CENTRO UNIVERSITARIO
DE TECNOLOGÍA Y ARTE DIGITAL

Estadística Descriptiva II

Distribuciones bidimensionales

PROBLEMAS TEMA 3

Mar Angulo Martínez

- **Problema 1**

	lunes	martes	miércoles	jueves	viernes	sábado	
Clientes potenciales (xi)	87	63	70	55	90	105	**470**
Volumen ventas (100 euros) (yi)	120	85	90	63	110	150	**618**
x_i^2	7.569	3.969	4.900	3.025	8.100	11.025	**38.588**
y_i^2	14.400	7.225	8.100	3.969	12.100	22.500	**68.294**
xi.yi	10.440	5.355	6.300	3.465	9.900	15.750	**51.210**

$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$= 470/6 = 78,333 $s_x^2 = \frac{\sum_{i=1}^{n} x_i^2}{n}$ - $\bar{x}^2$= 38.588/6-$(78{,}333)^2$=295,273→ ==s_x=17,18

$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$= 618/6 = 103 $s_y^2 = \frac{\sum_{i=1}^{n} y_i^2}{n}$ - $\bar{y}^2$= 68.294/6-$(103)^2$=773,33 → s_y=27,81

$s_{xy} = \frac{\sum_{i=1}^{n} x_i y_i}{n}$ - $\bar{x}\bar{y}$= 51.210/6-$(78{,}333x103)$ = 466,701

- r= $\frac{s_{xy}}{s_x.s_y}$ =$\frac{466,701}{17,18x27,81}$ =+ 0,9768 Correlación DIRECTA muy fuerte entre las variables,

- Recta de regresión y/x: y-103=1,581 $(x - 78,333)$

  y= 103 +1,581(130-78,333) =184,686--→la predicción de ventas son 18.468, 6 euros para un día en que entran 130 personas en el establecimiento

- Varianza residual $VR = s_e^2 = s_y^2.(1 - R^2)$ =

  773,33(1-0,954)= 35,57

- **Problema 2**

(xi) Antigüedad (años)	1	2	3	2	5	6	3	**22**
(yi) Nº infracciones	2	3	3	5	3	0	1	**17**
x_i^2	1	4	9	4	25	36	9	**88**
y_i^2	4	9	9	25	9	0	1	**57**
xi.yi	2	6	9	10	15	0	3	**45**

$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$= 22/7 = 3,14 años $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$= 17/7 = 2,43 infracciones

$s_x^2 = \frac{\sum_{i=1}^{n} x_i^2}{n}$ - $\bar{x}^2$= 88/7-$(3,14)^2$=2,712 → ==s_x=1,647

$s_y^2 = \frac{\sum_{i=1}^{n} y_i^2}{n}$ - $\bar{y}^2$= 57/7-$(2,43)^2$=2,238 → s_y=1,496

$s_{xy} = \frac{\sum_{i=1}^{n} x_i y_i}{n}$ - $\bar{x}\bar{y}$= 45/7-(3,14x2,43)= -1,2

- a) Si queremos predecir el número de infracciones(y) en función de la antigüedad (x) debemos utilizar la recta de regresión y/x:

  $y-2{,}43=\frac{-1{,}2}{2{,}712}(x-3{,}14)$ ⟶ $y(4)=2{,}43-0{,}44(4-3{,}14)=2{,}05$ infracciones

- b) Si queremos predecir la antigüedad (x) en función del nº infracciones (y) debemos utilizar la recta de regresión x/y:

  $x-3{,}14=\frac{-1{,}2}{2{,}238}(y-2{,}43)$ ⟶ $x(4)=3{,}14-0{,}536(4-2{,}43)=2{,}3$ años de antigüedad

- c) $b_{yx}=\frac{s_{xy}}{s_x^2}=\frac{-1{,}2}{2{,}712}=-0{,}44$ Por cada año más de antigüedad, el nº de infracciones disminuye en 0,44: Aproximadamente, cada dos años más de antigüedad, un conductor tiene una infracción menos.

