ESTADÍSTICA Y ANÁLISIS DE DATOS

Práctica del Tema 2. Variables estadísticas bidimensionales

Problemas

1. En la siguiente tabla aparecen, según la OIT (Organización Internacional del Trabajo), los parados de la CAV clasificados por sectores económicos (X) y por provincias (Y). Los números expresan miles de personas en Junio de 1992.

X \ Y	Araba	Bizkaia	Gipuzkoa
Agricultura	0	0,8	0,8
Industria	6	18,3	10,2
Construcción	1,4	8	5,3
Servicios	7,5	45,1	17,9
Busca 1^{er} empleo	8	39,6	17,7

Se pide:

a) Realizar las representaciones gráficas que nos muestren cómo se distribuían en 1992 los parados de la CAV por provincias y por sectores económicos, respectivamente.

b) Comparar, añadiendo los comentarios oportunos, la distribución de los parados según el sector en cada una de las tres provincias. ¿Estaba próxima a la independencia la distribución de estas dos variables en ese colectivo? Razona tu respuesta.

2. En la tabla adjunta aparece la clasificación de los municipios de Araba, Bizkaia, Gipuzkoa y los correspondientes a toda la CAV según el número de habitantes:

Nº habitantes (X)	Araba	Bizkaia	Gipuzkoa	CAV
$x \leq 500$	43,1 %	20,2 %	25,3 %	26,72 %
$x \leq 5000$	94,1 %	72,5 %	65,5 %	74,5 %
$x \leq 50000$	98 %	94,5 %	97,7 %	96,4 %
$x \leq 400000$	100 %	100 %	100 %	100 %

Como se puede observar, estas clasificaciones vienen dadas en términos de porcentajes acumulados.

a) Representar gráficamente la distribución de los municipios de la CAV según el número de habitantes. A la vista de este gráfico, razona cuál crees que será el grado de dispersión respecto de la media. Contrasta tu opinión a través de un estadístico apropiado. (Nota: utilizar, si es necesario, el valor medio de cada intervalo como marca de clase.)

b) A partir de los resultados obtenidos hasta ahora, y sin hacer más cálculos ¿cómo crees que será el grado de uniformidad de esta distribución?

c) Se ha decidido subvencionar un polideportivo al 50 % de los municipios, aquellos con menor número de habitantes, ¿cuál será el número de habitantes correspondiente al mayor municipio de los que reciben subvención?

d) Considerando la variable bidimensional (X, Y), $X=$ número de habitantes e $Y=$ provincia a la que pertenecen, analizar si estas dos variables están próximas o no a distribuirse de forma independiente en el colectivo de municipios de la CAV. Razona tu respuesta y comenta los resultados obtenidos.

3. Para las siguientes tablas bidimensionales de frecuencias absolutas, dedúzcase el valor del coeficiente de correlación en cada una:

X \ Y	y_1	y_2
x_1	3	0
x_2	0	2

X \ Y	y_1	y_2
x_1	0	1
x_2	1	0

X \ Y	y_1	y_2
x_1	1	2
x_2	2	4

En las tres tablas x_1 es menor que x_2 e y_1 es menor que y_2.

4. A lo largo de un año se ha medido, para 16 máquinas herramientas:

X: años de funcionamiento

Y: averías durante el año,

resultando la tabla de contingencia adjunta.

X \ Y	3	4	5	
1	0	2	3	
2	6	0	0	
3	0	2	3	

Se pide:

a) Distribución marginal de frecuencias relativas de la variable Y.

b) Distribución de la variable Y condicionada al valor $X = 2$, en términos de frecuencias relativas.

c) ¿Tienen una distribución independiente, en esta tabla, las variables X e Y? Razónese la respuesta.

d) Obténgase el coeficiente de correlación lineal r_{xy}. Coméntese brevemente el resultado obtenido.

