Estadística: estadística
Grado en Relaciones Laborales | Curso 2019-2020
Tema 2. Análisis descriptivo de dos variables

Alejandro Saavedra Nieves

Distribución de frecuencias bidimensional

En adelante, se trabaja con n pares de observaciones de la variable (X, Y) que pueden presentarse:

- Individualmente o en extensión: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.
- En tabla de frecuencias de doble entrada, representando una distribución bidimensional de frecuencias

Ejemplo Una muestra de 500 viviendas en Galicia se ha clasificado según su superficie en m^2 (X) y el número de dormitorios (Y) resultando la siguiente distribución conjunta de frecuencias:

X/Y	1	2	3	4
0-60	9	23	6	0
60-90	7	65	120	22
90-120	2	15	87	50
120-200	0	5	29	60

Distribución de frecuencias bidimensional

- La variable X toma r valores diferentes: $x_1, x_2, \ldots, x_r$.
- La variable Y toma c valores diferentes: $y_1, y_2, \ldots, y_c$.
- La frecuencia absoluta conjunta del valor x_i de X con el valor y_j de Y es n_{ij}, para todo $i = 1, \ldots, r$ y para todo $j = 1, \ldots, c$.

Propiedades

- El número total de observaciones es n:

$$n = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij}$$

- La frecuencia relativa conjunta para todo $i = 1, \ldots, r$ y para todo $j = 1, \ldots, c$ es

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

- La suma de todas las frecuencias relativas es igual a 1, es decir,

$$\sum_{i=1}^{r} \sum_{j=1}^{c} f_{ij} = 1.$$

Tabla de doble entrada

La distribución bidimensional de frecuencias suele presentarse en una tabla de doble entrada conocida como:

- Tabla de correlación si las variables son cuantitativas.
- Tabla de contingencia si alguna de las variables es cualitativa

Notación de la tabla

X/Y	y_1	y_2	$\dots$	y_c	
x_1	n_{11}	n_{12}	$\dots$	n_{1c}	
x_2	n_{21}	n_{22}	$\dots$	n_{2c}	
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	
x_r	n_{r1}	n_{r2}	$\dots$	n_{rc}	
					n

Tabla de doble entrada: ejemplo

Ejemplo Una muestra de 500 viviendas en Galicia se ha clasificado según su superficie en m^2 (X) y el número de dormitorios (Y):

X/Y	1	2	3	4
0-60	9	23	6	0
60-90	7	65	120	22
90-120	2	15	87	50
120-200	0	5	29	60

a) ¿Cuál será la distribución en frecuencias relativas?

X/Y	1	2	3	4
0-60	0.018	0.046	0.012	0
60-90	0.014	0.130	0.240	0.044
90-120	0.004	0.030	0.174	0.100
120-200	0	0.010	0.058	0.120

Tabla de doble entrada: ejemplo

Ejemplo Una muestra de 500 viviendas en Galicia se ha clasificado según su superficie en m^2 (X) y el número de dormitorios (Y):

X/Y	1	2	3	4
0-60	9	23	6	0
60-90	7	65	120	22
90-120	2	15	87	50
120-200	0	5	29	60

b) ¿Qué porcentaje de viviendas tiene más de 120 m^2 y 3 dormitorios?

$$\frac{29}{500} = 0.058 \longrightarrow 5.8\,\%$$

c) Entre las viviendas con superficie entre 60 y 90 m^2, ¿qué porcentaje tiene más de 2 dormitorios?

$$\frac{120+22}{7+65+120+22} = \frac{142}{214} = 0.6636 \longrightarrow 66.36\,\%$$

Tabla de doble entrada: ejemplo

Ejemplo Una muestra de 500 viviendas en Galicia se ha clasificado según su superficie en m^2 (X) y el número de dormitorios (Y):

