

---

## 2. Variables estadísticas bidimensionales

---

### 2.1. Variables bidimensionales. Tabulación de datos: tabla de correlación y tabla de contingencia

#### Definiciones

Sean  $X$  e  $Y$  dos variables estadísticas que hacen referencia a dos características de una población o muestra de tamaño  $N$ . Sean  $x_1, x_2, \dots, x_r$  y  $y_1, y_2, \dots, y_s$  las distintas categorías o valores que toman  $X$  e  $Y$  y sea  $\{(x_i, y_j)_k\}_{k=1}^N$  el conjunto de todos los pares de caracteres cuyo total es  $N$ . Estos pares son realizaciones de la variable bidimensional  $(X, Y)$ . Se definen

- **Frecuencia absoluta conjunta** del par  $(x_i, y_j)$  de caracteres: Número de pares  $(x_i, y_j)$  en el número total de observaciones de  $(X, Y)$  que se denota por  $n_{ij}$ , siendo

$$N = \sum_{j=1}^s \sum_{i=1}^r n_{ij}$$

- **Frecuencia relativa conjunta** del par  $(x_i, y_j)$ : Proporción de observaciones de este par en el conjunto de todos los pares

$$f_{ij} = n_{ij}/N$$

El conjunto formado por todos los pares y las frecuencias absolutas  $\{(x_i, y_j), n_{ij}\}$ , con  $i = 1, 2, \dots, r$  y  $j = 1, 2, \dots, s$  se denomina distribución de frecuencias bidimensional. Se representa en una tabla de doble entrada llamada **tabla de correlación** si las variables  $X$  e  $Y$  son cuantitativas y **tabla de contingencia** si alguna de las variables es cualitativa. El proceso de construcción de las tablas se llama **tabulación**.

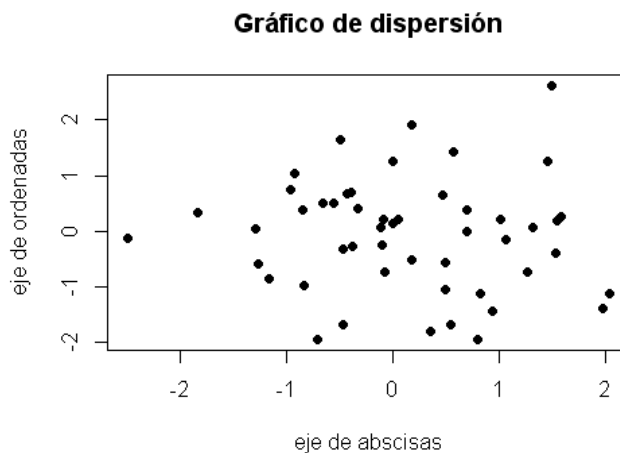
$X/Y$	$y_1$	$y_2$	$\cdots$	$y_s$
$x_1$	$n_{11}$	$n_{12}$	$\cdots$	$n_{1s}$
$x_2$	$n_{21}$	$n_{22}$	$\cdots$	$n_{2s}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$x_r$	$n_{r1}$	$n_{r2}$	$\cdots$	$n_{rs}$



Consultar cómo hacer la media de cada una de las variables ( $x$  e  $y$ )



Se puede representar una variable bidimensional utilizando un **diagrama de dispersión o nube de puntos**, donde en el plano se representa en el eje de abscisas



la primera variable y en el de ordenadas la segunda y cada  $(x_i, y_j)$  es un punto en el plano. Esta representación puede ayudar a descubrir visualmente la existencia de algún tipo de relación entre dos variables.

### 2.2. Distribuciones marginales. Distribuciones condicionadas. Independencia estadística

#### Definiciones

##### ■ Distribuciones de frecuencias marginales

Cuando interesa saber el número de pares que tienen un determinado valor de la variable estadística  $X$  sin importar lo que vale la variable estadística  $Y$  se obtiene la distribución de las frecuencias marginales de la variable  $X$ .

En concreto la frecuencia marginal absoluta de la modalidad o valor  $x_i$  de  $X$  es:

$$n_{i\cdot} = \sum_{j=1}^s n_{ij}, \quad i = 1, \dots, r$$

siendo la frecuencia marginal relativa

$$f_{i\cdot} = \frac{n_{i\cdot}}{N}$$

Análogamente, la frecuencia marginal absoluta de la modalidad o valor  $y_j$  de la variable estadística  $Y$  es

$$n_{\cdot j} = \sum_{i=1}^r n_{ij}, \quad j = 1, \dots, s$$

y su frecuencia marginal relativa:

$$f_{\cdot j} = \frac{n_{\cdot j}}{N}$$

$X/Y$	$y_1$	$y_2$	$\cdots$	$y_s$	$n_{i\cdot}$
$x_1$	$n_{11}$	$n_{12}$	$\cdots$	$n_{1s}$	$n_{1\cdot}$
$x_2$	$n_{21}$	$n_{22}$	$\cdots$	$n_{2s}$	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$x_r$	$n_{r1}$	$n_{r2}$	$\cdots$	$n_{rs}$	$n_{r\cdot}$
$n_{\cdot j}$	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot s}$	$N$

