1 | 1.1 | 0.5 | 1.1 | 1.5 | 1 | 1.4 | 1.4 | 1.3 | 0.4 | 0.3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $Y$ | 1.3 | 1.1 | 1.2 | 0.4 | 1.2 | 1.4 | 1.1 | 1.6 | 1.6 | 1.5 | 0.4 | 0.3 |

| $X$ | 0.3 | 1.5 | 1.4 | 1.1 | 1.2 | 1.2 | 0.4 | 0.5 | 1.3 | 1.5 | 1.2 | 0.2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $Y$ | 0.3 | 1.6 | 1.3 | 1.1 | 1.3 | 1.1 | 0.4 | 0.4 | 1.4 | 1.6 | 0.9 | 0.3 |

a) Construye la tabla de frecuencias conjunta. ¿Cuál es el porcentaje de ficheros que tardan menos de 1.5 segundos en el primer tipo y más de 1.4 en el segundo? ¿Cuántos ficheros tardan en grabarse entre 0.6 y 1.2 segundos en el primer tipo de memoria? ¿Cuánto tiempo tardan como mucho en gravarse al menos el 90 % de los ficheros en el segundo tipo de memoria?

b) Halla la tabla de frecuencias condicionadas de los tiempos del segundo tipo de memoria de aquellos programas que tardaron 1.2 en el primer tipo de memoria. ¿Cuál es la proporción de estos programas que tardan en grabarse más de 1.5 segundos en el segundo tipo de memoria?

c) Representa gráficamente los datos y comenta el resultado obtenido.

d) Si un fichero tarda 0.8 segundos en grabarse en el primer tipo de memoria, ¿cuántos segundos tardará en grabarse en el segundo tipo? Dar una medida de fiabilidad. ¿Confirma esta medida lo comentado en el apartado c)?

**25.** De un muelle se cuelgan pesos y obtenemos los alargamientos siguientes.

| **Peso gr $X$** | 0 | 10 | 30 | 60 | 90 | 120 | 150 | 200 | 250 | 350 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Alargamiento cm $Y$** | 0 | 0.5 | 1 | 3 | 5 | 6.5 | 8 | 10.2 | 12.5 | 18 |

Encuentra la recta de regresión de $Y$ sobre $X$ y estima el alargamiento que se conseguirá con pesos de 100 y 500 gr. ¿Cuál de las dos estimaciones es más fiable?

**26.** La tabla siguiente muestra el número de gérmenes patógenos por centímetro cubico de un determinado cultivo según el tiempo transcurrido.

| **Número de horas** | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| **Número de gérmenes** | 20 | 26 | 33 | 41 | 47 | 53 |

a) Calcula la recta de regresión para predecir el número de gérmenes por centímetro cubico en función del tiempo.

b) ¿Qué cantidad de gérmenes por centímetro cubico es previsible encontrar cuando transcurran 6 horas? ¿Es buena esta predicción?

**27.** En un depósito cilíndrico, la altura del agua que contiene varía a medida que pasa el tiempo según los datos recogidos en la tabla:

| **Tiempo: $h$** | 8 | 22 | 27 | 33 | 50 |
|---|---|---|---|---|---|
| **Altura: $m$** | 17 | 14 | 12 | 11 | 6 |

a) Encuentra el coeficiente correlación entre el tiempo y la altura. Da una interpretación de él.

b) ¿Qué altura se alcanzará cuando hayan transcurrido 40 horas?

c) Cuando la altura alcanza 2 m suena una alarma. ¿Cuánto tiempo tiene que pasar para que suene la alarma?

**28.** La evolución del IPC (índice de precios al consumo) y la tasa de inflación en los meses indicados de un determinado año, va a ser:

| | Enero | Febrero | Marzo | Abril | Mayo | Junio |
|---|---|---|---|---|---|---|
| **IPC** | 0.7 | 1.1 | 1.7 | 2 | 1.9 | 1.9 |
| **Tasa inflación** | 6 | 6 | 6.3 | 6.2 | 5.8 | 4.9 |

a) Representa la nube de puntos.

b) Calcula el coeficiente de correlación entre el IPC y la tasa de inflación.

c) ¿Se puede estimar la tasa de inflación a partir del IPC?

