

Relación de Problemas 2: Variables estadísticas bidimensionales

Estadística Descriptiva e Introducción a la Probabilidad

Primer curso del Doble Grado en Ingeniería Informática y Matemáticas

1. Se han lanzado dos dados varias veces, obteniendo los resultados que se presentan en la siguiente tabla, donde X designa el resultado del primer dado e Y el resultado del segundo:

X	1	2	2	3	5	4	1	3	3	4	1	2	5	4	3	4	4	5	3	1	6	5	4	6
Y	2	3	1	4	3	2	6	4	1	6	6	5	1	2	5	1	1	2	6	6	2	1	2	5

- a) Construir la tabla de frecuencias.
- b) Calcular las puntuaciones medias obtenidas con cada dado y ver cuales son más homogéneas.
- c) ¿Qué resultado del segundo dado es más frecuente cuando en el primero se obtiene un 3?
- d) Calcular la puntuación máxima del 50 % de las puntuaciones más bajas obtenidas con el primer dado si con el segundo se ha obtenido un 2 o un 5.
2. Medidos los pesos, X (en Kg), y las alturas, Y (en cm), a un grupo de individuos, se han obtenido los siguientes resultados:

$X \setminus Y$	160	162	164	166	168	170
48	3	2	2	1	0	0
51	2	3	4	2	2	1
54	1	3	6	8	5	1
57	0	0	1	2	8	3
60	0	0	0	2	4	4

- a) Calcular el peso medio y la altura media y decir cuál es más representativo.
- b) Calcular el porcentaje de individuos que pesan menos de 55 Kg y miden más de 165 cm.
- c) Entre los que miden más de 165 cm, ¿cuál es el porcentaje de los que pesan más de 52 Kg?
- d) ¿Cuál es la altura más frecuente entre los individuos cuyo peso oscila entre 51 y 57 Kg?
- e) ¿Qué peso medio es más representativo, el de los individuos que miden 164 cm o el de los que miden 168 cm?
3. En una encuesta de familias sobre el número de individuos que la componen (X) y el número de personas activas en ellas (Y) se han obtenido los siguientes resultados:

$X \setminus Y$	1	2	3	4
1	7	0	0	0
2	10	2	0	0
3	11	5	1	0
4	10	6	6	0
5	8	6	4	2
6	1	2	3	1
7	1	0	0	1
8	0	0	1	1

- a) Calcular la recta de regresión de Y sobre X .
- b) ¿Es adecuado suponer una relación lineal para explicar el comportamiento de Y a partir de X ?

4. Se realiza un estudio sobre la tensión de vapor de agua (Y , en ml. de Hg.) a distintas temperaturas (X , en $^{\circ}\text{C}$). Efectuadas 21 medidas, los resultados son:

$X \setminus Y$	(0.5, 1.5]	(1.5, 2.5]	(2.5, 5.5]
(1, 15]	4	2	0
(15, 25]	1	4	2
(25, 30]	0	3	5

Explicar el comportamiento de la tensión de vapor en términos de la temperatura mediante una función lineal. ¿Es adecuado asumir este tipo de relación?

5. Estudiar la dependencia o independencia de las variables en cada una de las siguientes distribuciones. Dar, en cada caso, las curvas de regresión y la covarianza de las dos variables.

$X \setminus Y$	1	2	3	4	5	$X \setminus Y$	1	2	3
10	2	4	6	10	8	-1	0	1	0
20	1	2	3	5	4	0	1	0	1
30	3	6	9	15	12	1	0	1	0
40	4	8	12	20	16				

6. Dada la siguiente distribución bidimensional:

$X \setminus Y$	1	2	3	4
10	1	3	0	0
12	0	1	4	3
14	2	0	0	2
16	4	0	0	0

- a) ¿Son estadísticamente independientes X e Y ?
 b) Calcular y representar las curvas de regresión de X/Y e Y/X .
 c) Cuantificar el grado en que cada variable es explicada por la otra mediante la correspondiente curva de regresión.
 d) ¿Están X e Y correladas linealmente? Dar las expresiones de las rectas de regresión.