La distribución marginal de frecuencias absolutas de  $X$  es  $\{x_i, n_{i\cdot}\}_{i=1,\dots,r}$ , donde  $n_{i\cdot}$  es el número de pares del conjunto total de los pares de valores que toman  $X$  e  $Y$ , para los que la característica  $X$  toma el valor  $x_i$  sea cual sea el valor de la característica  $Y$ .

Análogamente la distribución marginal de frecuencias absolutas de la variable estadística  $Y$  es  $\{y_j, n_{\cdot j}\}_{j=1,\dots,s}$ , donde  $n_{\cdot j}$  es el número de pares para los cuales la variable  $Y$  toma el valor  $y_j$  independientemente del valor que toma  $X$ .

#### ■ Distribuciones de frecuencias condicionadas

Se construyen para cada una de las dos variables cuando se fija un valor o valores concretos de la otra variable.

1. Por ejemplo, si se fija un valor de la variable  $Y$ , denotado por  $y_j$ , se define la distribución de  $X$  condicionada a que la variable  $Y$  tome el valor  $y_j$ , que se denota  $X/Y = y_j$ , como aquella cuya frecuencia absoluta de que  $X$  valga  $x_i$  (condicionada a que  $Y = y_j$ ) es  $n_{ij}$ , siendo la frecuencia relativa de que  $X$  tome el valor  $x_i$  condicionada a que la variable  $Y$  sea igual a  $y_j$

$$f_{i/j} = \frac{n_{ij}}{n_{\cdot j}}$$

Análogamente se puede fijar un valor de la variable estadística  $X$ ,  $x_i$ , y se puede definir la distribución de  $Y$  condicionada a que  $X = x_i$ , es decir  $Y/X = x_i$ .

$X/Y = y_j$	$n_{i/j}$	$f_{i/j}$
$x_1$	$n_{1/j} = n_{1j}$	$f_{1/j} = n_{1j}/n_{.j}$
$x_2$	$n_{2/j} = n_{2j}$	$f_{2/j} = n_{2j}/n_{.j}$
$\vdots$	$\vdots$	$\vdots$
$x_r$	$n_{r/j} = n_{rj}$	$f_{r/j} = n_{rj}/n_{.j}$
	$n_{.j}$	1

2. Las distribuciones de frecuencias condicionadas permite estudiar el comportamiento de una de la variables estadísticas cuando la otra variable cumple una cierta condición.

**Importante:** Las distribuciones marginales y condicionadas son distribuciones de frecuencias unidimensionales, luego para cada una de ellas pueden calcularse las medidas estudiadas en el tema 1.

- **Independencia estadística** Dos variables estadísticas  $X$  e  $Y$  son independientes si los valores que toma una de las variables no dependen de los valores que tome la otra variable. Cuando esto ocurre

$$f_{i/1} = f_{i/2} = \dots = f_{i/s} \quad i = 1, 2, \dots, r$$

$$f_{1/j} = f_{2/j} = \dots = f_{r/j} \quad j = 1, 2, \dots, s$$

lo que es equivalente a que

$$f_{ij} = f_{i.} \cdot f_{.j}$$

o bien

$$\frac{n_{ij}}{N} = \frac{n_{.j}}{N} \cdot \frac{n_{i.}}{N}$$

Para que dos variables sean estadísticamente independientes se ha de cumplir que cada una de las frecuencias conjuntas sea el producto de las correspondientes frecuencias marginales.

### Observación

Si alguna de las frecuencias conjuntas es igual a 0, las variables son dependientes.

## 2.3. Covarianza y coeficiente de correlación lineal

### Definición

La covarianza entre dos variables  $X$  e  $Y$  se define:

$$S_{XY} = \frac{\sum_{i=1}^r \sum_{j=1}^s (x_i - \bar{x})(y_j - \bar{y})n_{ij}}{N}$$

que es una medida del grado de relación lineal existente entre las variables  $X$  e  $Y$

### Propiedades

- De la definición de covarianza se obtiene una expresión alternativa:

$$S_{XY} = \frac{\sum_{i=1}^r \sum_{j=1}^s x_i y_j n_{ij}}{N} - \bar{x} \cdot \bar{y}$$

- Si las variables son independientes la covarianza vale cero.

- No le afectan los cambios de origen.