X/Y	1	2	3	4
0-60	9	23	6	0
60-90	7	65	120	22
90-120	2	15	87	50
120-200	0	5	29	60

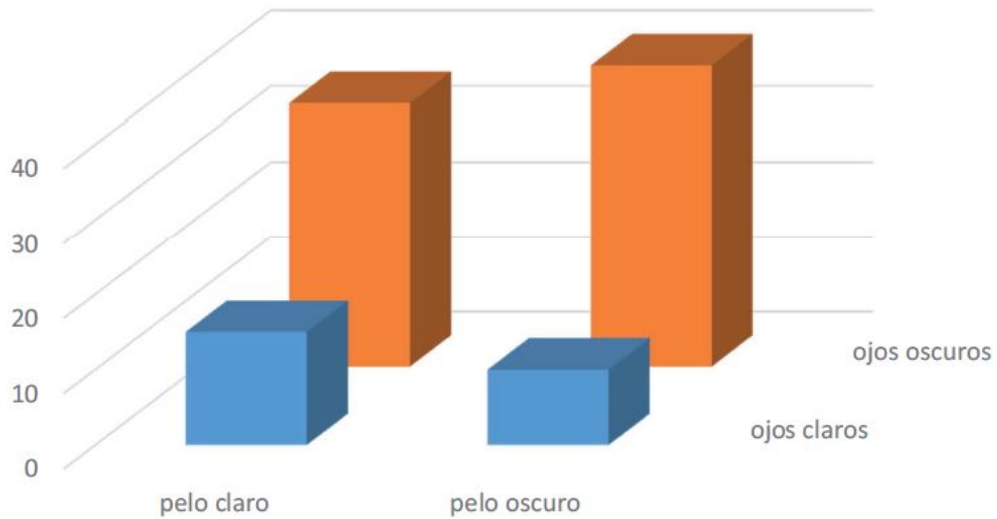
d) Entre las viviendas de menos de 3 dormitorios, ¿qué porcentaje tiene una superficie superior a 90 m^2?

$$\frac{2+15+5}{9+23+7+65+2+15+5} = \frac{22}{126} = 0.1746 \longrightarrow 17.46\,\%$$

e) Distribución de frecuencias de la variable X:

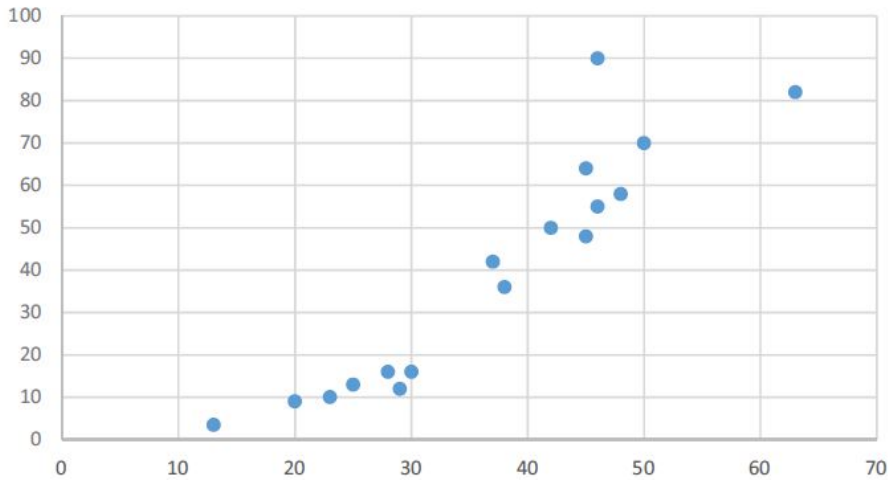
$L_{i-1} - L_i$	$n_{i\cdot}$	$f_{i\cdot}$
0-60	38	0.076
60-90	214	0.428
90-120	154	0.308
120-200	94	0.188
	500	1

- **Diagrama de barras**: Para variables cualitativas o cuantitativas sin agrupar.