5. Sea la siguiente distribución de frecuencias:

X \ Y	$(-0{,}5; 0{,}5)$	$(0{,}5; 1{,}5)$	
$(-1{,}5; -0{,}5)$	2	6	
$(-0{,}5; 0{,}5)$	2	n_{22}	
$(0{,}5; 1{,}5)$	n_{31}	n_{32}	

Suponiendo que el número de individuos sea 20, obténganse las frecuencias n_{22}, n_{31} y n_{32} de modo que se cumpla simultáneamente:

$$M_e(X) = 0 \qquad\qquad S_{xy} = -0{,}2$$

Tómense como marcas de clase los puntos medios.

6. La altura en centímetros y el peso en kilogramos de los alumnos de una determinada clase siguen la siguiente distribución de frecuencias conjunta:

	Peso		
Altura	$[50, 70)$	$[70, 90)$	$[90, 100)$
$[150, 170)$	9	3	1
$[170, 190)$	4	7	2

a) Representa gráficamente esta distribución bidimensional de frecuencias.

b) ¿Cuál es el número de alumnos cuyo peso está entre 70 y 90 kilos, independientemente de la altura que tengan?

c) De entre los alumnos cuya estatura está en el intervalo $[170, 190)$, ¿cuál es el porcentaje que pesa entre 50 y 70 kilogramos?

d) ¿Cuál es el porcentaje de alumnos cuya estatura está en el intervalo $[170, 190)$ y cuyo peso está en el intervalo $[50, 70)$?

e) Calcula la distribución de frecuencias relativas de la variable **Altura** condicionada a **Peso** $= [70, 90)$.

f) Calcula el coeficiente de correlación entre las variables **Peso** y **Altura**. Interpreta el resultado obtenido.

g) ¿Las variables **Peso** y **Altura** se distribuyen de forma independiente en este colectivo? Razónalo.

h) Consideremos ahora la variable **Nuevo peso** que resulta de disminuir el peso de todos los individuos un 10 %.

1) Obtén a partir de los valores típicos de **Peso** la media, la mediana y la desviación típica de **Nuevo peso**.
2) ¿Cuál es el valor del coeficiente de correlación entre **Nuevo peso** y **Altura**?
3) ¿Cuál es el valor de la covarianza entre **Nuevo peso** y **Peso**? ¿Y el valor de su correspondiente coeficiente de correlación?

7. La siguiente tabla recoge la información sobre el consumo medio de un coche, variable C medida en litros a los 100 km, para distintas velocidades, variable V medida en km por hora:

Consumo (C)	Velocidad (V)
11	20
9	40
6	60
5	80
6	100
8	120
11	140

a) Calcula el coeficiente de correlación entre las variables **Consumo** y **Velocidad**. ¿Crees que existe relación **lineal** entre ambas variables? Comenta el resultado.

b) Dibuja el diagrama de dispersión del **Consumo** con respecto a la **Velocidad**. A la vista del gráfico, ¿crees que ambas variables son independientes? Comenta en detalle.

c) ¿Son contradictorios los resultados de los apartados anteriores? Razona tu respuesta.

Cuestiones

1. Sea la siguiente tabla de contingencia. Sabiendo que $f_{11} = 0{,}25$, ¿qué valor tomará n_{23}?

X \ Y	Y_1	Y_2	Y_3	
X_1	5	6	3	
X_2	2	1	n_{23}	

(A) 0,15 (B) 3 (C) 7 (D) f_{11} no puede tomar el valor 0,25 (E) Todo falso

2. Si realizamos un cambio de variable del tipo $U = 4X + 9$, $V = 5Y + 4$:

(A)$r_{uv} = 16r_{xy}$ (B) $r_{uv} = r_{xy}$ (C)Todo falso (D) $r_{uv} = 20r_{xy} + 36$ (E) $r_{uv} = 20r_{xy}$

3. Dados los datos sobre las variables X=horas trabajadas e Y=producción obtenida, proporcionados en la siguiente tabla, la dependencia de la variable Y sobre la variable X será:

X	Y (%)
1	2
1	4
3	6
6	8
5	10
7	12

(A) Total, lineal y directa (B) Total, lineal e inversa (C) Fuerte, lineal y directa (D) Fuerte, lineal e inversa (E) Todo falso

4. Sabiendo que $S_x^2 = 3$, $S_y^2 = 8$ y $S_z^2 = 9$, donde $Z = X + Y$, podemos deducir que:

(A) $S_{xy} = 2$ (B) $S_{xy} = -2$ (C) X e Y están incorrelacionadas (D) $S_{xy} = 1$ (E) $S_{xy} = -1$

5. Dados los datos de la tabla, ¿cuál de los siguientes resultados es cierto?

X	Y (%)
0	3
1	4
-1	2
1,5	4,5

(A) $r_{xy} = -1$ (B) $r_{xy} = 0$ (C) $r_{xy} = 1$ (D) $\overline{x} = 1{,}5$ (E) Todo falso

6. La covarianza de una variable consigo misma es siempre igual a:

   (A) 1 (B) Su varianza (C) 0 (D) El inverso de su varianza (E) Todo falso

7. Para una distribución bidimensional se tiene la siguiente relación entre estadísticos: $m_{11} = 2(S_x \cdot S_y)$, ¿cómo será la relación lineal entre las variables X e Y?

   (A) Directa y fuerte (B) Inversa y fuerte (C) Directa y débil (D) Depende del signo y del valor de m_{11} (E) Estos datos son incompatibles

8. Si dos variables X e Y no se distribuyen de forma independiente en una tabla de datos puede decirse que:

   (A) Las variables X e Y pueden presentar incorrelación

   (B) Existe algún grado de relación lineal entre las variables

   (C) La correlación será total y negativa

   (D) La correlación será total y positiva

   (E) Todo falso

9. Si $X = aU + b$, $Y = eV + f$, $S_{xy} = S_{uv}$, es seguro que:

   (A) $|a| = |e| = 1$ (B) $ae = 1$ (C) Se cumple para todo valor de a y e (D) $a = e = 1$ (E) Todo falso

10. Respecto a la siguiente distribución de frecuencias relativas bidimensional

X \ Y	y_1	y_2	
x_1	0,3	0,2	
x_2	0,3	0,4	

   podemos decir que:

   (A) Es correcta, ya que se trata de frecuencias relativas

   (B) Es incorrecta, ya que las frecuencias de una tabla bidimensional son números naturales

   (C) Es correcta y la relación entre X e Y es directa

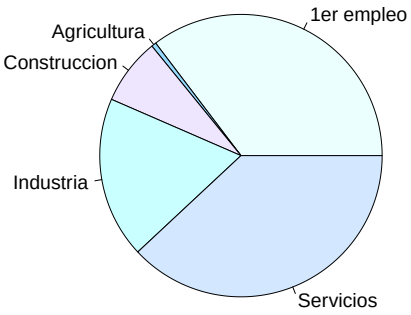
   (D) Es correcta y la relación entre X e Y es inversa

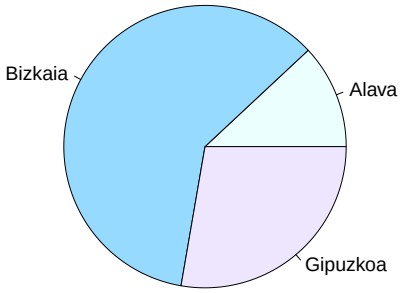
   (E) Todo falso

Solución a los problemas y cuestiones del Tema 2

Problemas

1. *a*) Gráficos de sectores.





