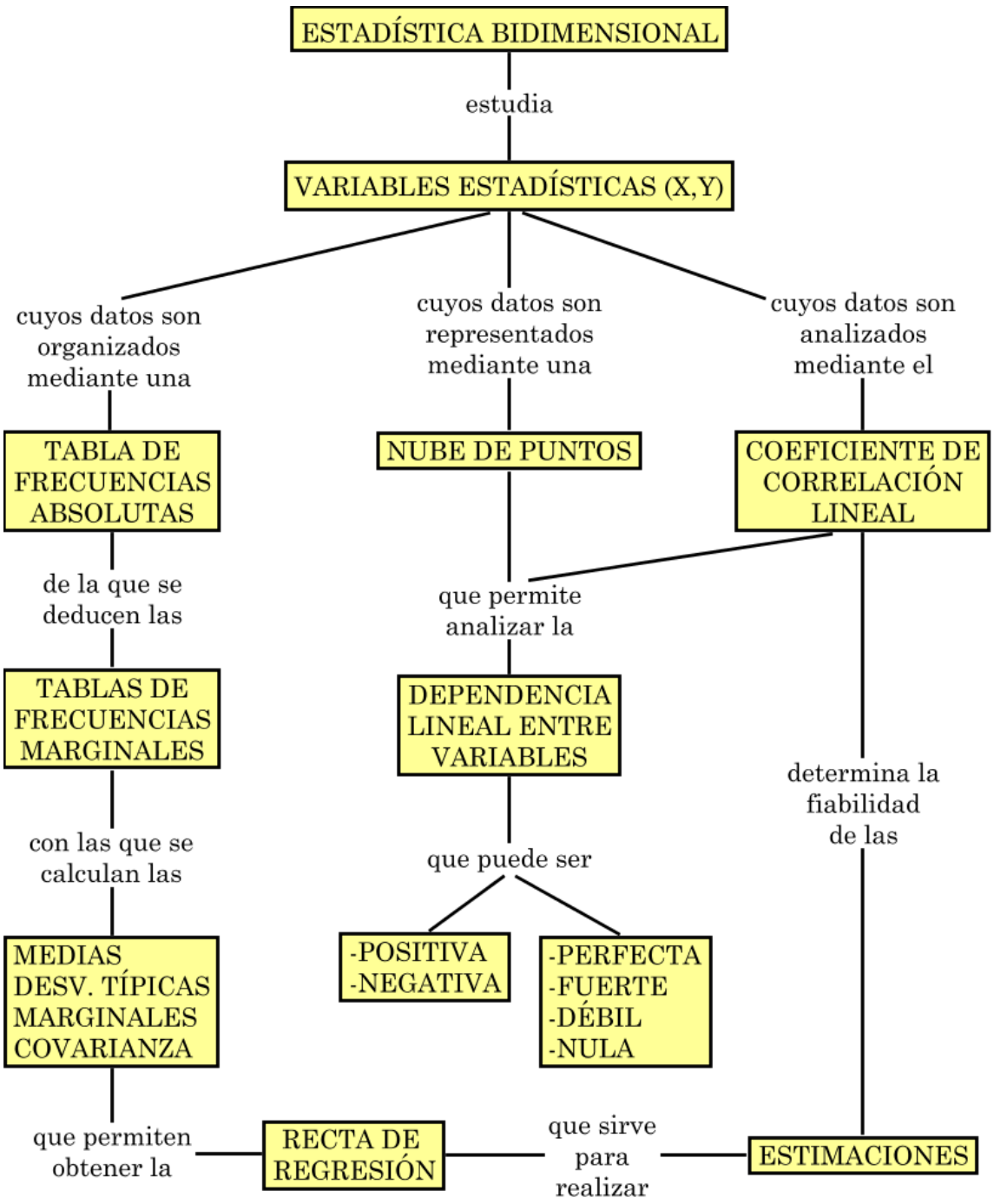
**DEPARTAMENTO DE MATEMÁTICAS.**

**MATEMÁTICAS B DE 4ºESO.**

**UNIDAD 8. VARIABLES BIDIMENSIONALES.**

IES LA BAHIA

## MAPA CONCEPTUAL DE LA UNIDAD



## 1. Variable estadística bidimensional.

Una **variable estadística bidimensional** es una pareja (X, Y) donde tanto X como Y son variables estadísticas que se estudian simultaneamente sobre una misma población.

