1. Se han estudiado los pesos (X) en Kg y la altura (Y) en cm de un grupo de personas, obteniéndose la información dada en la siguiente tabla. Se pide:

a) El peso y la altura media.

b) ¿Cuál es la altura más frecuente entre las personas cuyo peso oscila entre 51 Kg y 57 Kg?

c) Obtener el peso que es superado por el 70% de las personas que miden más de 165 cm.

d) ¿Qué peso medio es más representativo: el de las personas que miden 164 cm o el de las que miden 168 cm?

X \ 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) Peso y la altura media

X \ Y	160	162	164	166	168	170	Marginal X
48	3	2	2	1	0	0	8
51	2	3	4	2	2	1	14
54	1	3	6	8	5	1	24
57	0	0	1	2	8	3	14
60	0	0	0	2	4	4	10
Marginal Y	6	8	13	15	19	9	70

- Peso medio = Media de la D. Marginal de X

x_i	n_i	$n_i x_i$
48	8	384
51	14	714
54	24	1296
57	14	798
60	10	600
	70	3792

$$Peso\ Medio: \quad \bar{x} = \frac{\sum_{i=1}^{k} n_i x_i}{n} = \frac{3792}{70} = 54.171$$

a) Peso y altura media

X \ Y	160	162	164	166	168	170	Marginal X
48	3	2	2	1	0	0	8
51	2	3	4	2	2	1	14
54	1	3	6	8	5	1	24
57	0	0	1	2	8	3	14
60	0	0	0	2	4	4	10
Marginal Y	6	8	13	15	19	9	70

- Altura media = Media de la D. Marginal de Y

y_i	n_i	$n_i y_i$
160	6	960
162	8	1296
164	13	2132
166	15	2490
168	19	3192
170	9	1530
	70	11600

$$Estatura\ Media: \quad \bar{y} = \frac{\sum_{i=1}^{k} n_i y_i}{n} = \frac{11600}{70} = 165.714$$

b) ¿Cuál es la altura más frecuente entre las personas cuyo peso oscila entre 51 Kg y 57 Kg?

X \ Y	160	162	164	166	168	170	Marginal X
48	3	2	2	1	0	0	8
51	2	3	4	2	2	1	14
54	1	3	6	8	5	1	24
57	0	0	1	2	8	3	14
60	0	0	0	2	4	4	10
Marginal Y	6	8	13	15	19	9	70

- Moda de la Distribución Y / $51 \leq X \leq 57$

y_i	n_i
160	3
162	6
164	11
166	12
168	15
170	5

Moda = 168

c) Obtener el peso que es superado por el 70% de las personas que miden más de 165 cm.

X \ Y	160	162	164	166	168	170	Marginal X
48	3	2	2	1	0	0	8
51	2	3	4	2	2	1	14
54	1	3	6	8	5	1	24
57	0	0	1	2	8	3	14
60	0	0	0	2	4	4	10
Marginal Y	6	8	13	15	19	9	70

➢ Percentil 30 de la Distribución de X / Y > 165

x_i	n_i	N_i
48	1	1
51	5	6
54	14	20
57	13	33
60	10	43
	43	

$N_i = 20$

Percentil 30 = 54

d) ¿Qué peso medio es más representativo: el de las personas que miden 164 cm. o el de las que miden 168 cm.?

Distribución de X / Y = 164

x_i	n_i	$n_i x_i$	$n_i x_i^2$
48	2	96	4608
51	4	204	10404
54	6	324	17496
57	1	57	3249
60	0	0	0
	13	681	35757

$$Media: \quad \bar{x} = \frac{\sum_{i=1}^{k} n_i x_i}{n} = \frac{681}{13} = 52.3846$$

$$\sigma^2 = \frac{\sum_{i=1}^{k} n_i x_i^2}{n} - \bar{x}^2 = \frac{35757}{13} - 52.3846^2 = 6.3921$$

$$\sigma = \sqrt{6.3921} = 2.528 \qquad C.V. = \frac{\sigma}{\bar{x}} = \frac{2.528}{52.3846} = 0.0482$$