7. Para cada una de las distribuciones:

Distribución A			Distribución B			Distribución C						
$X \setminus Y$	10	15	20	$X \setminus Y$	10	15	20	$X \setminus Y$	10	15	20	25
1	0	2	0	1	0	2	0	1	0	3	0	1
2	1	0	0	2	1	0	0	2	0	0	1	0
3	0	0	3	3	0	0	3	3	2	0	0	0
4	0	1	0									

- a) ¿Dependen funcionalmente X de Y o Y de X ?
 b) Calcular las curvas de regresión y comentar los resultados.
8. De una muestra de 24 puestos de venta en un mercado de abastos se ha recogido información sobre el número de balanzas (X) y el número de dependientes (Y). Los resultados aparecen en la siguiente tabla:

$X \setminus Y$	1	2	3	4
1	1	2	0	0
2	1	2	3	1
3	0	1	2	6
4	0	0	2	3

- a) Determinar las rectas de regresión.
- b) ¿Es apropiado suponer que existe una relación lineal entre las variables?
- c) Predecir, a partir de los resultados, el número de balanzas que puede esperarse en un puesto con seis dependientes. ¿Es fiable esta predicción?
9. Se eligen 50 matrimonios al azar y se les pregunta la edad de ambos al contraer matrimonio. Los resultados se recogen en la siguiente tabla, en la que X denota la edad del hombre e Y la de la mujer:

$X \setminus Y$	(10, 20]	(20, 25]	(25, 30]	(30, 35]	(35, 40]
(15, 18]	3	2	3	0	0
(18, 21]	0	4	2	2	0
(21, 24]	0	7	10	6	1
(24, 27]	0	0	2	5	3

Estudiar la interdependencia lineal entre ambas variables.

10. Calcular el coeficiente de correlación lineal de dos variables cuyas rectas de regresión son:

$$x + 4y = 1$$

$$x + 5y = 2$$

11. Consideremos una distribución bidimensional en la que la recta de regresión de Y sobre X es $y = 5x - 20$, y $\sum y_j^2 n_{.j} = 3240$. Supongamos, además, que la distribución marginal de X es:

x_i	3	5	8	9
n_i	5	1	2	1

Determinar la recta de regresión de X sobre Y , y la bondad de los ajustes lineales.

12. De las estadísticas de "Tiempos de vuelo y consumos de combustible" de una compañía aérea, se han obtenido datos relativos a 24 trayectos distintos realizados por el avión DC-9. A partir de estos datos se han obtenido las siguientes medidas:

$$\begin{aligned} \sum y_i &= 219.719 & \sum y_i^2 &= 2396.504 & \sum x_i y_i &= 349.486 \\ \sum x_i &= 31.470 & \sum x_i^2 &= 51.075 & \sum x_i^2 y_i &= 633.993 \\ && \sum x_i^4 &= 182.977 & \sum x_i^3 &= 93.6 \end{aligned}$$

La variable Y expresa el consumo total de combustible, en miles de libras, correspondiente a un vuelo de duración X (el tiempo se expresa en horas, y se utilizan como unidades de orden inferior fracciones decimales de la hora).

- a) Ajustar un modelo del tipo $Y = aX + b$. ¿Qué consumo total se estimaría para un programa de vuelos compuesto de 100 vuelos de media hora, 200 de una hora y 100 de dos horas? ¿Es fiable esta estimación?
- b) Ajustar un modelo del tipo $Y = a + bX + cX^2$. ¿Qué consumo total se estimaría para el mismo programa de vuelos del apartado a)?
- c) ¿Cuál de los dos modelos se ajusta mejor? Razonar la respuesta.
13. La curva de Engel, que expresa el gasto en un determinado bien en función de la renta, adopta en ocasiones la forma de una hipérbola equilátera. Ajustar dicha curva a los siguientes datos, en los que X denota la renta en miles de euros e Y el gasto en euros. Cuantificar la bondad del ajuste:

X	10	12.5	20	25
Y	50	90	160	180

14. Se dispone de la siguiente información referente al gasto en espectáculos (Y , en euros) y la renta disponible mensual (X , en cientos de euros) de 6 familias:

Y	30	50	70	80	120	140
X	9	10	12	15	22	32

Explicar el comportamiento de Y por X mediante:

- a) Relación lineal.
- b) Hipérbola equilátera.
- c) Curva potencial.
- d) Curva exponencial.