- Le afectan los cambios de escala:

Sean las variables  $U = a + bX$  y  $V = c + dY$  construidas a partir de las variables  $X$  e  $Y$ , con  $a, b, c$  y  $d$  parámetros, entonces

$$S_{UV} = bdS_{XY}$$

### Observación

Si la covarianza es cero, no hay relación lineal entre las variables, pero las variables pueden estar relacionadas de otra manera.

### Definiciones

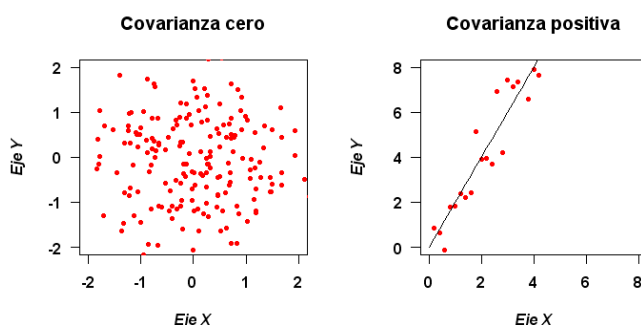
- Se define el *coeficiente de correlación lineal* de las variables estadísticas  $X$  e  $Y$  como:

$$r_{XY} = \frac{S_{XY}}{S_X S_Y}$$

- Las variables  $X$  e  $Y$  son *incorreladas* cuando su coeficiente de correlación lineal es cero (lo que implica que su covarianza es cero).

### Propiedades

- El coeficiente de correlación lineal de dos variables estadísticas toma valores entre  $-1$  y  $+1$ .
- El signo de  $r_{XY}$  coincide con el signo de la covarianza.
- Sean las variables estadísticas  $X$  e  $Y$  y las variables obtenidas de una transformación lineal (supone un cambio de origen y de escala)  $U = a + bX$  y  $V = c + dY$ , se cumple que  $r_{UV} = r_{XY}$  si  $b$  y  $d$  tienen el mismo signo, mientras que  $r_{UV} = -r_{XY}$  si  $b$  y  $d$  tienen distinto signo.
- Si dos variables son estadísticamente independientes, entonces están incorreladas (el recíproco no es cierto).
- Si  $S_{XY} \neq 0$ , entonces  $X$  e  $Y$  son dependientes. El recíproco no es cierto.



## 2.4. Recta de regresión. Estimación de coeficientes. Bondad de ajuste. Predicción

Sean dos variables estadísticas  $X$  e  $Y$  con distribución conjunta de frecuencias  $\{(x_i, y_j), n_{ij}\}$  que se representa gráficamente mediante un diagrama de dispersión o nube de puntos.

El objetivo del procedimiento llamado **regresión lineal** es determinar la recta que mejor representa dicha nube de puntos según un criterio.

La regresión lineal mínimo cuadrática explica el comportamiento de  $Y$ , variable dependiente, a partir de  $X$ , variable independiente, utilizando la recta  $y = ax + b$  cuyo criterio consiste en minimizar la suma de los cuadrados de la diferencia entre el valor observado de la variable dependiente ( $y_i$ ) y el valor estimado mediante la recta, llamada recta de regresión, de la variable dependiente ( $\hat{y}_i = ax_i + b$ ). Esta diferencia se llama residuo  $\hat{y}_i - y_i = e_i$ .

Esquemáticamente el proceso para obtener la recta de regresión utilizando mínimos cuadrados es:

- Recta de regresión:

La recta buscada adopta la forma:  $\hat{y} = a + bx$  siendo  $a$  y  $b$  los parámetros a determinar.

- Residuo para el par  $(x_i, y_i)$ :

$$e_i = y_i - \hat{y}_i = y_i - (ax_i + b)$$

- Objetivo de la regresión mínimo cuadrática: Encontrar los valores de  $a$  y  $b$  tal que se minimice la suma del cuadrado de los residuos.

$$\min_{a,b} \sum_{i=1}^N e_i^2 = \min_{a,b} \sum_{i=1}^N (y_i - ax_i - b)^2 = \min_{a,b} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

- Solución del problema: utilizando las técnicas de optimización se obtiene el resultado

$$b = \frac{S_{XY}}{S_X^2}, \quad a = \bar{y} - b\bar{x} = \bar{y} - \frac{S_{XY}}{S_X^2} \bar{x}$$

- Luego la **recta de regresión** de  $Y$  sobre  $X$  se puede expresar:

$$\hat{y} - \bar{y} = \frac{S_{XY}}{S_X^2} (x - \bar{x}) \quad \text{Fórmula recta de regresión}$$

### Definición

En un ajuste lineal de dos variables estadísticas  $X$  e  $Y$  el **coeficiente de determinación** se define como:

$$R^2 = \frac{S_{XY}^2}{S_X^2 S_Y^2} = r_{XY}^2$$

Se tiene que  $0 \leq R^2 \leq 1$ , que da el porcentaje en el que la regresión es fiable.