Distribución de X / Y = 168

x_i	n_i	$n_i x_i$	$n_i x_i^2$
48	0	0	0
51	2	102	5202
54	5	270	14580
57	8	456	25992
60	4	240	14400
	19	1068	60174

$$Media: \quad \bar{x} = \frac{\sum_{i=1}^{k} n_i x_i}{n} = \frac{1068}{19} = 56.21$$

$$\sigma^2 = \frac{\sum_{i=1}^{k} n_i x_i^2}{n} - \bar{x}^2 = \frac{60174}{19} - 56.21^2 = 7.4885$$

$$\sigma = \sqrt{7.4885} = 2.7361 \qquad C.\,V. = \frac{\sigma}{\bar{x}} = \frac{2.7361}{56.21} = 0.0486$$

$$\left.\begin{array}{l}\text{En la Distribución de X / Y = 164, C. V. = 0.0482}\\ \text{En la Distribución de X / Y = 168, C. V. = 0.0486}\end{array}\right\} \Rightarrow$$

La media de X / Y = 164 es más representativa

2. Se ha medido la edad, X , y la tensión arterial máxima, Y , de un grupo de personas.

a) Calcular la tensión arterial media de las personas con más de 20 años.

b) Calcular la edad media de las personas con tensión arterial entre 100 y 120.

c) En el conjunto de personas con tensión arterial entre 100 y 120, calcular la edad mínima del 30% de las personas con más edad.

X \ Y	90 – 100	100 – 120	120 –140
10 – 15	6	3	1
15 – 20	5	10	2
20 – 25	4	1	7
25 –30	2	2	4

a) Calcular la tensión arterial media de las personas con más de 20 años

X \ Y	90 – 100	100 – 120	120 –140
10 – 15	6	3	1
15 – 20	5	10	2
20 – 25	4	1	7
25 –30	2	2	4

- Media de la Distribución de Y / X > 20

Y	y_i	n_i	$n_i y_i$
90 – 100	95	6	570
100 – 120	110	3	330
120 – 140	130	11	1430
		20	2330

$$Media: \quad \bar{y} = \frac{\sum_{i=1}^{k} n_i y_i}{n} = \frac{2330}{20} = 116.5$$

b) Calcular la edad media de las personas con tensión arterial entre 100 y 120

X \ Y	90 – 100	100 – 120	120 –140
10 – 15	6	3	1
15 – 20	5	10	2
20 – 25	4	1	7
25 –30	2	2	4

- Media de la Distribución de X / 100 < Y < 120

X	x_i	n_i	$n_i x_i$
10 – 15	12.5	3	37.5
15 – 20	17.5	10	175
20 – 25	22.5	1	22.5
25 –30	27.5	2	55
		16	290

$$Media: \quad \bar{x} = \frac{\sum_{i=1}^{k} n_i x_i}{n} = \frac{290}{16} = 18.125$$

c) En el conjunto de personas con tensión arterial entre 100 y 120, calcular la edad mínima del 30% de las personas con más edad

X \ Y	90 – 100	100 – 120	120 –140
10 – 15	6	3	1
15 – 20	5	10	2
20 – 25	4	1	7
25 –30	2	2	4

➢ Percentil 70 de la Distribución de X / 100 < Y < 120

(16x70)/100 =11.2

X	n_i	N_i
10 – 15	3	3
15 – 20	10	13
20 – 25	1	14
25 –30	2	16
	16	

$\boxed{N_i = 13}$

$$P_{70} = e_{i-1} + \frac{11.2 - N_{i-1}}{n_i} a_i = 15 + \frac{11.2 - 3}{10} \times 5 = 19.1$$

3. Dadas las siguientes distribuciones, ¿son independientes las variables X e Y?

a)

X \ Y	10	15	20
1	0	3	0
2	1	0	0
3	0	0	5
4	0	1	0

b)

X \ Y	10	15	20	25
1	0	3	0	4
2	0	0	1	0
3	2	0	0	0

c)

X \ Y	10	15	20
1	0	5	0
2	3	0	0
3	0	0	2

d)

X \ Y	10	15	20
1	3	2	0
2	1	0	2
3	0	1	1

¿Son independientes las variables X e Y?

a.-

X \ Y	10	15	20	Marginal X
1	0	3	0	3
2	1	0	0	1
3	0	0	5	5
4	0	1	0	1
Marginal Y	1	4	5	10

$$n_{12} = 3 \neq \frac{n_{1\bullet}\, n_{\bullet 2}}{n} = \frac{3 \times 4}{10} = 1.2 \quad \Rightarrow$$