- d) Variabilidad en el número de infracciones que no queda explicada por la antigüedad de los conductores

Es la varianza residual $s_e^2 = s_y^2 (1 - R^2)$ = 2,238.(1-0,2372)= 1,707 unidades de variabilidad en las infracciones quedan sin explicar

h)

- Coeficiente de correlación r= $\frac{s_{xy}}{s_x . s_y} = \frac{-1,2}{1,647 x 1,496}$ =-0,487 ⟹ R^2 =0,2372
- Interdependencia entre las variables muy débil e inversa (a > antigüedad, < nº de infracciones)
- Valores observados lejos de los valores teóricos y recta de regresión alejada de la nube
- Ajuste bastante malo y predicciones muy poco fiables (23,72%)
- Muy poca (23,72%) variabilidad en las infracciones queda explicada por la antigüedad mediante el ajuste

e) Se plantea bonificar un 10% de la prima del seguro al 15% de conductores con mayor antigüedad. ¿Qué antigüedad mínima ha de tener un conductor para tener acceso a la bonificación?

- Se trata de calcular el P_{85} de la variable X
  - Calculamos $\frac{85n}{100}$= 5,95 ⟹ P_{85}= 5 *años de antigüedad*

(xi) Antigüedad (años)	1	2	3	5	6		
Nº conductores (n_i)	1	2	2	1	1	7	
N_i	1	3	5	6	7		
$x_i^2 n_i$	1	8	18	25	36	88	

a) Al 5% de conductores con mayor nº de infracciones se les va a penalizar con un 10% de recargo en la prima del seguro. ¿Hasta qué número de infracciones un conductor queda libre de dicho recargo?

(xi) Nº infracciones	0	1	2	3	5		
Nº conductores (n_i)	1	1	1	3	1		
N_i	1	2	3	6	7		
$x_i^2 n_i$	0	1	4	27	25		

- f) Se trata de calcular el P_{95} de la variable Y
  - Calculamos $\frac{95n}{100}$= 6,65 $P_{95}= 5\ infracciones$

g) ¿Qué distribución es más homogénea, la de la antigüedad o la de las infracciones?

$\bar{x}$ =3,14 años $\bar{y}$ =2,43 infracciones

s_x=1,647 $CV_X = \frac{1,647}{3,14}$=0,5245

s_y=1,496 $CV_Y = \frac{1,496}{2,43}$=0,6156

- Es más homogénea la distribución de la antigüedad

- **Problema 3**
  - X: coste salarial mensual (en 10^4 euros) Y: indicador de productividad (en puntos)
  - Datos: n=12 $\bar{x} = 17$ $\bar{y} = 12$

    $\sum_{i=1}^{n} x_i^2$=5732 $\sum_{i=1}^{n} y_i^2$=3593 $\sum_{i=1}^{n} x_i\, y_i$ =4135

- Calculamos varianzas, desviaciones típicas y covarianza:

$$s_x^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2 = \frac{5732}{12} - (17)^2 = 188{,}67 \rightarrow == s_x = 13{,}736$$

$$s_y^2 = \frac{\sum_{i=1}^{n} y_i^2}{n} - \bar{y}^2 = \frac{3593}{12} - (12)^2 = 155{,}417 \rightarrow s_y = 12{,}467$$

$$s_{xy} = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \bar{x}\bar{y} = \frac{4135}{12} - 17x12 = 140{,}583$$

- b) Queremos predecir la productividad (y) para un valor conocido del coste salarial (x=20) debemos utilizar la recta de regresión y/x:

  $y-12=\frac{140,583}{188,67}(x-17)$ ⟶ $y-12=+0,745(x-17)$

  Si x=20: y=14,235 puntos de productividad

- a) $r=\frac{s_{xy}}{s_x.s_y}=\frac{140,583}{13,736x12,467}$ =+0,821 Interdependencia directa y fuerte entre ambas variables; la medida en que una explica la otra nos la proporciona el coeficiente de determinación R^2=0,674