Los valores que toma la variable bidimensional (X, Y) son de la forma $(x_i, y_j)$, siendo:

$x_1, x_2, \ldots, x_m$ los valores que toma X

$y_1, y_2, \ldots, y_n$ los valores que toma Y

valores tomados sobre cada individuo de la población, respectivamente.

### Tabla de frecuencias absolutas y marginales

Dada una población de N individuos sobre la que se estudia una variable (X, Y)

Se llama **frecuencia absoluta** $f_{ij}$ del par $(x_i, y_j)$ al número de veces que se repite dicho par.

Se llama **frecuencia marginal** $f_i$ del valor $x_i$ al número de veces que se repite el valor $x_i$.

Se llama **frecuencia marginal** $f_j$ del valor $y_j$ al número de veces que se repite el valor $y_j$.

Para organizar los datos, se utiliza una **tabla de frecuencias de doble entrada**, en la que se colocan las frecuencias absolutas $f_{ij}$ de cada pareja $(x_i, y_j)$ de datos en las casillas interiores.

En el supuesto de que las variables tengan los datos agrupados por intervalos, se escribe como $x_i$ ó $y_j$ **la respectiva marca de clase** del intervalo.

| Y \ X | $x_1$ | $x_2$ | … | $x_m$ | Frecuencias marginales de Y |
|---|---|---|---|---|---|
| $y_1$ | $f_{11}$ | $f_{21}$ | … | $f_{m1}$ | $\sum f_{i1}$ |
| $y_2$ | $f_{12}$ | $f_{22}$ | … | $f_{m2}$ | $\sum f_{i2}$ |
| … | … | … | … | … | … |
| $y_n$ | $f_{1n}$ | $f_{2n}$ | … | $f_{mn}$ | $\sum f_{in}$ |
| Frecuencias marginales de X | $\sum f_{1j}$ | $\sum f_{2j}$ | … | $\sum f_{mj}$ | N |

Ejemplo: en un grupo de 30 alumnos se han tomado los datos sobre la variable (X, Y), siendo:

X = número de horas diarias de estudio  Y = número de suspensos en junio

(2,0) (2,2) (0,5) (3,1) (1,2) (2,1) (3,1) (4,0) (0,4) (2,2) (3,1) (2,1) (4,0) (3,1) (2,4)

(3,1) (1,2) (2,1) (2,0) (3,0) (3,1) (2,2) (2,2) (3,1) (0,6) (1,3) (2,2) (3,1) (1,3) (1,4)

Tabla de frecuencias de doble entrada

| Y \ X | 0 | 1 | 2 | 3 | 4 | Total |
|---|---|---|---|---|---|---|
| 0 | | | 2 | 1 | 2 | 5 |
| 1 | | | 3 | 8 | | 11 |
| 2 | | 2 | 5 | | | 7 |
| 3 | | 2 | | | | 2 |
| 4 | 1 | 1 | 1 | | | 3 |
| 5 | 1 | | | | | 1 |
| 6 | 1 | | | | | 1 |
| Total | 3 | 5 | 11 | 9 | 2 | 30 |

Tabla de frecuencias marginales de X

| $x_i$ | $f_i$ |
|---|---|
| 0 | 3 |
| 1 | 5 |
| 2 | 11 |
| 3 | 9 |
| 4 | 2 |
| | 30 |