¿Qué ajuste es más adecuado?

1. Se han lanzado dos dados varias veces, obteniendo los resultados que se presentan en la siguiente tabla, donde X designa el resultado del primer dado e Y el resultado del segundo:

X	1	2	2	3	5	4	1	3	3	4	1	2	5	4	3	4	4	5	3	1	6	5	4	6
Y	2	3	1	4	3	2	6	4	1	6	6	5	1	2	5	1	1	2	6	6	2	1	2	5

- a) Construir la tabla de frecuencias.
- b) Calcular las puntuaciones medias obtenidas con cada dado y ver cuáles son más homogéneas.
- c) ¿Qué resultado del segundo dado es más frecuente cuando en el primero se obtiene un 3?
- d) Calcular la puntuación máxima del 50 % de las puntuaciones más bajas obtenidas con el primer dado si con el segundo se ha obtenido un 2 o un 5.

a)

x_i	1	2	3	4	5	6	$n_{i,j}$	$n_{i,\cdot}$	$n_{\cdot j}$
1	0	1	0	0	0	3	4	4	
2	1	0	1	0	1	0	3	6	
3	1	0	0	2	1	1	4	12	
4	2	3	0	0	0	1	6	24	
5	2	1	1	0	0	0	4	20	
6	0	1	0	0	1	0	2	12	
$n_{\cdot j}$	6	6	2	2	3	5	24		
$n_{\cdot \cdot}$	6	12	6	8	15	30			

0) $\bar{x} = \frac{1}{n} \sum n_{i,j} \cdot x_i = \frac{4+6+12+24+20+12}{24} = 3,25$

$$\bar{y} = \frac{1}{n} \sum n_{j,j} \cdot y_j = \frac{6+12+6+8+15+30}{24} = 3,2083$$

Ahora calc. var Pearson

c) Tel u.

d) Percentil 35% y 75%

2. Medidos los pesos, X (en Kg), y las alturas, Y (en cm), a un grupo de individuos, se han obtenido los siguientes resultados:

$X \setminus Y$	160	162	164	166	168	170
48	3	2	2	1	0	0
51	2	3	4	2	2	1
54	1	3	6	8	5	1
57	0	0	1	2	8	3
60	0	0	0	2	4	4

- a) Calcular el peso medio y la altura media y decir cuál es más representativo.
- b) Calcular el porcentaje de individuos que pesan menos de 55 Kg y miden más de 165 cm.
- c) Entre los que miden más de 165 cm, ¿cuál es el porcentaje de los que pesan más de 52 Kg?
- d) ¿Cuál es la altura más frecuente entre los individuos cuyo peso oscila entre 51 y 57 Kg?
- e) ¿Qué peso medio es más representativo, el de los individuos que miden 164 cm o el de los que miden 168 cm?

vo tiene
ver geo'ser
porcentual

$$G_x^2 = m_2 - m_1^2 = \frac{\sum}{n} - \bar{x}^2 \rightarrow G_x = \sqrt{G_x^2}$$

$X \setminus Y$	160	162	164	166	168	170	m_i	$u_i \cdot x x_i$	$u_i(x_i - \bar{x})^2$
48	3	2	2	1	0	0	8	384	304,65
51	2	3	4	2	2	1	14	714	140,77
54	1	3	6	8	5	1	24	1296	0,70
57	0	0	1	2	8	3	14	798	112,04
60	0	0	0	2	4	4	10	600	332,77
m_j	6	8	13	15	19	9			

$$\sum u_i \cdot x x_i = 960 + 1296 + 2132 + 2400 + 3192 + 1830$$

$$\sum u_i (x_i - \bar{x})^2 = 195,89 + 110,35 + 38,19 + 1,22 + 99,29 + 165,32$$

$$\bar{x} = \frac{384 + 714 + 1296 + 798 + 600}{70} = 54,171 \text{ kg}$$

$$\bar{y} = \frac{960 + 1296 + 2132 + 2400 + 3192 + 1830}{70} = 165,714 \text{ cm}$$