Es decir, **el coste salarial explica el 67,4% de la variabilidad en el indicador de productividad de ese grupo de trabajadores**

- c) Para comparar dispersiones utilizamos el CV de Pearson

$$CV_x=\frac{s_x}{\bar{x}}=\frac{13,736}{17}=0,808 \quad < \quad CV_y=\frac{s_y}{\bar{y}}=\frac{12,467}{12}=1,04$$

Es menos dispersa la distribución del coste salarial que la distribución de productividades.

d) **Si hemos de recortar la masa salarial en 3.000 euros ¿Cómo dirías que se verá afectada la productividad de los empleados? Razónalo**

- Si x se reduce en 3 unidades, ¿Qué variación experimenta la y?
- Coeficiente de regresión b_{yx} :
  representa el incremento de y por cada aumento unitario de x:

$$b_{yx} = \frac{s_{xy}}{s_x^2} = \frac{140{,}583}{188{,}67} = 0{,}745$$

- Es decir, por cada 10.000 ptos más de coste salarial, la productividad se incrementa en 0,745 puntos
- **Si la masa salarial se reduce en 0,3 unidades, la productividad caerá en 0,3x0,745=0,22 ptos.**

Problema 4

En una empresa se pretende analizar la relación existente entre: las ventas totales (X) y las exportaciones (Y), expresadas en Millones de euros.
Disponemos de los siguientes datos:
Las dos rectas de regresión se cortan en el punto (6,4).
Por cada millón de euros que se incrementan las exportaciones, el incremento que experimentan las ventas totales, bajo el modelo lineal de Y sobre X, es de 1,15 M euros.
Si se incrementan en un millón de euros las ventas totales, las exportaciones aumentarían en 810.000 euros.
La dispersión relativa de las exportaciones (Y), medida en términos de su coeficiente de variación, ha sido del 50 %.
a) Con estos supuestos, predecir las exportaciones de la empresa en un período en que las ventas totales han sido de 25 Millones de euros y dar una medida de la fiabilidad de dicha predicción
b) Calcular el coeficiente de correlación e interpretarlo
c) Comparar la dispersión relativa de ambas distribuciones
d) Calcular e interpretar la varianza residual y la varianza explicada

- X: ventas totales Y: exportaciones
- $\bar{x} = 6 \quad \bar{y} = 4$
- $b_{xy} = \frac{s_{xy}}{s_y^2} = 1{,}15 \qquad b_{yx} = \frac{s_{xy}}{s_x^2} = 0{,}81$
- Dispersión relativa de Y: $CV_Y = \frac{s_y}{\bar{y}} = 0{,}5 \longrightarrow s_y = 0{,}5 \times 4 = 2 \longrightarrow s_{xy} = 1{,}15 \times 4 = 4{,}6 \longrightarrow s_x^2 = \frac{4{,}6}{0{,}81} = 5{,}68 \longrightarrow s_x = 2{,}38$

a) Con estos supuestos, predecir las exportaciones de la empresa en un período en que las ventas totales han sido de 25 Millones de euros y dar una medida de la fiabilidad de dicha predicción

- Recta de regresión (de y sobre x):

  $y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x})$ ⟹ y-4=0,81(x-6) ⟹ y(25)=19,39 M euros

- La fiabilidad de la predicción es del 93,4% (R^2)

b) Calcular el coeficiente de correlación e interpretarlo

$$r = \frac{s_{xy}}{s_x . s_y} = \frac{4,6}{2,38x2} = +0,966 \Longrightarrow R^2 = 0,934$$