Tabla de frecuencias marginales de Y

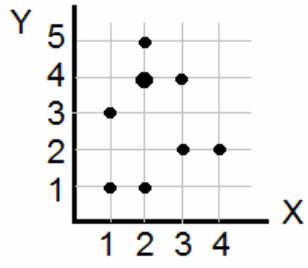
| $y_j$ | $f_j$ |
|---|---|
| 0 | 5 |
| 1 | 11 |
| 2 | 7 |
| 3 | 2 |
| 4 | 3 |
| 5 | 1 |
| 6 | 1 |
| | 30 |

## 2. Dependencia lineal entre variables.

Se llama **diagrama de dispersión** o **nube de puntos** de una variable bidimensional (X, Y) al gráfico obtenido al representar, en unos ejes de coordenadas, los N pares de datos $(x_i, y_j)$ que toma la variable. Si la frecuencia absoluta de un par $(x_i, y_j)$ es mayor que uno, se aumenta el tamaño del punto que lo representa de forma proporcional a su frecuencia.

Ejemplo:

| Y \ X | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 1 | 1 | | |
| 2 | | | 1 | 1 |
| 3 | 1 | | | |
| 4 | | 3 | 1 | |
| 5 | | 1 | | |




Al analizar una variable bidimensional (X, Y), se puede establecer el grado de dependencia que existe entre las variables X e Y que la forman.

– Si la nube de puntos se ajusta a una recta, se dice que X e Y están en **dependencia lineal**.

– Si la nube de puntos se ajusta a otro tipo de curva, se dice que X e Y están en dependencia **funcional** (el calificativo dependerá del tipo de curva que sea: cuadrática, exponencial …).

Por un lado, la **dependencia lineal** puede ser:

– **Perfecta**, si los puntos de la nube se ajustan exactamente a una recta.

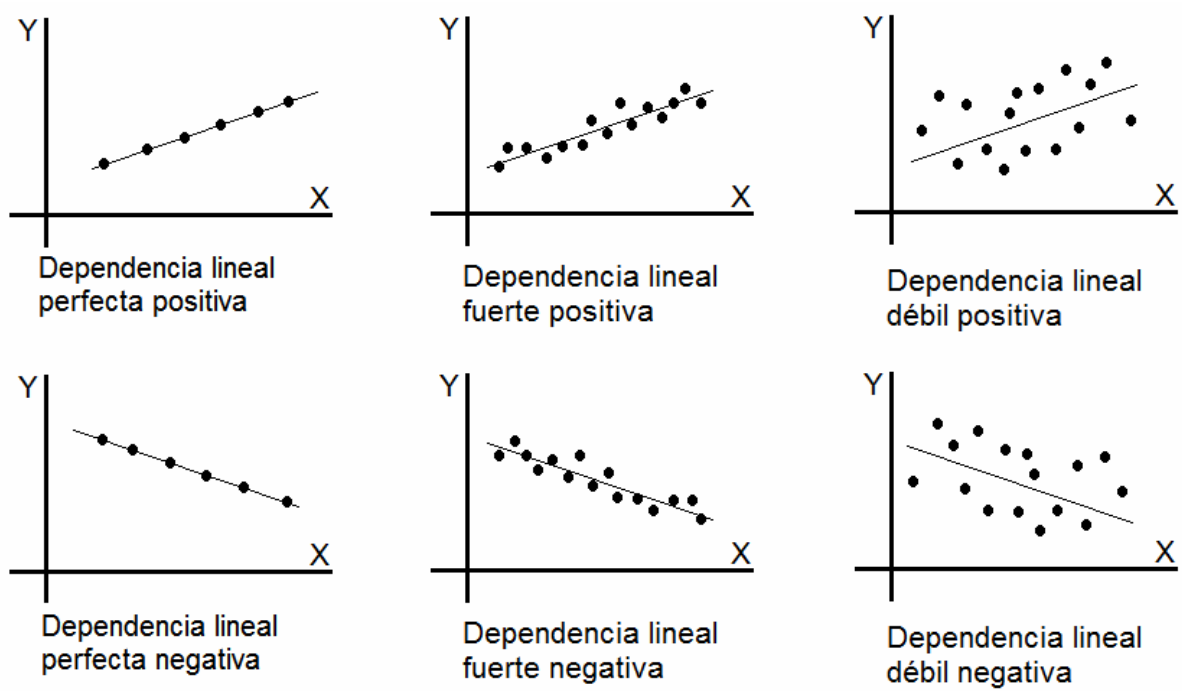
– **Fuerte**, si los puntos de la nube no se ajustan exactamente a una recta, pero se encuentran muy próximos, a muy poca distancia de una recta. En este caso, la nube de puntos es estrecha.