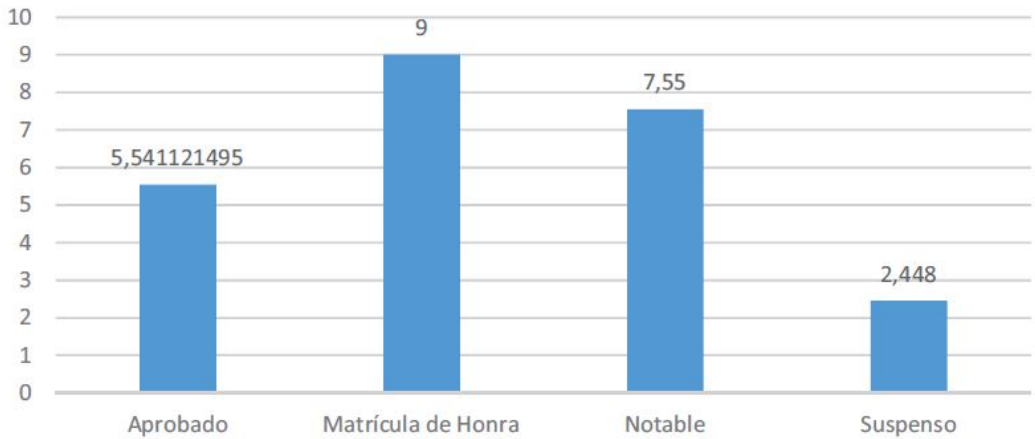
Representaciones gráficas

- **Diagrama de dispersión**: Para variables cuantitativas sin agrupar.



Representaciones gráficas

- **Gráficos de resumen**: Una variable es explicada en función de la otra.



Distribuciones marginales

A partir de la tabla de doble entrada podemos obtener las distribuciones de X o de Y.

X/Y	y_1	y_2	$\dots$	y_c	
x_1	n_{11}	n_{12}	$\dots$	n_{1c}	$n_{1\cdot}$
x_2	n_{21}	n_{22}	$\dots$	n_{2c}	$n_{2\cdot}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	
x_r	n_{r1}	n_{r2}	$\dots$	n_{rc}	$n_{r\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$		$n_{\cdot c}$	n

- La frecuencia marginal de x_i, $i = 1, \dots, r$:

$$n_{i\cdot} = n_{i1} + n_{i2} + \cdots + n_{ic} = \sum_{j=1}^{c} n_{ij}$$

(se corresponde con sumar por filas la tabla)

- La frecuencia marginal de y_j, $j = 1, \dots, c$:

$$n_{\cdot j} = n_{1j} + n_{2j} + \cdots + n_{rj} = \sum_{i=1}^{r} n_{ij}$$

(se corresponde con sumar por columnas la tabla)

Distribuciones marginales

Llamaremos distribuciones marginales a las distribuciones unidimensionales de frecuencias de las variables X e Y ; respectivamente:

$$(x_i; n_{i\cdot}),\ i = 1, 2, \ldots, r \text{ y } (y_j; n_{\cdot j}),\ j = 1, 2, \ldots, c.$$

- La frecuencia relativa marginal de X,

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

- La frecuencia relativa marginal de Y,

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

Tabla de doble entrada: ejemplo

a) Superficie media y mediana de las viviendas y varianza de la superficie (X):

$(L_{i-1}, L_i]$	$n_{i\cdot}$	c_i	$c_i n_{i\cdot}$	$N_{i\cdot}$	$c_i^2 n_{i\cdot}$
0-60	38	30	1140	38	34200
60-90	214	75	16050	252	1203750
90-120	154	105	16710	406	1697850
120-200	94	160	15040	500	2406400
	500		48400		5342200

Media:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{r} c_i n_{i\cdot} = \frac{48400}{500} = 96.8 \ m^2$$

Mediana:

$$\frac{n}{2} = 250 \rightarrow Me_x \in (60, 90]$$

$$Me_X = L_{i-1} + \frac{n/2 - N_{i-1}}{n_i} \times a_i = 60 + \frac{250 - 38}{214} \times 30 = 89.72 \ m^2$$