- Interpretación de r:
  - Interdependencia muy fuerte entre las ventas y las exportaciones; y es directa (a mayores ventas, mayores exportaciones, y viceversa
  - Valores teóricos muy próximos a los valores observados y recta de regresión que pasa muy cerca de la nube de puntos
  - El ajuste es por tanto muy bueno y las predicciones que hagamos con él tendrán una alta fiabilidad (93,4%)
  - Rectas de regresión casi coincidentes

c) Comparar la dispersión relativa de ambas distribuciones

$CV_X = \frac{s_x}{\bar{x}} = \frac{2,38}{6} = 0,3967$ $\qquad$ $CV_Y = \frac{2}{4} = 0,5$

- Existe menor dispersión en la distribución de ventas que en la distribución de exportaciones; las ventas medias ofrece por tanto una media más representativa de su distribución

d) Calcular e interpretar la varianza residual y la varianza explicada

- Varianza residual $s_y^2 . (1 - R^2)$ = 4x(1-0,934)=0,264
- Varianza explicada $s_y^2 . (R^2)$ = 4x0,934=3,736
- Sólo 0,264 unidades de variabilidad en las exportaciones no quedan explicadas por la variación en las ventas mediante el ajuste realizado. Esa variabilidad residual (no explicada) representa escasamente un 6,6% del total.

Problema 5

En una empresa se analiza la posible existencia de una relación lineal entre la edad (X) de sus trabajadores y el número de días de baja a lo largo de un año.

Se han obtenido los siguientes datos correspondientes a 5 trabajadores

$\bar{x} = 44{,}2$ $\bar{y} = 8{,}4$ $\sum_{i=1}^{n} x_i^2$=10.439 $\sum_{i=1}^{n} y_i^2$=368 $\sum_{i=1}^{n} x_i\, y_i$ =1.937

a) Obtener una medida de la correlación entre ambas variables e interpretarla
b) Hallar una predicción del tiempo de baja de un trabajador de 32 años
c) ¿Cuál es la variabilidad en los períodos de baja que se explica por la edad de los trabajadores? ¿Qué porcentaje representa esa variabilidad?
d) Calcular la pendiente de la recta de regresión de y sobre x e interpretarla en el contexto del problema
e) Razonar si la edad media es más o menos representativa que el número medio de días de baja para los trabajadores de dicha empresa
f) Si nos facilitan además las edades de los 5 trabajadores de la empresa que son:

50 44 37 60 30

Dar una medida que permita analizar la simetría de la edad de esos trabajadores e interpretarla. Interpretar también los valores que utilices en la fórmula correspondiente.

- Calculamos varianzas, desviaciones típicas y covarianza

$\bar{x} = 44{,}2\ años$ $\bar{y} = 8{,}4\ días$ $\sum_{i=1}^{n} x_i^2$=10.439 $\sum_{i=1}^{n} y_i^2$=368 $\sum_{i=1}^{n} x_i\, y_i$ =1.937

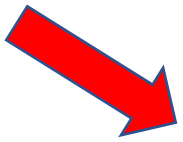
$s_x^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2$= 10.439/5-$(44{,}2)^2$=133,56 → s_x=11,56 años

$s_y^2 = \frac{\sum_{i=1}^{n} y_i^2}{n} - \bar{y}^2$= 368/5-$(8{,}4)^2$=3,04 → s_y=1,74 días

$s_{xy} = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \bar{x}\bar{y}$= 1.937/5-$(44{,}2x8{,}4)$ = 16,12

- a) Interdependencia entre las variables: coeficiente de correlación de Pearson

r= $\frac{s_{xy}}{s_x . s_y} = \frac{16{,}12}{11{,}56x1{,}74}$ =0,8



- Interdependencia relativamente fuerte y directa: el número de días de baja de estos 5 trabajadores es mayor a medida que aumenta la edad.
- Los valores observados están cerca de los teóricos y la recta de regresión pasará por tanto cerca de la nube de puntos
- El ajuste es relativamente bueno y las predicciones tendrán una fiabilidad del 64% (R^2)
- Las dos rectas de regresión forman un ángulo pequeño.