X \ Y	Araba	Bizkaia	Gipuzkoa	
Agricultura	0	0,8	0,8	1,6
Industria	6	18,3	10,2	34,5
Construcción	1,4	8	5,3	14,7
Servicios	7,5	45,1	17,9	70,5
Busca 1^{er} empleo	8	39,6	17,7	65,3
	22,9	111,8	51,9	186,6

Según estos datos, en Junio de 1992 el mayor porcentaje de parados de la CAV pertenecía al sector servicios (37,78 %), seguido muy de cerca por los que buscaban el primer empleo (34,99 %); en tercer lugar el perteneciente al sector industrial, luego a la construcción y, por último, al sector agrícola (tan sólo un 0,86 %).

Por provincias, más de la mitad de los parados vivían en Bizkaia, casi el 30 % en Gipuzkoa y sólo el 12,27 % en Araba.

b) Para comparar la distribución de los parados según el sector (X) en cada una de las tres provincias calculamos las distribuciones condicionadas $X|Y = \text{Araba}$, $X|Y = \text{Bizkaia}$ y $X|Y = \text{Gipuzkoa}$ en términos de frecuencias relativas:

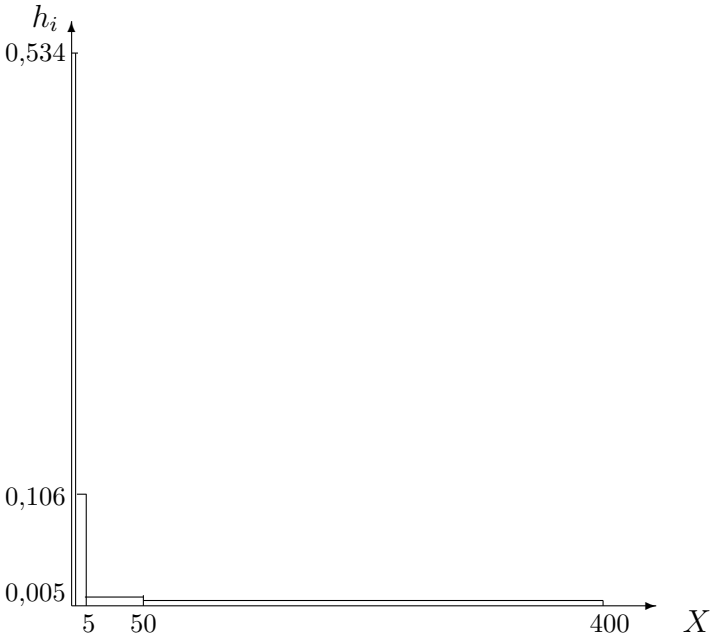
Sector	$f_{i\|Y=\text{Araba}}$	$f_{i\|Y=\text{Bizkaia}}$	$f_{i\|Y=\text{Gipuzkoa}}$
Agricultura	0	0,0072	0,015
Industria	0,262	0,16	0,19
Construcción	0,061	0,0715	0,10
Servicios	0,327	0,40	0,345
Busca 1^{er} Empleo	0,349	0,35	0,341

Según estos datos, en las tres provincias los mayores porcentajes de parados eran los correspondientes al sector Servicios y Primer empleo (para el primero ligeramente superior en Bizkaia, 40 %, seguido de Gipuzkoa, 34,5 %, y por último Araba, 32,7 %; para el segundo alrededor del 35 % en las tres provincias); después el asociado al sector Industrial (ligeramente superior en Araba, 26,2 %) y los sectores con menor porcentaje de parados eran Construcción y Agricultura (próximo a cero en Gipuzkoa y Bizkaia, así como nulo en Araba).

La distribución de los parados según el sector, a pesar de tener ligeras diferencias entre provincias, presentaba los mismos rasgos básicos en las tres provincias y éstos coincidían con los de la distribución para toda la CAV. Por tanto, podemos considerar que la distribución de los parados de la CAV según el sector estaba próxima a ser independiente de la provincia en 1992.