Para ver cuál es más representativa hay el coeficiente de variación de Pearson $C.V(x) = \frac{G_x}{\bar{x}}$

$$C.V(y) = \frac{G_y}{\bar{y}}$$

$$G_x^2 = \frac{304,65 + 140,77 + 0,70 + 112,04 + 332,77}{70} = 12,82$$

$$G_x = 3,58 \sim \text{repetir cálculo}$$

$$G_y^2 = \frac{195,89 + 110,35 + 38,19 + 1,22 + 99,29 + 165,32}{70} = 8,718$$

$$G_y = 2,95 \sim \text{más representativa}$$

$$e) \frac{1+2+1+2+8+5+1}{70} \cdot 100 = 28,57\%$$

c) Alors ob tang ma dimension

$$\bar{x}/y > 1,6 \text{ sm}$$

x_i	n_i
48	1
51	5
54	14
57	13
60	10
	43

$$\frac{37}{43} \cdot 100 = 86,04\%$$

d) $y / s_1 \leq x \leq s_7$

y_j	n_j
160	3
162	6
164	11
166	12
168	15
170	5

$$\bar{m}_0 = 168 \text{ cm}$$

e)

3. En una encuesta de familias sobre el número de individuos que la componen (X) y el número de personas activas en ellas (Y) se han obtenido los siguientes resultados:

$X \setminus Y$	1	2	3	4
1	7	0	0	0
2	10	2	0	0
3	11	5	1	0
4	10	6	6	0
5	8	6	4	2
6	1	2	3	1
7	1	0	0	1
8	0	0	1	1

$n_{ij} \neq \frac{n_i \cdot n_j}{n}$
 Si hay o no hay
 independencia

a) Calcular la recta de regresión de Y sobre X .

b) ¿Es adecuado suponer una relación lineal para explicar el comportamiento de Y a partir de X ?

$$y - \bar{y} = \frac{6xy}{6x} (x - \bar{x})$$

$$0.3145x + 0.5331$$

a) Si fueran independientes no tiene sentido hacer recta de regresión.

$$y = ax + b \rightarrow a = \frac{6xy}{6x^2} \quad b = \bar{y} - a\bar{x}$$

$$6xy = m_{11} - \bar{x}\bar{y}$$

Esto es lo único bidimensional

$$m_{11} = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^k n_{ij} x_i y_0 = A/n$$

$X \setminus Y$	1	2	3	4	$n_{..}$	$\sum_{i=1}^k n_{i..} x_i$	$\sum_{i=1}^k \sum_{j=1}^n n_{ij} x_i y_0$	$\sum_{i=1}^k n_{i..} (x_i - \bar{x})^2$
0,845 1	7	0	0	0	7	7	7	56,535
1,16 2	10	2	0	0	12	24	28	40,71
1,415 3	11	5	1	0	17	51	72	12,03
1,79 4	10	6	6	0	22	88	160	0,549
2,105 5	8	6	4	2	20	100	200	26,81
2,42 6	1	2	3	1	7	42	108	32,59
2,735 7	1	0	0	1	2	14	35	19,94
3,05 8	0	0	1	1	2	16	56	34,89
					89	$\bar{x} = 3,842$	748	$6x^2 = 2,81$
	$n_{..}$	48	21	15	5	89		
	$\sum_{i=1}^k n_{i..} x_i$	48	42	15	20	$\bar{y} = 1,741$		
	$n_{..} y_0 - \bar{y}$	26,355	1,408	13,776	2,5515	$6y^2 = 0,865$		

$$r^2 = \frac{(6xy)^2}{6x^2 \cdot 6y^2} =$$

↓ no es adecuado

$$6xy = m_{11} - \bar{x}\bar{y} = 0,791$$

$$a = \frac{0,791}{2,81} = 0,315$$

$$y = 0,315x + 0,5331$$

1) Si x, y indep

$$i - \overline{O_{xy}} = 0$$

$$ii - y = \bar{y}; \quad x = \bar{x}$$

$$iii - m_{rs} = m_{r_0} \cdot m_{os}$$

$$\mu_{rs} = \mu_{ro} \cdot \mu_{os}$$

4. Se realiza un estudio sobre la tensión de vapor de agua (Y , en ml. de Hg.) a distintas temperaturas (X , en °C). Efectuadas 21 medidas, los resultados son:

$X \setminus Y$	(0.5, 1.5]	(1.5, 2.5]	(2.5, 5.5]
(1, 15]	4	2	0
(15, 25]	1	4	2
(25, 30]	0	3	5

Explicar el comportamiento de la tensión de vapor en términos de la temperatura mediante una función lineal. ¿Es adecuado asumir este tipo de relación?