- b) Hallar una predicción del tiempo de baja de un trabajador de 32 años
- Utilizaremos la recta de regresión de y sobre x (queremos predecir el valor de y para x=32)

  $y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x})$ $y-8,4=\frac{16,12}{133,56}(x - 44,2)$ $y=8,4 + 0,12(32 - 44,2)=6,936$

  Aproximadamente 7 días de baja es nuestra predicción para un trabajador de 32 años.

- c) ¿Cuál es la variabilidad en los períodos de baja que se explica por la edad de los trabajadores? ¿Qué porcentaje representa esa variabilidad?

- La variabilidad en el número de días de baja que se explica por la edad de los trabajadores (x) es la varianza explicada $s_y^2 . r^2$ = *3,04x0,64=1,9456 unidades de varianza que representan el 64%de la variabilidad total del producto.*

- d) Calcular la pendiente de la recta de regresión de y sobre x e interpretarla en el contexto del problema
- La pendiente de la recta de regresión se llama coeficiente de regresión
- $b_{yx} = \frac{16,12}{133,56}$=0,12
- Por cada año más de edad, el número de días de baja se incrementa en 0,12

- e) Razonar si la edad media es más o menos representativa que el número medio de días de baja para los trabajadores de dicha empresa

$CV_X = \frac{s_x}{\bar{x}} = \frac{11,56}{44,2} = 0,2615$ $CV_y = \frac{s_y}{\bar{y}} = \frac{1,74}{8,4} = 0,2071$

  - Presenta mayor dispersión la distribución de edades que la de días de baja por tanto la media de esta última es más representativa que la edad media.

a) f) Si nos facilitan además las edades de los 5 trabajadores de la empresa que son:

50 44 37 60 30

Dar una medida que permita analizar la simetría de la edad de esos trabajadores e interpretarla. Interpretar también los valores que utilices en la fórmula correspondiente.

- Utilizamos el coeficiente de Bowley: $A_B = \frac{(Q_3 - Me) - (Me - Q_1)}{Q_3 - Q_1} =$
- Ordenamos los valores 30 37 44 50 60 $\frac{n}{4}$=1,25 Q_1=37 y $\frac{n}{2}$=2,5 Me=44

N_i 1 2 3 4 5 $\frac{3n}{4}$=3,75 Q_3=50

$A_B = \frac{(Q_3 - Me) - (Me - Q_1)}{Q_3 - Q_1} = \frac{(50-44)-(44-37)}{50-37} = -0,077$ Ligerísima asimetría hacia la izquierda

Problema 6

El número de cetano se emplea como indicador de la calidad de ignición del combustible utilizado en un motor diésel.

En un estudio se trató de ajustar mediante un modelo lineal dicho indicador (Y) en función del índice de yodo X (en gramos)

Se analizó una muestra de 14 combustibles y se obtuvieron los siguientes cálculos:

$\sum_{i=1}^{n} x_i$=1.307,5 $\sum_{i=1}^{n} y_i$=779,2 $\sum_{i=1}^{n} x_i^2$=128.913,93

$\sum_{i=1}^{n} y_i^2$=43.745,22 $\sum_{i=1}^{n} x_i\, y_i$ =71.347,3

a) Predecir utilizando un modelo lineal el número de cetano para un combustible que tiene un índice de yodo de 115 y explicar el significado del coeficiente de regresión en ese modelo.
¿Cuál es la fiabilidad de dicha predicción?

a) Dar una medida de la bondad del ajuste e interpretarla

c) ¿Cuánta variabilidad en el número de cetano no se explica por el índice de yodo? Indicar también qué % representa ese valor

Se han obtenido los siguientes datos adicionales sobre el número de cetano de la muestra de 14 combustibles

i. La mitad de ellos tiene un índice de cetano comprendido entre 87 y 114
ii. La mitad de los combustibles de la muestra tienen un índice de cetano menor o igual que 100
iii. Los índices de cetano mínimo y máximo obtenidos en la muestra de combustibles han resultado de 51 y 173
iv. El índice de cetano más frecuente ha sido de 92.

d) Construir un box plot para el número de cetano e interpretarlo

e) ¿Qué valor medio es más representativo, el del número de cetano o el del índice de yodo? Explicar por qué.