2. *a*) Representación gráfica: histograma.

x_i (miles de hb.)	h_i
$(0; 0{,}5]$	0,534
$(0{,}5; 5]$	0,106
$(5; 50]$	0,005
$(50; 400]$	0,0001



En el gráfico se observa una gran dispersión, los valores están muy separados de la media aritmética ($\overline{x} = 15{,}5$ miles de hb.). Para comprobarlo calculamos el coeficiente de variación:

$$g_0(x) = \frac{41{,}85}{15{,}5} = 2{,}7$$

Obtenemos un valor mayor que la unidad, indicando que efectivamente hay un alto grado de dispersión respecto de la media y, por tanto, la media es poco representativa del número de habitantes que realmente hay en los municipios de este colectivo.

b) Desde el punto de vista estadístico, sabemos que un alto grado de dispersión va asociado a un alto grado de concentración de la masa de la distribución. En

este caso, la masa de la distribución es el total de habitantes en la CAV y los resultados obtenidos nos indican que no se reparte uniformemente entre los N municipios sino que se concentra en un número pequeño de ellos (en los más grandes).

c) Calculamos el valor de la Mediana M_e:

$$M_e \in [0{,}5; 5) \rightarrow M_e = 2{,}693 \text{ miles de hb.}$$

El mayor municipio de los que reciben subvención tendrá, aproximadamente, 2693 habitantes.

d) La distribución de los municipios según el número de habitantes (X) no es independiente de la provincia (Y) en este colectivo ya que las distribuciones condicionadas no son semejantes entre sí, ni semejantes a la marginal:

x_i	$f_{i\|Y=\text{Araba}}$	$f_{i\|Y=\text{Bizkaia}}$	$f_{i\|Y=\text{Gipuzkoa}}$	$f_{i\bullet}$
$[0; 0{,}5)$	0,43	0,202	0,253	0,2672
$[0{,}5; 5)$	0,51	0,523	0,402	0,4778
$[5; 50)$	0,039	0,220	0,322	0,219
$[50; 400)$	0,02	0,055	0,023	0,036

Entre otras diferencias se observa que en Araba el 43 % de los municipios tienen menos de 500 hb. mientras que en Bizkaia y Gipuzkoa esto sólo ocurre en el 20,2 % y el 25 % de los municipios, respectivamente.

3. $r_{xy} = 1$ $\quad r_{xy} = -1$ $\quad r_{xy} = 0$

4. *a*) $f_{\bullet j}$: 6/16 4/16 6/16

   b) $f_{j|X=2}$: 1 0 0

   c) No son independientes en distribución porque las distribuciones condicionadas no coinciden con la marginal.

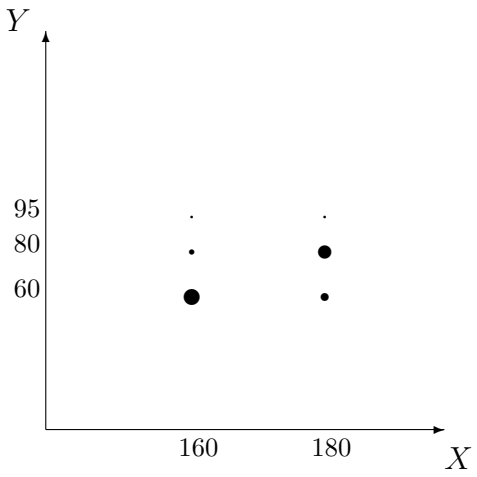
   d) Si consideramos $U = X-2$ y $V = Y-4$ se obtiene fácilmente que $r_{xy} = r_{uv} = 0$, es decir, las variables están incorrelacionadas.

5. $n_{22} = 2 = n_{32}$ $\quad n_{31} = 6$

6.