Varianza:

$$S_X^2 = \frac{1}{n}\sum_{i=1}^{r} c_i^2 n_{i\cdot} - \bar{x}^2 = \frac{5342200}{500} - 96.8^2 = 1314.16 (m^2)^2.$$

Tabla de doble entrada: ejemplo

b) Número medio y más frecuente de dormitorios en una vivienda y varianza del no de dormitorios (Y):

y_j	$n_{\cdot j}$	$y_j n_{\cdot j}$	$y_j^2 n_{\cdot j}$
1	18	18	18
2	108	216	432
3	242	726	2178
4	132	528	2112
	500	1488	4740

Media:

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{c} y_j n_{\cdot j} = \frac{1488}{500} = 2.976 \ m^2$$

Moda: $Mo_Y = 3$

Varianza:

$$S_Y^2 = \frac{1}{n}\sum_{j=1}^{c} y_j^2 n_{\cdot j} - \bar{y}^2 = \frac{4740}{500} - 2.976^2 = 0.6234$$

c) ¿Cuál de las dos variables presenta mayor dispersión?

$$CV_X = \frac{S_X}{\bar{x}} = \frac{36.251}{96.8} = 0.3745 \text{ y } CV_Y = \frac{S_Y}{\bar{y}} = \frac{0.7896}{2.976} = 0.2653$$

Ejemplo. Distribución de la superficie para las viviendas de 2 dormitorios:

$X \mid Y=2$	$n_{i \mid Y=2}$	$f_{i \mid Y=2} = \frac{n_{i \mid Y=2}}{108}$
0-60	23	0.213
60-90	65	0.602
90-120	15	0.139
120-200	5	0.0462
	108	1

Ejemplo. Distribución de la superficie para las viviendas de 3 dormitorios:

$X \mid Y=3$	$n_{i \mid Y=3}$	$f_{i \mid Y=3} = \frac{n_{i \mid Y=3}}{242}$
0-60	6	0.025
60-90	120	0.496
90-120	87	0.360
120-200	29	0.120
	242	1

Distribuciones condicionadas

Distribución de X condicionada al valor y_j de la variable Y $(X|Y = y_j)$

- Se representa por $(x_i; n_{i|Y=y_j} = n_{ij})$, con $i = 1, 2, \ldots, r$.
- Número total de observaciones: $\sum_{i=1}^{r} = n_{\cdot j}$
- Frecuencias relativas: $f_{i|Y=y_j} = \frac{n_{i|Y=y_j}}{n_{\cdot j}} = \frac{n_{ij}}{n_{\cdot j}}$

Distribución de Y condicionada al valor x_i de la variable Y $(Y|X = x_i)$

- Se representa por $(y_j; n_{j|X=x_i} = n_{ij})$, con $i = 1, 2, \ldots, c$.
- Número total de observaciones: $\sum_{j=1}^{c} = n_{i\cdot}$
- Frecuencias relativas: $f_{i|Y=y_j} = \frac{n_{j|X=x_i}}{n_{i\cdot}} = \frac{n_{ij}}{n_{i\cdot}}$

Distribuciones condicionadas

- Pueden definirse también distribuciones condicionadas a un conjunto o intervalo de valores; por ejemplo, la distribución de $X|Y \leq y_j$ o la de $Y|X > x_i$.
- En las distribuciones condicionadas el estudio se reduce a la parte de la tabla determinada por la condición.
- Las distribuciones condicionadas son distribuciones unidimensionales.

Ejemplo Número medio de dormitorios en las viviendas de más de 90 m^2, esto es, se trata de la distribución de $Y|X > 90$.

y_j	$n_{j\|X>90}$	$y_j n_{j\|X>90}$
1	2	2
2	20	40
3	116	348
4	110	440
	248	830

Media:

$\bar{y}_{|X>90} = \frac{1}{248}\sum_{j=1}^{c} y_j n_{j|X>90} = \frac{830}{248} = 3.347$, que es mayor que la media global (2.976).