## CURIOSIDADES. REVISTA

### 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.

### RELACION FUNCIONAL – CORRELACIÓN

Si lanzamos una piedra hacia arriba llegará más alto cuando más fuerte sea lanzada. Existe una fórmula que nos permite calcular, exactamente la altura conseguida en función de la velocidad con que es lanzada. Estamos ante una relación funcional.

Las personas, en general, pesan más cuando más altos son. Pero no se puede dar una fórmula que nos permita dar el peso de una persona con exactitud conociendo su altura, sólo podremos conseguir una fórmula que nos dé un valor aproximado y conocer la eficacia de esa fórmula. La relación entre las variables peso-estatura es una relación estadística. Diremos que hay una correlación entre estas variables.

También vamos a encontrar correlación entre la distancia a que un jugador de baloncesto se coloca de la cesta y el número de cestas que consigue. Pero en este caso, al contrario del anterior, hay una correlación negativa, ya que a más distancia, menor número de cestas.

## CONTRA LA SUPERSTICIÓN, ESTADÍSTICA

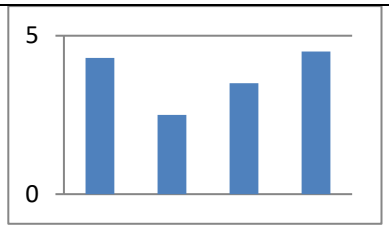
Vivimos en un mundo dominado por la ciencia y la tecnología, a pesar de ello las supersticiones y las creencias seudocientíficas siguen dominando entre la población general, incluso más que en otras épocas. La Estadística es un arma importante para desenmascarar algunas afirmaciones que circulan impunemente 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.

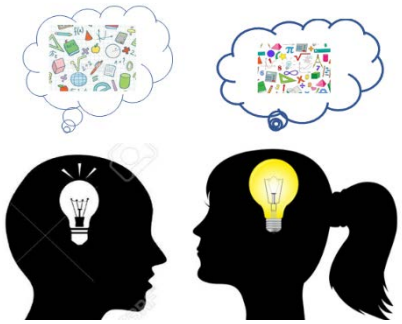
## Estadística

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.

# RESUMEN

| | | |
|---|---|---|
| **Histograma** | Representación gráfica de los datos agrupados en intervalos. |   |
| **Media aritmética** | $\bar{x} = \frac{\sum_i x_i n_i}{n} = \sum_{i=1}^{k} x_i f_i$ | $\bar{x} = \frac{0 \cdot 2 + 1 \cdot 4 + 2 \cdot 21 + 3 \cdot 15 + 4 \cdot 6 + 5 \cdot 1 + 6 \cdot 1}{50}$ $= \frac{126}{50} = 2.52$ |
| **Mediana** | Valor tal que en la distribución hay tantos datos menores que él como mayores que él. | |
| **Moda** | Dato con mayor frecuencia, el que más veces se repite. | |
| **Varianza** | $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = \frac{\sum_{i=1}^{k}(x_i)^2 \cdot f_i}{n} - \bar{x}^2$ | |
| **Desviación típica** | $s = \sqrt{Varianza}$ | |
| **Covarianza** | $S_{xy} = \frac{\sum_i \sum_j (x_i - \bar{x}) \cdot (y_i - \bar{y}) \cdot n_{ij}}{n}$ $= \frac{\sum_i \sum_j x_i \cdot y_i \cdot n_{ij}}{n} - \bar{x} \cdot \bar{y}$ | |
| **Coeficiente correlación** | $r_{xy} = \frac{S_{xy}}{s_x \cdot s_y}$ $-1 \le r \le 1$ | |
| **Dependencia lineal** | $r = -1$ dependencia funcional lineal negativa<br>$-1 < r < 0$ dependencia negativa<br>$r = 0$ no existe dependencia lineal, ni funcional<br>$0 < r < 1$ dependencia positiva<br>$r = 1$ dependencia funcional lineal positiva | |
| **Recta regresión $Y$ sobre $X$** | $y = \bar{y} + \frac{S_{xy}}{{s_x}^2}(x - \bar{x})$ | |

# EJERCICIOS Y PROBLEMAS



## Estadística descriptiva unidimensional

**1.** Se conoce el volumen semanal de residuos sólidos recogidos en $m^3$ 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.

**2.** 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.

**3.** Se ha preguntado a 40 estudiantes por el número de hermanas y hermanos que tenía, y se ha obtenido

| **Número de hermanas/os** | **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.

**4.** Se ha preguntado a 50 estudiantes de 1º de Bachillerato por el número de hermanas/os 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.