Las variables X e Y no son Independientes

b.-

X \ Y	10	15	20	25	Marginal X
1	0	3	0	4	7
2	0	0	1	0	1
3	2	0	0	0	2
Marginal Y	2	3	1	4	10

$$n_{23} = 1 \neq \frac{n_{2\bullet}\, n_{\bullet 3}}{n} = \frac{1 \times 1}{10} = 0.1 \quad \Rightarrow$$

Las variables X e Y no son Independientes

¿Son independientes las variables X e Y?

c.-

X \ Y	10	15	20	Marginal X
1	0	5	0	5
2	3	0	0	3
3	0	0	2	2
Marginal Y	3	5	2	10

$$n_{11} = 0 \neq \frac{n_{1\bullet} n_{\bullet 1}}{n} = \frac{5 \times 3}{10} = 1.5 \Rightarrow$$

Las variables X e Y no son Independientes

d.-

X \ Y	10	15	20	Marginal X
1	3	2	0	5
2	1	0	2	3
3	0	1	1	2
Marginal Y	4	3	3	10

$$n_{21} = 1 \neq \frac{n_{2\bullet} n_{\bullet 1}}{n} = \frac{3 \times 4}{10} = 1.2 \Rightarrow$$

Las variables X e Y no son Independientes

4. Se quiere estudiar la posible asociación entre el nivel de estudios de un grupo de personas y el hábito de fumar. Las personas se han seleccionado de forma aleatoria y los datos se presentan en la tabla adjunta.

a) Calcular la moda de la distribución de los estudios

b) ¿Son dichas variables independientes?

c) En caso negativo, estudiar el grado de asociación entre las variables

Fumar \ n_estudios	Primarios	Medios	Superiores
SI	20	10	4
NO	16	12	2

a) Calcular la moda de la distribución del nivel de estudios

y_j	$n_{j/i=1}$
Primarios	20
Medios	10
Superiores	4
	34

La moda son los estudios Primarios

b) ¿Son dichas variables independientes?

Fumar \ n_estudios	Primarios	Medios	Superiores	
SI	20	10	4	34
NO	16	12	2	30
	36	22	6	64

$$n_{13} = 4 \neq \frac{34 \times 6}{64} = 3.1875$$

Las variables no son independientes

$$\Downarrow$$

Existe algún grado de asociación entre el hábito de fumar y el nivel de estudios

c) Estudiar el grado de asociación entre las variables

Fumar \ n_estudios	Primarios	Medios	Superiores	
SI	20	10	4	34
NO	16	12	2	30
	36	36	6	64

Fumar	Nivel Estudios	t_{ij}	$\frac{(t_{ij}-n_{ij})^2}{t_{ij}}$
SI	Primarios	19.125	0.04
SI	Medios	19.125	4.3538
SI	Superiores	3.1875	0.207
NO	Primarios	16.875	0.045
NO	Medios	16.875	1.4083
NO	Superiores	2.8125	0.2347

donde $$t_{ij}=\frac{n_{i.}n_{.j}}{n}$$

Vamos a calcular los indicadores de asociación

Coeficiente χ^2

$$\chi^2 = \sum_{ij} \frac{\left(t_{ij} - n_{ij}\right)^2}{t_{ij}} = 6,2888, \text{con}$$

$$0 \leq \chi^2 \leq N \min\left\{p-1, q-1\right\} = 64$$, ya que

$$N \min\left\{p-1, q-1\right\} = 64 \times \min\left\{2-1, 2-1\right\} = 64$$

Coeficiente de contingencia de Pearson

$$C = \sqrt{\frac{\chi^2}{n+\chi^2}} = 0.2991 \quad 0 \leq C \leq \sqrt{\frac{k-1}{k}} = 0.707$$, puesto que

$$k = \min\left\{p, q\right\} = 2$$

Coeficiente T de Tschuprow

$$T = \frac{\chi^2}{n\sqrt{(p-1)(q-1)}} = 0.06948 \qquad 0 \leq T \leq 1$$

Por tanto, el grado de asociación entre el hábito de fumar y el nivel de estudios es pobre o débil

5. Se quiere estudiar la relación entre la edad de los obreros y el tipo de accidente laboral en una industria. Se consideran dos categorías para la edad: A1 (menor de 40 años) y A2 (de 40 años en adelante), mientras que, en el tipo de accidente, se consideran tres: sobreesfuerzo (B1), caída de personal (B2) y golpes por herramientas (B3). Los datos se presentan en la tabla adjunta.