Independencia Estadística

Dos variables X, Y se dice que son independientes estadísticamente si el comportamiento de una de ellas no se ve afectado por los valores que toma la otra, es decir:

$$f_{i|Y=y_j} = f_{i\cdot} \text{ para cualquier par de valores } (x_i, y_j)$$

$$y$$

$$f_{j|X=x_i} = f_{\cdot j} \text{ para cualquier par de valores } (x_i, y_j).$$

Equivalentemente, X e Y son independientes si y sólo si

$$f_{ij} = f_{i\cdot} f_{\cdot j} \text{ y } n_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n} \text{ para todo } i, j$$

Independencia Estadística: ejemplo

X/Y	y_1	y_2	y_3	$n_{i\cdot}$	$f_{i\cdot}$
x_1	1	3	5	9	9/54=1/6
x_2	2	6	10	18	18/54=1/3
x_3	3	9	15	27	27/54=1/2
$n_{\cdot j}$	6	18	30	$n = 54$	

Tenemos que ver que $f_{i|Y=y_j} = f_{i\cdot}$ con $i, j = 1, 2, 3$.

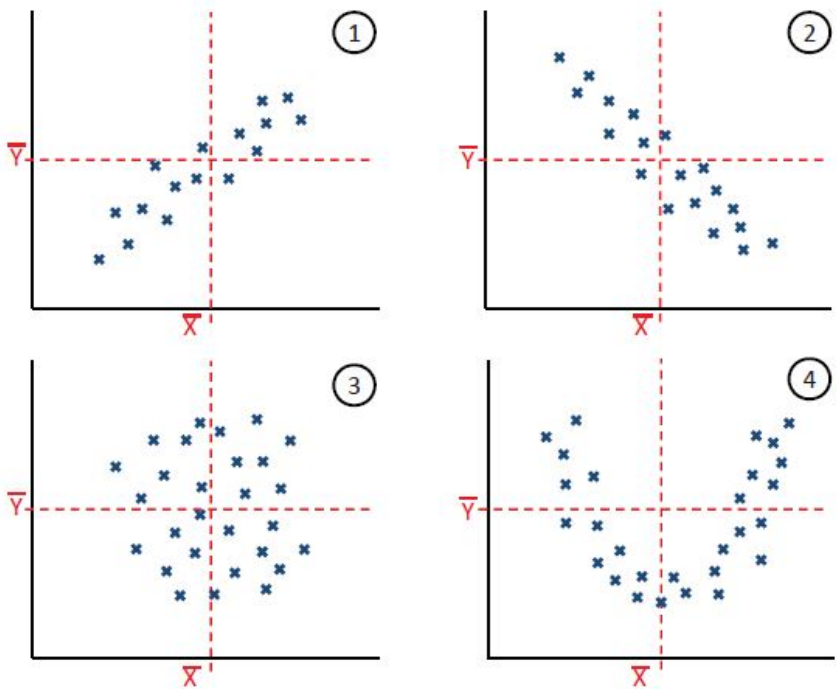
X	$n_{i\|Y=y_1}$	$f_{i\|Y=y_1}$	$n_{i\|Y=y_2}$	$f_{i\|Y=y_2}$	$n_{i\|Y=y_3}$	$f_{i\|Y=y_3}$
x_1	1	1/6	3	3/18=1/6	5	5/30=1/6
x_2	2	2/6=1/3	6	6/18=1/3	10	10/30=1/3
x_3	3	3/6=1/2	9	9/18=1/2	15	15/30=1/2
	$n_{\cdot 1} = 6$		$n_{\cdot 2} = 18$		$n_{\cdot 3} = 30$	

Covarianza: asociación entre variables cuantitativas

En caso de que las variables no sean independientes, vamos a estudiar cómo medir la posible relación **lineal** entre ambas.