$X \setminus Y$	(0.5, 1.5]	(1.5, 2.5]	(2.5, 5.5]	n_i	$\frac{1}{n} \sum_{i=1}^k x_i n_i$	$\frac{1}{n} \sum_{i=1}^k n_i (x_i - \bar{x})^2$	$\frac{1}{n} \sum_{i,j} n_{ij} x_i y_j$
8 (1, 15]	4	2	0	6	4.8	782.498	64
20 (15, 25]	1	4	2	7	14.0	2.354	0.0
27,5 (25, 30]	0	3	5	8	22.0	522.201	71.0
	5	9	7	21	19.41	$6_x^2 - 62.20$	53.23
	5	18	28		2.42	$6_y^2 = 1.387$	
	$\frac{1}{n} \sum_{i,j} n_{ij} y_j$	$\frac{1}{n} \sum_{i,j} n_{ij} (y_j - \bar{y})^2$	$\frac{1}{n} \sum_{i,j} n_{ij} (x_i - \bar{x})(y_j - \bar{y})$				
	5	18	28		10.082	1.887	12.470

$$6_{xy} = 53.23 - 2.42 \cdot 19.41 = 6.233$$

$$a = \frac{6^2_{xy}}{6^2_x} = 0.1$$

$$y = 0.1x + 0.478$$

$$r = \bar{y} - a\bar{x} = 0.478$$

$$r^2 = 0.45 \rightarrow \text{no es adecuado}$$

5. Estudiar la dependencia o independencia de las variables en cada una de las siguientes distribuciones. Dar, en cada caso, las curvas de regresión y la covarianza de las dos variables.

$X \setminus Y$	1	2	3	4	5
10	2	4	6	10	8
20	1	2	3	5	4
30	3	6	9	15	12
40	4	8	12	20	16

$X \setminus Y$	1	2	3
-1	0	1	0
0	1	0	1
1	0	1	0

150

Se dice que el carácter Y es independiente condicionado del carácter X , si las distribuciones de Y condicionadas a cada valor de X , son idénticas para todos x_i ; $i=1, 2, \dots, k$.

$$g_j^i = g_j / i$$

$$\frac{n_{ij}}{n_{i \cdot}} = \frac{n_{2j}}{n_{2 \cdot}} = \dots = \frac{n_{ij}}{n_{i \cdot}} = \dots = \frac{n_{kj}}{n_{k \cdot}} \quad \forall j = 1, 2, \dots, p$$

$$j=1 \quad \frac{2}{30} = \frac{1}{15} = \frac{3}{45} = \frac{4}{60}$$

$$j=2 \quad \frac{4}{30} = \frac{2}{15} = \frac{6}{45} = \frac{8}{60}$$

$$j=3 \quad \frac{6}{30} = \frac{3}{15} = \frac{9}{45} = \frac{12}{60}$$

$$j=4 \quad \frac{10}{30} = \frac{5}{15} = \frac{15}{45} = \frac{20}{60}$$

$$j=5 \quad \frac{8}{30} = \frac{4}{15} = \frac{12}{45} = \frac{16}{60}$$

Y es independiente del carácter X , y por tanto, al reciproco también es cierto.

No tiene sentido estudiar curvas de regresión si son independientes.

$$\sigma_{xy} = 0$$

$X \setminus Y$	1	2	3
-1	0	1	0
0	1	0	1
1	0	1	0

Es evidente que no son independientes puesto que si algún $n_{ij} = 0$, la igualdad

$$\frac{n_{ij}}{n_{i\cdot}} = \frac{n_{2j}}{n_2\cdot} = \dots = \frac{n_{ij}}{n_{i\cdot}} = \dots = \frac{n_{xj}}{n_k\cdot} \quad f_j = 1, 2, \dots, p$$

No se va a dar a veces que $n_{ij} = 0 \neq t_{ij}$, lo cual no tendría sentido porque sería establecer una variable que no ha presentado ninguna distribución de frecuencia.