	Tipo de accidente		
Edad	B1	B2	B3
A1	17	15	20
A2	21	25	12

Se pide

a) ¿Son dichas variables independientes?

c) Estudia el grado de asociación entre las variables

b) ¿Son dichas variables independientes?

	Tipo de accidente			
Edad	B1	B2	B3	
A1	17	15	20	52
A2	21	25	12	58
	38	40	32	110

$$n_{21} = 21 \neq \frac{38 \times 58}{110} = 20.036$$

La edad y el tipo de accidente
no son independientes

$$\Downarrow$$

Existe algún grado de asociación
entre dichas variables

c) Estudia el grado de asociación entre las variables

	Tipo de accidente			
Edad	B1	B2	B3	
A1	17	15	20	52
A2	21	25	12	58
	38	40	32	110

Edad	Tipo accidente	t_{ij}	$\frac{\left(t_{ij}-n_{ij}\right)^2}{t_{ij}}$
A1	B1	17.96	0.051
A1	B2	18.91	0.808
A1	B3	15.13	1.568
A2	B1	20.04	0.046
A2	B2	21.09	0.725
A2	B3	16.87	1.406

donde $$t_{ij}=\frac{n_{i.}n_{.j}}{n}$$

Medidas de asociación

Coeficiente χ^2

$$\chi^2 = \sum_{ij} \frac{\left(t_{ij} - n_{ij}\right)^2}{t_{ij}} = 4.604$$ con

$$0 \le \chi^2 \le N \min\left\{p-1, q-1\right\} = 110$$, ya que

$$N \min\left\{p-1, q-1\right\} = 110 \times \min\left\{2-1, 2-1\right\} = 110$$

Coeficiente de contingencia de Pearson

$$C = \sqrt{\frac{\chi^2}{n+\chi^2}} = 0.2004 \quad 0 \le C \le \sqrt{\frac{k-1}{k}} = 0.707$$, puesto que

$$k = \min\left\{p, q\right\} = 2$$

Coeficiente T de Tschuprow

$$T = \frac{\chi^2}{n\sqrt{(p-1)(q-1)}} = 0.0296 \qquad 0 \le T \le 1$$

Por tanto, el grado de asociación entre la edad y el tipo de accidente en una industria es débil

6. Los siguientes datos muestran la tasa de paro media (en %) y el aumento porcentual en casos de depresión en cinco ciudades andaluzas

a) Determina el aumento de depresiones esperado para un nivel de paro igual al 18 %.

b) Bondad de ajuste.

Nivel de Paro	Aumento depresión
16.5	9
17.8	14
20	23
18.5	19
22	22

a) Aumento de la depresión para una tasa de paro del 18%

x_i	y_i	$x_i y_i$	x_i^2	y_i^2
16.5	9	148.5	272.25	81
17.8	14	249.2	316.84	196
20	23	460	400	529
18.5	19	351.5	342.25	361
22	23	484	484	484
Σ=94.8	Σ=87	Σ= 1693.2	Σ= 1815.34	Σ= 1651

$$\bar{x} = \frac{94.8}{5} = 18.96 \ ; \quad \bar{y} = \frac{87}{5} = 17.4$$

$$\sigma_{xy} = \frac{\sum n_i x_i y_i}{n} - \overline{x}\,\overline{y} = \frac{1693.2}{5} - 18.96 \times 17.4 = 8.73$$

$$\sigma_x^2 = \frac{\sum n_i x_i^2}{n} - \bar{x}^2 = \frac{1815.34}{5} - 18.96^2 = 3.5864$$

$$\sigma_y^2 = \frac{\sum n_i y_i^2}{n} - \bar{y}^2 = \frac{1651}{5} - 17.4^2 = 27.44$$

$$y = a + bx$$

$$b = \frac{Cov\left[X,Y\right]}{Var\left[X\right]} = \frac{\sigma_{xy}}{{\sigma_x}^2} = \frac{8.736}{3.5864} = 2.4359$$

$$a = \overline{y} - b\overline{x} = 17.4 - 2.4359 \times 18.96 = -28.78$$

$$y = a + b\,x = -150.5068 + 1.258\,x$$

Para $x = 18 \quad \Rightarrow$

$$y = a + bx = -28.78 + 2.4359 \times 18 = 15.0662$$

b)

$$r^2 = \frac{\sigma_{xy}^2}{\sigma_x^2\,\sigma_y^2} = \frac{8.736^2}{3.5864x27.44} = 0.7755$$

7. Las puntuaciones finales en Estadística y Economía de 10 estudiantes elegidos al azar, aparecen en la tabla adjunta:

a) Si un estudiante tiene una nota 75 en Estadística, ¿qué nota tendrá en Economía?

b) Calcula los valores de los coeficientes de correlación y de determinación. Interpretación.