Covarianza: $S_{XY} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{c}(x_i - \overline{x})(y_j - \overline{y})n_{ij}$, para datos tabulados

$$S_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_j - \overline{y}), \text{ para } n \text{ pares de datos}$$



1. Relación directa: $S_{XY} > 0$
2. Relación inversa: $S_{XY} < 0$
3. Independientes: $S_{XY} \approx 0$
4. Sin relación lineal: $S_{XY} \approx 0$

Covarianza: propiedades

1. $S_{XX} = S_X^2$.
2. Si X e Y son independientes, entonces la covarianza es cero. El recíproco no es cierto en general.
3. Si $S_{XY} > 0$, existe una relación lineal positiva: las variables varían en en el mismo sentido.
4. Si $S_{XY} < 0$, existe una relación lineal negativa: las variables varían en sentido contrario.
5. $S_{XY} = \left(\frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{c} x_i y_j n_{ij} \right) - \overline{x} \cdot \overline{y}$, para datos tabulados.

   $S_{XY} = \frac{1}{n} \sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$, para n pares de datos
6. Depende de las unidades de medida y no está acotada.

Coeficiente de correlación: asociación entre variables cuantitativas

¿Cómo sabemos si la relación lineal existente entre las variables es intensa o no?

Usaremos el **coeficiente de correlación de Pearson**:

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

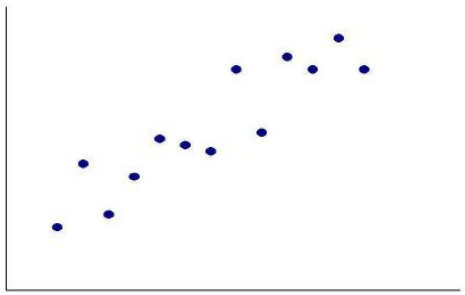
Propiedades

1. Es una medida adimensional.
2. $-1 \leq r_{XY} \leq 1$.
3. $r_{aX+b,cY+d} = r_{XY}$.
4. Cuando $r_{XY} = 0$ se dice que las variables están **incorreladas**: no existe relación lineal entre ellas.

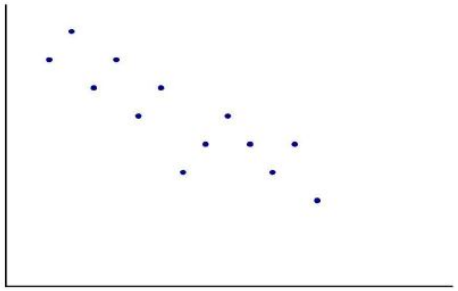
Más comentarios

- Su signo coincide con el de la covarianza.
- Si X e Y son independientes, $r_{XY} = 0$. El recíproco no es cierto en general.
- Si $r_{XY} \neq 0$ existe asociación lineal, más fuerte cuanto más se acerque el coeficiente a 1 o a -1.