	Peso			
Altura	$[50, 70)$	$[70, 90)$	$[90, 100)$	$n_{i.}$
$[150, 170)$	9	3	1	13
$[170, 190)$	4	7	2	13
$n_{.j}$	13	10	3	26
$N_{.j}$	13	23	26	

   a) Diagrama de dispersión.



b) $n_{\cdot 2}$=10. Hay 10 alumnos que pesan entre 70 y 90 kg independientemente de su altura.

c) $f_{j=1|X\in[170,190)} = \dfrac{n_{21}}{n_{2\cdot}} = \dfrac{4}{13} = 0{,}3076$. Aproximadamente el 30 % de los alumnos cuya estatura está en el intervalo [170,190) pesan entre 50 y 70 kg.

d) $f_{21} = \dfrac{n_{21}}{N} = \dfrac{4}{26} = 0{,}1538$. Aproximadamente el 15 % de los alumnos tiene una estatura comprendida en el intervalo [170,190) y pesa entre 50 y 70 kg.

e) Sea Altura|Peso $\in [70, 90) \equiv X|Y \in [70, 90) \equiv X|Y = y_2$.

$X\|Y = y_2$	$n_{i\|Y=y_2}$	$f_{i\|Y=y_2}$
160	3	3/10=0,3
180	7	7/10=0,7
	10	1

f) Coeficiente de correlación: $r_{xy} = \dfrac{S_{xy}}{S_x S_y}$

$\bar{x} = \dfrac{160 \cdot 13 + 180 \cdot 13}{26} = 170\ cm.$

$\bar{y} = \dfrac{60 \cdot 13 + 80 \cdot 10 + 95 \cdot 3}{26} = \dfrac{1865}{26} = 71{,}73\ Kg.$

$a_{20} = \dfrac{1}{26}(160^2 \cdot 13 + 180^2 \cdot 13) = \dfrac{754000}{26} = 29000\ cm^2.$

$a_{02} = \dfrac{1}{26}(60^2 \cdot 13 + 80^2 \cdot 10 + 95^2 \cdot 3) = \dfrac{137875}{26} = 5302{,}88\ kg^2.$

$S_x^2 = a_{20} - \bar{x}^2 = 29000 - 170^2 = 100\ cm^2.$

$S_y^2 = a_{02} - \bar{y}^2 = 5302{,}88 - 71{,}73^2 = 157{,}68\ kg^2.$

$a_{11} = \dfrac{1}{26}(160 \cdot 60 \cdot 9 + 160 \cdot 80 \cdot 3 + 160 \cdot 95 \cdot 1 + 180 \cdot 60 \cdot 4 + 180 \cdot 80 \cdot 7 + 180 \cdot 95 \cdot 2) =$

$\dfrac{318200}{26} = 12238{,}46\ cmkg.$

$S_{xy} = a_{11} - \bar{x}\bar{y} = 12238{,}46 - 170 \cdot 71{,}73 = 44{,}36\ cmkg.$

$$r_{xy} = \frac{44{,}36}{\sqrt{100}\sqrt{157{,}68}} = 0{,}35.$$

Las variables altura y peso tienen una relación lineal directa o positiva, de manera que a mayor altura, en general, mayor es el peso. Sin embargo, la relación lineal no es muy fuerte.

g) No se distribuyen de manera independiente ya que al ser $r_{xy} > 0$ existe una relación lineal positiva.

h) Sea Z la variable Nuevo peso. $Z = Y - 0{,}1Y = 0{,}9Y$.

1) $\bar{z} = 0{,}9\bar{y} = 0{,}9 \cdot 71{,}73 = 64{,}557\ kg.$

$Me(Z) = 0{,}9Me(Y) = 0{,}9 \cdot 70 = 63\ kg.$

$Me(Y) \in [50, 70) \qquad Me(Y) = 50 + \frac{13-0}{13-0} \cdot 20 = 70\ kg.$

$S_z^2 = 0{,}9^2 \cdot S_y^2 = 0{,}9^2 \cdot 157{,}68 = 127{,}72\ kg^2 \Rightarrow S_z = \sqrt{S_z^2} = 11{,}30\ kg.$

2) $r_{zx} = \dfrac{0{,}9 \cdot S_{yx}}{0{,}9 \cdot S_y \cdot S_x} = r_{xy} = 0{,}35.$

3) $S_{zy} = 0{,}9 \cdot S_{yy} = 0{,}9 \cdot S_y^2 = 0{,}9 \cdot 157{,}68 = 141{,}912$

$$r_{zy} = \frac{0{,}9 \cdot S_y^2}{0{,}9 \cdot S_y \cdot S_y} = \frac{0{,}9S_y^2}{0{,}9S_y^2} = 1$$

7.