### Observaciones importantes

- $r_{XY} = 1$  relación lineal perfecta positiva (todos los puntos están sobre una recta de pendiente positiva).
- $r_{XY} = -1$  relación lineal perfecta negativa (todos los puntos están sobre una recta de pendiente negativa).
- $r_{XY} = 0$  **inexistencia de relación lineal** (covarianza cero). Las variables son incorreladas o independientes.
- $-1 < r_{XY} < 0$  relación lineal negativa.
- $0 < r_{XY} < 1$  relación lineal positiva.

### Predicción

Dada la recta de regresión de  $Y$  sobre  $X$  obtenida a partir de  $N$  pares

$$(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$$

$$\hat{y} - \bar{y} = \frac{S_{XY}}{S_X^2}(x - \bar{x})$$

si se tiene un nuevo valor de la variable  $X$ ,  $x_{N+1}$ , la predicción del correspondiente valor de la variable dependiente  $Y$ ,  $\hat{y}_{N+1}$ , se obtiene como :

$$\hat{y}_{N+1} = \bar{y} + \frac{S_{XY}}{S_X^2}(x_{N+1} - \bar{x})$$

La bondad de esta predicción dependerá del valor del coeficiente de determinación del ajuste.

### Observaciones

- La predicción será más fiable cuanto más cerca esté  $x_{N+1}$  del rango de variación de los datos utilizados para estimar la recta de regresión.
- La predicción será mejor cuanto mayor sea el número de datos.



## 2.5. Ejemplos

### Ejemplo 1

- a) Construir la tabla de correlación correspondiente a los siguientes datos sobre el número de hijos de la familia ( $X$ ) de 16 alumnos y su edad ( $Y$ )

$X$	2	2	3	3	3	3	4	4	3	3	3	4	2	4	2	4
$Y$	18	18	19	20	18	22	19	20	20	20	22	20	19	23	19	18

- b) ¿Cuál es el porcentaje de alumnos con 18 años y solo dos hijos en su familia?
- c) Distribuciones marginales. ¿Qué porcentaje de alumnos forman parte de una familia con tres hijos? ¿Cuál es el porcentaje de familias en las que hay un estudiante menor de 20 años?
- d) Construir la distribución de la edad para los alumnos que pertenecen a una familia de cuatro hijos. ¿cuál es la moda de esta distribución? ¿cuál es el número medio de hijos en las familias de los alumnos de más de 20 años?

### Solución

$X Y$	18	19	20	22	23	$n_{i.}$
2	2	2	0	0	0	4
3	1	1	3	2	0	7
4	1	1	2	0	1	5
$n_{.j}$	4	4	5	2	1	16

- a)
- b) El 50 % de los alumnos de 18 años pertenece a una familia con 2 hijos.

$X$	$n_{i.}$
2	4
3	7
4	5

- c)

**Ejemplo 2**

Sea la tabla de frecuencias conjunta de la variable estadística bidimensional

$X/Y$	1	2	3	$n_{i.}$
1	2	3	1	6
2	4	6	2	12
3	6	9	3	18
$n_{.j}$	12	18	6	36

Calcular  $S_{XY}$ . ¿Son las variables estadísticas  $X$  e  $Y$  independientes?

**Solución**

$S_{XY} = 0$  y las variables son independientes, lo que se observa fácilmente.

**Ejemplo 3**

Sea la tabla de frecuencias conjunta de la variable estadística bidimensional  $(X, Y)$

$X/Y$	1	2	3	$n_{i.}$
-1	0	1	1	0
0	1	0	1	2
1	0	1	0	1
$n_{.j}$	1	2	1	4

Comprobar que  $X$  e  $Y$  no son independientes, siendo  $S_{XY} = 0$ .

**Solución**

No son independientes porque  $f_{i.} \cdot f_{.j} \neq f_{ij}$ . Sin embargo,  $S_{XY} = 0$ .

**Ejemplo 4**

Se dispone de información sobre las subvenciones recibidas por un sector ( $X$ , en millones de euros) durante diez años consecutivos, así como del número de contrataciones llevadas a cabo por las empresas de dicho sector ( $Y$ ) durante el mismo periodo

- Obtener la recta de regresión que expresa el volumen de empleo generado en función de la subvención recibida.
- Estimar la bondad del ajuste.

$X$	$Y$
1.52	145.00
1.74	180.00
1.83	182.00
1.75	155.00
1.92	200.00
2.06	220.00
2.14	240.00
2.08	200.00
2.06	179.00
1.96	164.00

- c) ¿Qué cantidad de nuevos empleos esperarían obtenerse el próximo año si la subvención planificada es de dos millones de euros?