## Utiliza una hoja de cálculo con el ordenador

**5.** 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:

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.

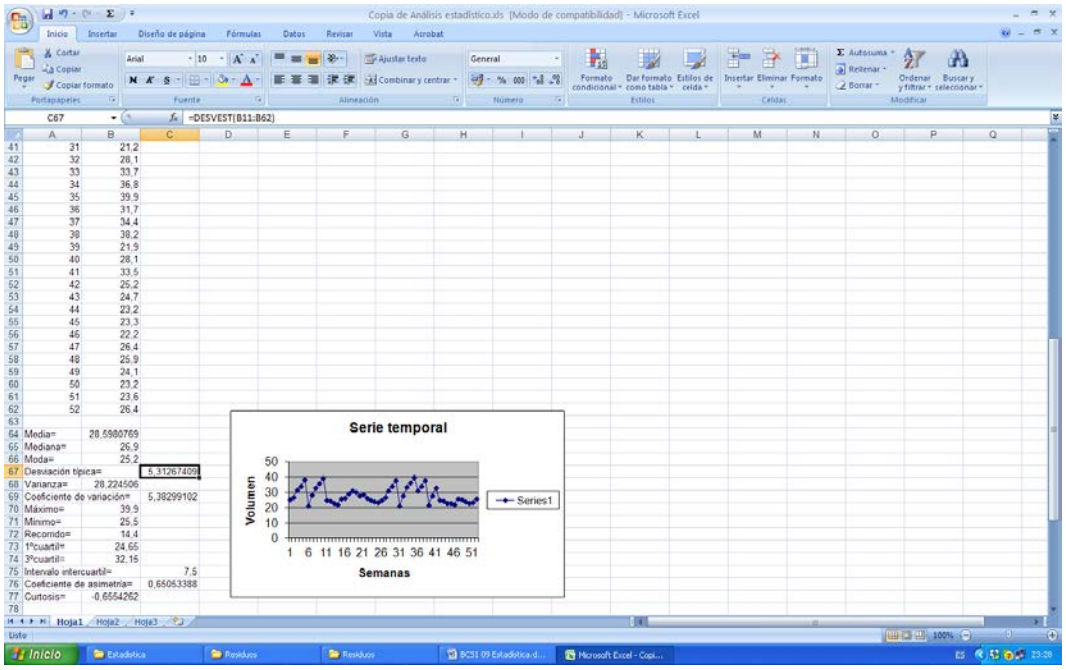
d) Haz una representación gráfica en **serie temporal**, que permita observar tendencias, ciclos y fluctuaciones. Recuerda que, en una serie temporal, en el eje de abscisas está el tiempo de observación y en el eje de ordenadas la magnitud de observación.

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.

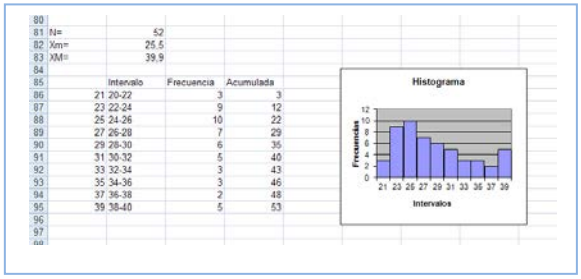
Observa un trozo de pantalla con algunos resultados:



**6.** 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?

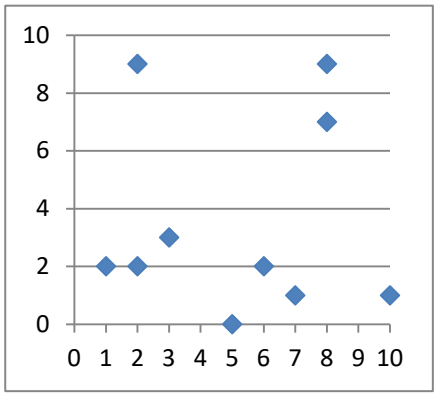
- **Otra investigación:** Vamos a investigar la distribución de la media. Para ello vamos a tomar muestras de tamaño 5. Utiliza la columna G. En G11 escribe =PROMEDIO(B11:B15), en G12 la media de B16 a B20, y así hasta el final. Tenemos calculadas las 10 medias de muestras de tamaño 5. Calcula la media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