Observando ahora la posible dependencia funcional, vemos que X , no depende funcionalmente de Y porque a la modalidad 2 de Y , le corresponden dos posibles modalidades de X (señalado arriba).

Calculamos entonces la covarianza

$$6_{xy} = m_{11} - \bar{x}\bar{y} \quad \text{Como } \bar{x} = 0, \text{ la covarianza será } 6_{xy} = m_{11}$$

$X \setminus Y$	1	2	3	$m_{11} = \frac{1}{n} \sum \sum n_{ij} x_i y_j$
-1	0	1	0	-1
0	1	0	1	0
1	0	1	0	2

$$6_{xy} = 0$$

Ahora continuemos la cerca de regresión de tipo I:

- Pasa por (x_i, \bar{y}_i) $i = 1, \dots, n$

$$\text{Punto 1} : (-1, 2)$$

$$\text{" 2 : } (0, 2)$$

$$\text{" 3 : } (1, 2)$$

10. Calcular el coeficiente de correlación lineal de dos variables cuyas rectas de regresión son:

$$x + 4y = 1$$

$$x + 5y = 2$$

Supongamos que $x + 4y = 1 \Rightarrow$ la recta para $x/4$ y a $x + 5y = 2$ para $x/5$

Además, sabemos que ambas rectas tienen el mismo signo en la pendiente

$$(i) x = -4y + 1 \quad (ii) y = -\frac{1}{5}x + \frac{2}{5}$$

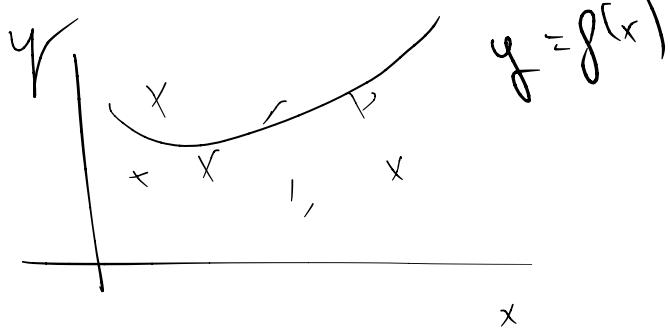
$$a_1 = \frac{6xy}{6^2y} < 0 \quad a_2 = \frac{6xy}{6^2x} < 0$$

$$a_1 \cdot a_2 = r^2 = \frac{6xy^2}{6^2 \cdot 6^2 y} = \frac{4}{5}$$

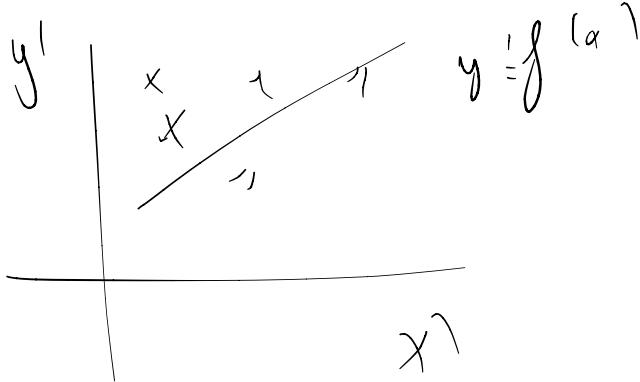
$$\text{Nos piden } r = \sqrt{r^2} = -\frac{\sqrt{4}}{\sqrt{5}} = -\frac{2}{\sqrt{5}}$$

Dicho resultado tiene sentido porque $-1 \leq -\frac{2}{\sqrt{5}} \leq 1$

Último problema



$$y = f(x)$$



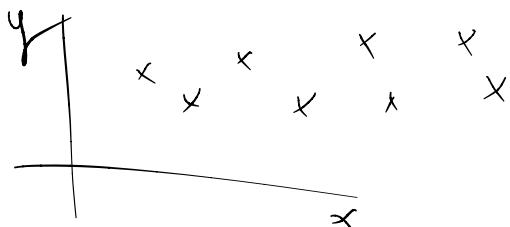
$$y = f^{(\alpha)}$$

Como compare la /cordad

$$y - y/x^2$$

$$y^2 - y/x^2 = r^2$$

Esto en el último problema no hace falta



La más pequeña varianza residual, la variante de y no varía · la menor es la preferida