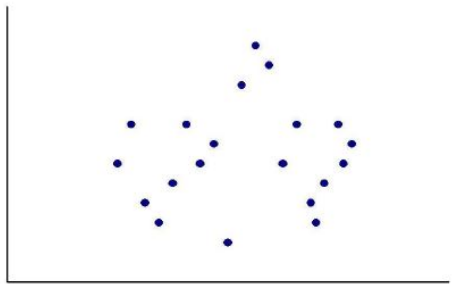
Coeficiente de correlación



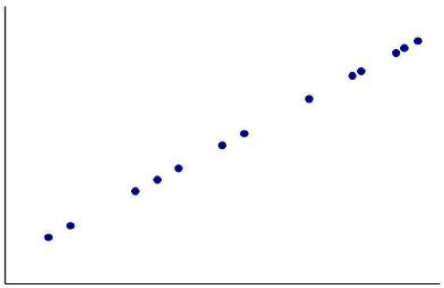
$r_{XY} > 0$



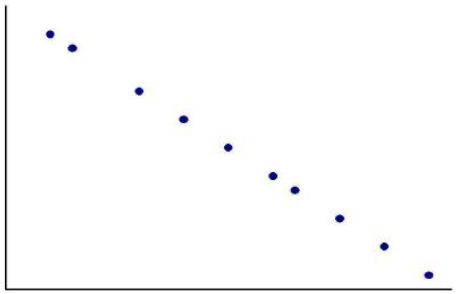
$r_{XY} < 0$



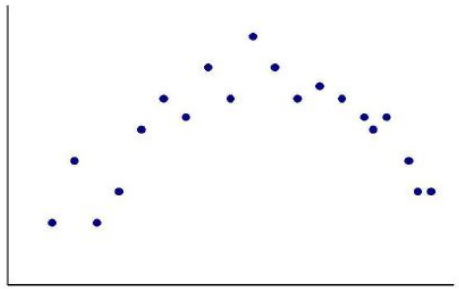
r_{XY} cercano a 0



$r_{XY} = 1$



$r_{XY} = -1$



r_{XY} cercano a 0

Ejercicio

En la tabla conjunta se tiene información sobre la edad (X) y el salario por hora (en euros) (Y) de un grupo de trabajadores:

$X \setminus Y$	6-10	10-14	14-20
20-30	22	10	0
30-40	14	24	8
40-50	5	17	20

1. Determina la edad más frecuente.
2. Calcula el salario medio de los trabajadores mayores de 30 años.
3. Si nos restringimos a aquellos trabajadores que que cobran más de 10 euros por hora ¿qué porcentaje de ellos tiene más de 27 años?
4. ¿Son independientes ambas variables?

Coeficiente χ^2 de Pearson

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - \frac{n_{i\cdot}\, n_{\cdot j}}{n})^2}{\frac{n_{i\cdot}\, n_{\cdot j}}{n}} = n\left(\sum_{i=1}^{r}\sum_{j=1}^{c} \frac{n_{ij}^2}{n_{i\cdot}\, n_{\cdot j}} - 1\right)$$

- Toma valores entre 0 y $n(\text{mín}\{r, c\} - 1)$.
- Si X e Y son independientes, entonces $\chi^2 = 0$.
- Si $\chi^2 \neq 0$, existe asociación entre las variables, de mayor intensidad cuanto más alto sea su valor.

Coeficiente V de Crámer

- Se define como $V = \sqrt{\frac{\chi^2}{n(\text{mín}\{r,c\}-1)}}$.
- $0 \leq V \leq 1$.
- Mayor grado de dependencia o asociación cuanto más próximo a 1.

Asociación entre variables cualitativas: ejemplo

De la población adulta con edad comprendida entre 25 y 65 años se ha seleccionado una muestra de 500 personas clasificándolas según la frecuencia de asistencia al cine y el nivel educativo (educación superior o no):

	Superior	No superior
Cada semana	8	2
Cada mes	35	18
Alguna vez al año	77	83
Nunca	35	242

Vamos a analizar la asociación utilizando el coeficiente V de Crámer, para lo cual necesitamos calcular el coeficiente χ^2 de Pearson.

Asociación entre variables cualitativas: ejemplo

	Superior	No superior	$n_{i\cdot}$
Cada semana	8	2	10
Cada mes	35	18	53
Alguna vez al año	77	83	160
Nunca	35	242	277
$n_{\cdot j}$	155	345	500

$$\chi^2 = n\left(\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{n_{ij}^2}{n_{i\cdot}\, n_{\cdot}} - 1\right) = 500\left(\frac{8^2}{10\times 155} + \frac{2^2}{10\times 345} + \cdots + \frac{242^2}{277\times 345} - 1\right) = 107.255$$

$$V = \sqrt{\frac{\chi^2}{n(\text{mín}\{r,c\}-1)}} = \sqrt{\frac{107.255}{500\times 1}} = 0.463$$