## Estadística descriptiva bidimensional

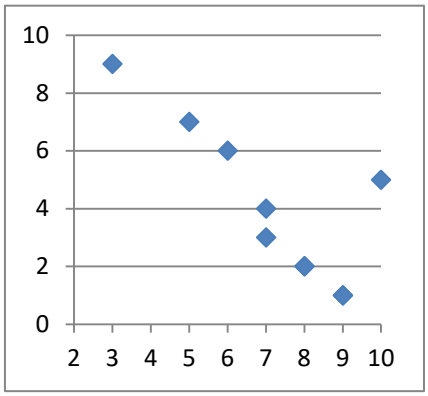
**7.** En una muestra de 10 personas miramos su color de ojos y pelo y encontramos que hay 5 morenos de ojos marrones, 1 moreno de ojos verdes, 3 rubios de ojos azules y 1 rubio de ojos verdes. A) Representa en una tabla de doble entrada esta situación. B) Escribe la tabla de frecuencias relativas. C) Escribe las frecuencias absolutas y relativas marginales. D) Escribe la distribución de frecuencias condicionadas.

**8.** Lola ha calculado los coeficientes de correlación de las tres nubes de puntos adjuntas, y ha obtenido: –0.8, 0.85 y 0.03, pero ahora no recuerda cuál es de cada una. ¿Puedes ayudar a decidir qué coeficiente corresponde con cada nube?

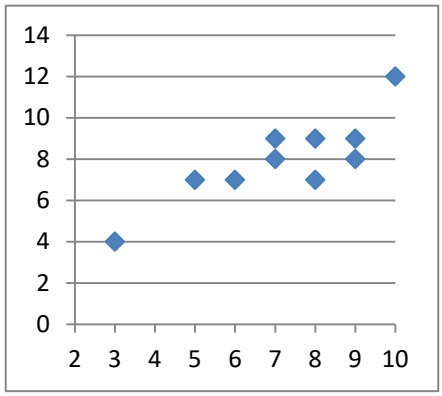
A




B




C




**9.** En una tienda quieren estudiar las ventas del pan de molde en función del precio. Para ello prueban cada semana con un precio distinto y calculan las ventas realizadas. Han obtenido los siguientes datos:

| **Precio (euros)** | 0.5 | 0.7 | 1 | 1.2 | 1.3 | 1.5 | 1.7 | 1.8 | 2 |
|---|---|---|---|---|---|---|---|---|---|
| **Ventas (medias)** | 20.2 | 19.2 | 18.1 | 15.3 | 11.6 | 6 | 4 | 0 | 0 |

a) Representa los datos en un diagrama de dispersión (nube de puntos) e indica a qué conclusiones crees que se va a llegar.

b) Calcula la covarianza, el coeficiente de correlación, la recta de regresión y el coeficiente de determinación

c) Deciden poner un precio de 1.4 euros, ¿cuáles opinas que serían las ventas medias semanales?

**10.** Una compañía aérea realiza un estudio sobre la relación entre las variables $X$, tiempo de un vuelo, en horas; e $Y$, consumo de combustible (gasóleo) para dicho vuelo, en litros, y se han obtenido los siguientes datos.

| ***X (horas)*** | 0.5 | 1 | 1.5 | 2 | 2.5 | 3 |
|---|---|---|---|---|---|---|
| ***Y (litros)*** | 2 250 | 3 950 | 5 400 | 7 300 | 8 500 | 10 300 |

a) Representa los datos en un diagrama de dispersión.

b) Calcula la covarianza y el coeficiente de correlación entre ambas variables. Interpreta los resultados.

c) Calcula la ecuación de las rectas de regresión.

d) Calcula el coeficiente de determinación.

**11.** Preguntamos a 10 estudiantes de 1º de Bachillerato 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.

| **Calificaciones de Matemáticas** | 10 | 3 | 8 | 8 | 5 | 10 | 10 | 8 | 5 | 8 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Minutos diarios que ve la TV** | 0 | 90 | 30 | 20 | 70 | 10 | 0 | 20 | 60 | 30 |
| **Horas semanales de estudio** | 15 | 0 | 10 | 10 | 10 | 15 | 15 | 10 | 5 | 5 |
| **Estatura (en cm)** | 175 | 166 | 155 | 161 | 161 | 177 | 182 | 177 | 167 | 172 |

Queremos estudiar la relación entre las calificaciones de Matemáticas y las otras tres variables. Para ello dibuja los diagramas de dispersión, y calcula los coeficientes de correlación y determinación. Calcula las rectas de regresión.