C^V	20	40	60	80	100	120	140	$n_{i.}$
5	0	0	0	1	0	0	0	1
6	0	0	1	0	1	0	0	2
8	0	0	0	0	0	1	0	1
9	0	1	0	0	0	0	0	1
11	1	0	0	0	0	0	1	2
$n_{\cdot j}$	1	1	1	1	1	1	1	7

a) $r_{cv} = \dfrac{S_{cv}}{S_c \cdot S_v}$

$$\bar{c} = \frac{5 \cdot 1 + 6 \cdot 2 + 8 \cdot 1 + 9 \cdot 1 + 11 \cdot 2}{7} = 8\ l/100Km$$

$$\bar{v} = \frac{20 + 40 + 60 + 80 + 100 + 120 + 140}{7} = 80\ Km/h$$

$$a_{20} = \frac{1}{7}(5^2 \cdot 1 + 6^2 \cdot 2 + 8^2 \cdot 1 + 9^2 \cdot 1 + 11^2 \cdot 2) = \frac{484}{7} = 69{,}14$$

$$a_{02} = \frac{1}{7}(20^2 + 40^2 + 60^2 + 80^2 + 100^2 + 120^2 + 140^2) = \frac{56000}{7} = 8000$$

$$S_c^2 = a_{20} - \bar{c}^2 = 69{,}14 - 8^2 = 5{,}14$$

$$S_v^2 = a_{02} - \bar{v}^2 = 8000 - 80^2 = 1600$$

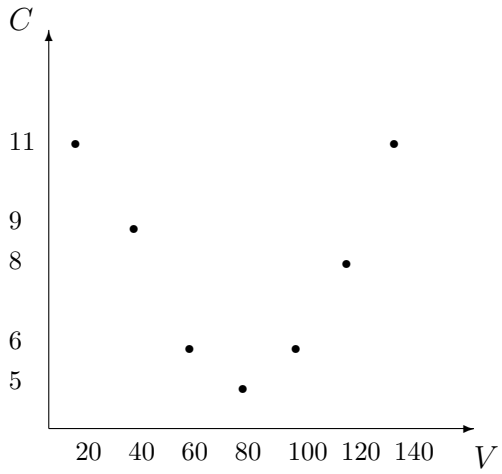
$$a_{11} = \frac{1}{7}(5 \cdot 80 + 6 \cdot 60 + 6 \cdot 100 + 8 \cdot 120 + 9 \cdot 40 + 11 \cdot 20 + 11 \cdot 140) = \frac{4440}{7} = 634{,}28$$

$S_{cv} = a_{11} - \overline{c}\overline{v} = 634{,}28 - 8 \cdot 80 = -5{,}72$

$$r_{cv} = \frac{-5{,}72}{\sqrt{5{,}14}\sqrt{1600}} = \frac{-5{,}72}{90{,}68} = -0{,}063.$$

La relación lineal entre el consumo y la velocidad es prácticamente nula puesto que el coeficiente de correlación es aproximadamente cero.

b) Diagrama de dispersión. A la vista del gráfico las variables consumo y velocidad presentan una relación, aunque no es lineal.



c) Los resultados no son contradictorios ya que se trata de variables dependientes que están relacionadas de una forma no lineal.

Cuestiones

1.B	2.B	3.C	4.E	5.C
6.B	7.E	8.A	9.B	10.E