Estadística (X)	Economía (Y)
75	82
80	78
93	86
65	73
87	91
71	80
98	95
68	72
84	89
77	74

a) Si un estudiante tiene una nota 75 en Estadística, ¿qué nota tendrá en Economía?

x_i	y_i	$x_i y_i$	x_i^2	y_i^2
75	82	6150	5625	6724
80	78	6240	6400	6084
93	86	7998	8649	7396
65	73	4745	4225	5329
87	91	7917	7569	8281
71	80	5680	5041	6400
98	95	9310	9604	9025
68	72	4896	4624	5184
84	89	7476	7056	7921
77	74	5698	5929	5476
Σ=798	Σ=820	Σ=66110	Σ= 64722	Σ= 67820

$$\bar{x} = \frac{798}{10} = 79.8 \; ; \qquad \bar{y} = \frac{820}{10} = 82$$

$$\sigma_{xy} = \frac{\sum n_i x_i y_i}{n} - \overline{x}\overline{y} = \frac{6610}{10} - 79.8 \times 82 = 67.4$$

$$\sigma_x^2 = \frac{\sum n_i x_i^2}{n} - \bar{x}^2 = \frac{64722}{10} - 79.8^2 = 104.16$$

$$\sigma_y^{2} = \frac{\sum n_i y_i^{2}}{n} - \bar{y}^{2} = \frac{67820}{10} - 82^{2} = 58$$

$$y = a + bx$$

$$b = \frac{Cov[X, Y]}{Var[X]} = \frac{\sigma_{xy}}{\sigma_x^{2}} = \frac{67.4}{104.26} = 0.6471$$

$$a = \bar{y} - b\bar{x} = 82 - 0.6471 \times 79.8 = 30.36$$

$$y = a + b\,x = 30.36 + 0.6471x$$

Para $x = 75 \quad \Rightarrow$

$$y = a + bx = 30.36 + 0.6471 \times 75 = 78.89$$

b)

$$r = \frac{\sigma_{xy}}{\sigma_x\,\sigma_y} = \frac{67.4}{\sqrt{104.16}x\sqrt{58}} = 0.8671$$

Hay asociación lineal fuerte y positiva entre las calificaciones en Estadística y las calificaciones en Economía.

$$r^{2} = 0.7519$$

El 75.19 % de la variación en las calificaciones en Economía se pueden explicar por su relación lineal con las notas en Estadística

Ejercicios propuestos

Ejercicio 1:

Una compañía de seguros considera que el número de vehículos (y) que circulan por una determinada autopista a más de 120 km/h , puede ponerse en función del número de accidentes (x) que ocurren en ella. Durante 5 días obtuvo los siguientes resultados:

Accidentes	5	7	2	1	9
Número de vehículos	15	18	10	8	20

a) Calcula el coeficiente de correlación lineal.

b) Si ayer se produjeron 6 accidentes, ¿cuántos vehículos podemos suponer que circulaban por la autopista a más de 120 km/h?

c) ¿Es buena la predicción? Razona la respuesta

Ejercicio 2:

Los siguientes datos corresponden a los gastos en desplazamiento que tienen que hacer un grupo de alumno de un centro de secundaria para llegar a clase cada día y sus calificaciones en Matemáticas. Los datos se han seleccionado aleatoriamente.

Gasto	2	4	6	8	10	12
Notas	1	7.5	3	4	8	9

Se pide:

a) Obtén la nota esperada para un alumno que gaste 20 euros en llegar a clase cada día

b) Calcula el valor del coeficiente de correlación lineal. Interpretación.

Ejercicio 3:
Se quiere estudiar la asociación entre la edad (medida en 3 niveles) y el nivel de colesterol en sangre (en 3 niveles) de un grupo de personas sanas. Los datos disponibles se presentan en la siguiente tabla:

Edad \ n_colesterol	Bajo	Medio	Alto
Joven	6	9	28
Edad Intermedia	9	10	14
Mayor	4	25	44

Se pide:

a) ¿Qué edad es la más habitual entre los que tienen una nivel de colesterol alto?

b) ¿Son independientes estas variables?

c) En caso negativo, calcula los indicadores de asociación.