**12.** Haz un trabajo. Pasa una encuesta a tus compañeros y compañeras de clase. Elige una muestra de 10 personas y hazles dos preguntas con datos numéricos, como por ejemplo, cuánto mide su mano, qué número de zapato calza, el número de libros que lee en un mes, el número de horas que ve la televisión a la semana, dinero que gasta al mes en comprar música, la calificación en Matemáticas de su último examen… Representa los datos obtenidos en una tabla de doble entrada. Haz un estudio completo. Puedes utilizar el ordenador:

a) Escribe en tu cuaderno una tabla de doble entrada de frecuencias absolutas, frecuencias relativas. Obtén las distribuciones marginales y condicionadas.
b) Con las distribuciones unidimensionales, dibuja los diagramas de barras, diagramas de líneas y diagramas de sectores. Calcula las medias, medianas y modas. Calcula las varianzas y las desviaciones típicas. Calcula los cuartiles y los intervalos intercuartílicos.
c) Con las distribuciones bidimensionales, dibuja un diagrama de dispersión, y calcula la covarianza, el coeficiente de correlación y la recta de regresión.
d) Reflexiona sobre los resultados y escribe un informe.

## Utiliza una hoja de cálculo con un ordenador

**13.** El objetivo de esta práctica es estudiar la dispersión entre dos variables, mediante una nube de puntos o diagrama de dispersión, el coeficiente de correlación, la recta de regresión y la regresión cuadrática.

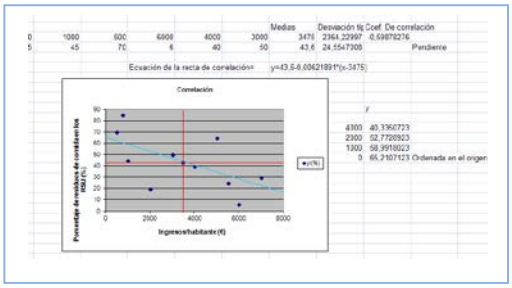
En 10 países se anotan los ingresos medios, en euros, por habitante y año, y el porcentaje medio en los residuos sólidos de comida.

Se obtiene:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $x_i$ (€) | 750 | 5 000 | 7 000 | 2 000 | 5 500 | 1 000 | 500 | 6 000 | 4 000 | 3 000 |
| $y_i$ (%) | 85 | 65 | 30 | 20 | 25 | 45 | 70 | 6 | 40 | 50 |

a) Abre una hoja de cálculo. Copia los datos. Calcula la media y la desviación típica de las $x$, y la media y la desviación típica de las $y$.
b) Representa la nube de puntos. Selecciona los datos, incluyendo a las medias. Aprieta el botón de asistente de gráficos y elige **XY (Dispersión)**. En títulos escribe como **Título del gráfico** *Correlación*, en **Eje de valores (X)** describe la variable $x$ sin olvidar decir las unidades, escribe: *Ingresos/habitante (€),* en **Eje de valores (Y)** describe la variable $y$ sin olvidar decir las unidades, escribe: *Porcentaje de residuos de comida en los RSU (%)*. En **Leyenda** elige no mostrar leyenda.

c) Observa que si $x - \bar{x}$ e $y - \bar{y}$ tienen el mismo signo quedan en los cuadrantes I y III y si lo tienen distinto en II y IV. Cuenta los puntos que quedan en los cuadrantes I y III, cuenta los que quedan en los cuadrantes II y IV. Nos puede dar una idea de la correlación. ¿Va a ser positiva o negativa? ¿Es una correlación fuerte o débil? ¿Entre que valores puede variar el coeficiente de correlación? Estima a ojo un valor para esa correlación.



d) Organiza en Excel una hoja de cálculo que te permita calcular la correlación. Escribe los datos en las filas 3 y 4. En L3 y L4 calcula las medias utilizando la función **PROMEDIO**. En M3 y M4 calcula la desviación típica utilizando la función **DESVEST**. En N3 calcula el coeficiente de correlación, utilizando la función:

**COEF.DE.CORREL**(B3:K3;B4:K4)

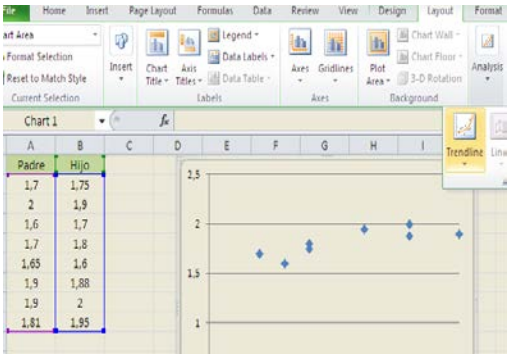
e) Ahora vamos a mejorar nuestro gráfico. Observa que si colocas al ratón encima de un punto indica las coordenadas. Traza las rectas $x = \bar{x}$, $y = \bar{y}$ que indican las medias. Utiliza para ello la paleta de dibujo. Dibújalas en color rojo.

f) La recta de regresión es la recta que hace mínimas las distancias de la nube de puntos. Es la recta: $y = \bar{y} + \rho \frac{s_y}{s_x}(x - \bar{x})$. Calcula en N4 la pendiente de la recta. Escribe la ecuación de la recta. Observa el gráfico. ¿Cómo la habrías estimado a ojo? Evalúa la pendiente y la ordenada en el origen.

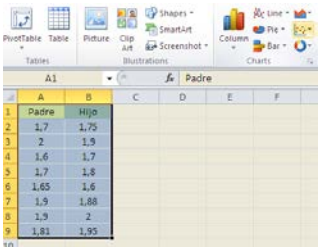
g) Usa la hoja de cálculo para volver a calcular la recta de regresión, ya aproxima de nuevos la nube con otras funciones, parábola, exponencial, polinómica…

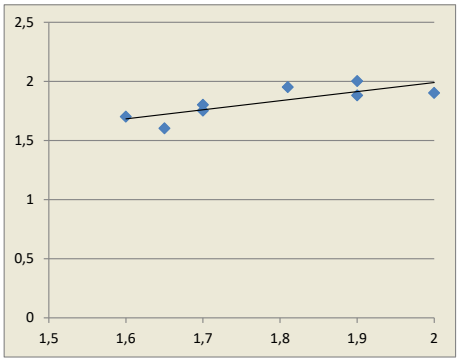
**14.** Se recoge en una tabla la altura (en metros) de un padre y de la de su hijo con 15 años de edad.

| **Padre** | 1.7 | 2 | 1.6 | 1.7 | 1.65 | 1.9 | 1.9 | 1.81 |
|---|---|---|---|---|---|---|---|---|
| **Hijo** | 1.75 | 1.9 | 1.7 | 1.8 | 1.6 | 1.88 | 2 | 1.95 |



a) Utiliza el ordenador para representar el diagrama de dispersión. Copia los datos en una hoja de cálculo en las columnas A y B. Señala las dos series y elige *insertar gráfico de dispersión*. Automáticamente verás que aparece el diagrama de dispersión (nube de puntos). Juega con las opciones para modificar el título, el formato, la escala de los ejes…





b) Dibuja la recta de regresión. Pincha sobre un punto de la nube, y elige "*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*". Al final, si lo has hecho bien, el dibujo debe ser más o menos algo similar a esto:

c) Utiliza la recta para determinar que altura del hijo correspondería a una altura del padre de 1.75 m.

# AUTOEVALUACIÓN

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 |

**1.** La media de errores es

a) 6.75 b) 7 c) 7.9 d) 6.9

**2.** La media de tiempos es

a) 6.75 b) 7 c) 7.9 d) 6.9

**3.** La desviación típica de errores es

a) 1 b) 1.41 c) 1.33 d) 1.2

**4.** La desviación típica de tiempos es

a) 1 b) 1.41 c) 1.33 d) 1.2

**5.** 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

6. 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

**7.** La covarianza es:

a) 1.21 b) –1.5 c) –1.4 d) 1.425

**8.** El coeficiente de correlación es:

a) 0.8 b) –0.8 c) –0.7 d) 0.7

**9.** La recta de regresión lineal de los errores sobre el tiempo es:

a) $y = 3.1 - 0.71x$ b) $y = 3.1 + 0.71x$ c) $y = 0.4 + 0.8x$ d) $y = 0.4 - 0.8x$

**10.** La recta de regresión lineal del tiempo sobre los errores es:

a) $y = 3.1 - 0.71x$ b) $y = 3.1 + 0.7$ c) $y = 0.4 + 0.8x$ d) $y = 0.4 - 0.8x$