– **Débil**, si los puntos de la nube se encuentran muy alejados de una recta que se pudiera trazar entre ellos. En este caso, la nube de puntos es ancha.

Por otro lado, la **dependencia lineal** puede ser:

– **Positiva**, si la nube de puntos, de izquierda a derecha, se orienta en sentido creciente.

– **Negativa**, si la nube de puntos, de izquierda a derecha, se orienta en sentido decreciente.



Ejemplos:

1. Están en dependencia lineal perfecta el peso colgado de un muelle y su alargamiento.

2. Están en dependencia lineal fuerte el número de partidos ganados y la posición final en la liga de los equipos de fútbol.

3. Están en dependencia lineal débil la edad y el tiempo diario dedicado a la lectura por la población de una ciudad.

## 3. Correlación lineal.

La palabra correlación significa "correspondencia o relación recíproca entre dos o más cosas".

El grado de dependencia lineal que existe entre dos variables se puede medir mediante el **coeficiente de correlación lineal** de Pearson.

Para el cálculo de este coeficiente se necesitan calcular previamente los siguientes parámetros:

**Media marginal de X** $\quad \overline{x} = \dfrac{x_1 f_1 + x_2 f_2 + \ldots + x_m f_m}{N} = \dfrac{\sum x_i f_i}{N}$

**Media marginal de Y** $\quad \overline{y} = \dfrac{y_1 f_1 + y_2 f_2 + \ldots + y_n f_n}{N} = \dfrac{\sum y_j f_j}{N}$

**Desviación típica marginal de X** $\quad \sigma_x = \sqrt{\dfrac{\sum (x_i - \overline{x})^2 \cdot f_i}{N}} = \sqrt{\dfrac{\sum x_i^2 \cdot f_i}{N} - \overline{x}^2}$

**Desviación típica marginal de Y** $\quad \sigma_y = \sqrt{\dfrac{\sum (y_j - \overline{y})^2 \cdot f_j}{N}} = \sqrt{\dfrac{\sum y_j^2 \cdot f_j}{N} - \overline{y}^2}$

Se llama **covarianza** de una variable (X, Y) a la media aritmética de los productos de las desviaciones respecto a la media de las variables X e Y. La covarianza se representa por $\sigma_{xy}$.

$$\sigma_{xy} = \frac{\sum f_{ij} \cdot (x_i - \overline{x}) \cdot (y_j - \overline{y})}{N} = \frac{\sum f_{ij} \cdot x_i \cdot y_j}{N} - \overline{x} \cdot \overline{y}$$

Se llama **coeficiente de correlación lineal** al número real $r = \dfrac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$ $\qquad [-1 \le r \le 1]$

El coeficiente de correlación lineal siempre es un número real comprendido entre –1 y 1.

**Interpretación del coeficiente de correlación**

Si $r = 1 \quad \Rightarrow$ X e Y están en **dependencia lineal perfecta positiva**.

Si $r = -1 \quad \Rightarrow$ X e Y están en **dependencia lineal perfecta negativa**.

Si $r = 0 \quad \Rightarrow$ X e Y son **independientes linealmente**.

Si r es un número cercano a 1 $\quad \Rightarrow$ X e Y están en **dependencia lineal fuerte positiva**, tanto más fuerte cuanto más próximo esté r de 1.

Si r es un número cercano a –1 $\quad \Rightarrow$ X e Y están en **dependencia lineal fuerte negativa**, tanto más fuerte cuanto más próximo esté r de –1.

Si r es un número cercano a 0 $\quad \Rightarrow$ X e Y tienen **poca dependencia o dependencia débil,** tanto más débil cuanto más próximo esté r de 0.

Ejemplo: se quiere investigar si existe alguna relación entre las calificaciones en Matemáticas y en Física de un grupo de 12 personas. Éstos son los datos recogidos:

| X = nota de Matemáticas | 2 | 3 | 4 | 4 | 5 | 6 | 6 | 7 | 7 | 8 | 10 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Y = nota de Física | 1 | 3 | 2 | 4 | 4 | 4 | 6 | 4 | 6 | 7 | 9 | 10 |

Para medir el grado de dependencia lineal entre X e Y, se siguen los siguientes pasos:

– En primer lugar, hay que calcular $\overline{x}$, $\sigma_x$ en la tabla de frecuencias marginales de X:

| $x_i$ | $f_i$ | $x_i \cdot f_i$ | $x_i^2 \cdot f_i$ |
|---|---|---|---|
| 2 | 1 | 2 | 4 |
| 3 | 1 | 3 | 9 |
| 4 | 2 | 8 | 32 |
| 5 | 1 | 5 | 25 |
| 6 | 2 | 12 | 72 |
| 7 | 2 | 14 | 98 |
| 8 | 1 | 8 | 64 |
| 10 | 2 | 20 | 200 |
| | 12 | 72 | 504 |

La media es $\overline{x} = \frac{72}{12} = 6$

$\sigma_x^2 = \frac{504}{12} - 6^2 = 42 - 36 = 6$

La desviación típica es $\sigma_x = \sqrt{6} \cong 2{,}45$

– En segundo lugar, hay que calcular $\overline{y}$, $\sigma_y$ en la tabla de frecuencias marginales de Y:

| $y_j$ | $f_j$ | $y_j \cdot f_j$ | $y_j^2 \cdot f_j$ |
|---|---|---|---|
| 1 | 1 | 1 | 1 |
| 2 | 1 | 2 | 4 |
| 3 | 1 | 3 | 9 |
| 4 | 4 | 16 | 64 |
| 6 | 2 | 12 | 72 |
| 7 | 1 | 7 | 49 |
| 9 | 1 | 9 | 81 |
| 10 | 1 | 10 | 100 |
| | 12 | 60 | 380 |

La media es $\overline{y} = \frac{60}{12} = 5$

$\sigma_y^2 = \frac{380}{12} - 5^2 = 31{,}67 - 25 = 6{,}67$

La desviación típica es $\sigma_y = \sqrt{6{,}67} \cong 2{,}58$

– En tercer lugar, hay que calcular $\sigma_{xy}$ en la tabla de frecuencias de (X, Y):

| $x_i$ | $y_j$ | $f_{ij}$ | $x_i \cdot y_j \cdot f_{ij}$ |
|---|---|---|---|
| 2 | 1 | 1 | 2 |
| 3 | 3 | 1 | 9 |
| 4 | 2 | 1 | 8 |
| 4 | 4 | 1 | 16 |
| 5 | 4 | 1 | 20 |
| 6 | 4 | 1 | 24 |
| 6 | 6 | 1 | 36 |
| 7 | 4 | 1 | 28 |
| 7 | 6 | 1 | 42 |
| 8 | 7 | 1 | 56 |
| 10 | 9 | 1 | 90 |
| 10 | 10 | 1 | 100 |
| | | 12 | 431 |

La covarianza de (X, Y) es

$\sigma_{xy} = \frac{\sum f_{ij} \cdot x_i \cdot y_j}{N} - \overline{x} \cdot \overline{y} = \frac{431}{12} - 6 \cdot 5 = 35{,}92 - 30 = 5{,}92$

– Finalmente se calcula el coeficiente de correlación lineal

$r = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = \frac{5{,}92}{2{,}45 \cdot 2{,}58} \cong 0{,}94$

Por lo tanto, como $r$ es un número cercano a 1, se puede afirmar que X e Y están en **dependencia lineal fuerte positiva**,

## 4. Recta de regresión. Estimaciones.

Dada una variable bidimensional (X, Y), se llama **recta de regresión** a la recta que mejor se aproxima a los puntos de su diagrama de dispersión o nube de puntos.

Nota: la recta de regresión es la que hace mínima la suma de las distancias entre las ordenadas de cada punto y las de la recta.

La recta de regresión se puede expresar de dos formas:

1. Si se expresa Y en función de X , se llama **recta de regresión de Y sobre X**

$y - \overline{y} = \dfrac{\sigma_{xy}}{\sigma_x^2} \cdot (x - \overline{x})$ A la pendiente $\dfrac{\sigma_{xy}}{\sigma_x^2}$ se le llama coeficiente de regresión.

2. Si se expresa X en función de Y, se llama **recta de regresión de X sobre Y**

$x - \overline{x} = \dfrac{\sigma_{xy}}{\sigma_y^2} \cdot (y - \overline{y})$ A la pendiente $\dfrac{\sigma_{xy}}{\sigma_y^2}$ se le llama coeficiente de regresión.

### Estimación de resultados

La recta de regresión permite obtener, de forma aproximada, el valor esperado para una de las variables a partir de un valor de la otra. Al valor obtenido se le llama **estimación**.

La **fiabilidad** o grado de aproximación de la estimación depende de dos aspectos:

1. De que el valor sustituido se encuentre dentro del rango de valores obtenidos en el estudio estadístico, o muy próximos al mismo.

2. De que el coeficiente r de correlación sea próximo a $\pm 1$, tanto más fiable cuanto más se acerque a 1 ó –1. Por el contrario, la estimación carece de validez si el coeficiente r es un valor próximo a cero.

Ejemplo: se quiere investigar si existe alguna relación entre la estatura de un grupo de diez personas y su peso. Éstos son los datos recogidos y los correspondientes cálculos:

| X = Estatura (en cm) | 161 | 167 | 168 | 169 | 172 | 173 | 176 | 177 | 182 | 191 |
|---|---|---|---|---|---|---|---|---|---|---|
| Y = Peso (en kg) | 56 | 61 | 64 | 70 | 65 | 61 | 68 | 69 | 76 | 79 |

$\overline{x} = 173{,}6$ cm $\quad \overline{y} = 66{,}9$ kg $\quad \sigma_x = 8{,}05 \quad \sigma_y = 6{,}67 \quad \sigma_{xy} = 48{,}16$

$r = \dfrac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = \dfrac{48{,}16}{8{,}05 \cdot 6{,}67} \cong 0{,}897 \quad \sigma_x^2 = 64{,}8025$

Como el coeficiente r es próximo a 1, la altura y el peso de este grupo de personas están en **dependencia lineal fuerte positiva**. Se puede afirmar que en este grupo, una mayor estatura va ligada a un mayor peso.

La recta de regresión de Y sobre X tiene por ecuación $y - 66{,}9 = \dfrac{48{,}16}{64{,}8025} \cdot (x - 173{,}6)$

$\Rightarrow \quad y - 66{,}9 = 0{,}7432 \cdot (x - 173{,}6) \quad \Rightarrow \quad y = 0{,}7432 \cdot x - 62{,}1195$

Por otra parte, como el coeficiente r es próximo a 1, se puede realizar una **estimación fiable** del peso esperado para una persona que tenga estatura comprendida dentro del rango de la muestra [161, 191].
Por ejemplo, se puede estimar el peso que se espera para una persona con 180 cm de estatura. Basta con sustituir x por 180 en la ecuación de la recta de regresión:

$y = 0{,}7432 \cdot 180 - 62{,}1195 \quad \Rightarrow \quad y \cong 71{,}66$ kg