- Calculamos medias, varianzas, desviaciones típicas y covarianza

$\sum_{i=1}^{n} x_i$=1.307,5 $\sum_{i=1}^{n} y_i$=779,2 $\sum_{i=1}^{n} x_i^2$=128.913,93 $\sum_{i=1}^{n} y_i^2$=43.745,22 $\sum_{i=1}^{n} x_i\, y_i$ =71.347,3

$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1.307,5}{14} = 93,39\ grs\ de\ yodo$ $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{779,2}{14} = 55,66\ número\ medio\ de\ cetano$

$s_x^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2$= 128.913,93/14-$(93,39)^2$=486,446 → s_x=22,056 grs

$s_y^2 = \frac{\sum_{i=1}^{n} y_i^2}{n} - \bar{y}^2$= 43.745,22/14-$(55,66)^2$=26,61 → s_y=5,16

$s_{xy} = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \bar{x}\bar{y}$= 71.347,3/14-(93,39$x$55,66) = -101,85

- a) Predecir utilizando un modelo lineal el número de cetano para un combustible que tiene un índice de yodo de 115 y explicar el significado del coeficiente de regresión en ese modelo. ¿Cuál es la fiabilidad de dicha predicción?
- Queremos predecir el valor de Y cuando el índice de yodo (X) es de 115: sustituiremos en la rectas de regresión de y sobre x

$y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x})$ $y-55,66=\frac{-101,85}{486,446}(x - 93,39)$ $y=55,66 - 0,209(115 - 93,39)=51,143$

- b) Dar una medida de la bondad del ajuste e interpretarla
- Una medida de la bondad del ajuste nos la da el coeficiente de correlación lineal de Pearson

$$r=\frac{s_{xy}}{s_x.s_y}=\frac{-101,85}{22,056x5,16}=-0,89$$

- Interdependencia inversa bastante fuerte: el número de cetano disminuiría al aumentar el índice de yodo.
- Los valores observados están cerca de los teóricos y la recta de regresión pasará bastante cerca de la nube de puntos
- El ajuste es bastante bueno y las predicciones tendrán una fiabilidad del 79,21%
- Las dos rectas de regresión forman un ángulo pequeño.

- c) ¿Cuánta variabilidad en el número de cetano no se explica por el índice de yodo?

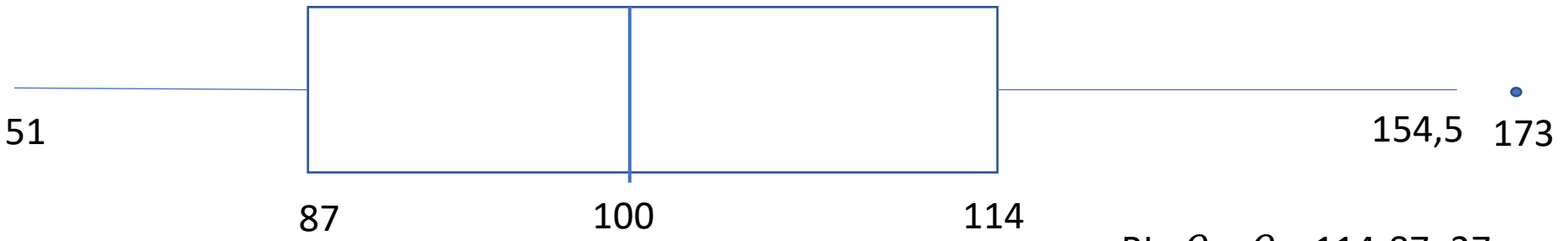
  Indicar también qué % representa ese valor

- La variabilidad en el número de cetano que no queda explicada por el índice de yodo mediante el ajuste (x) es la varianza residual y será prácticamente toda la s_y^2
- *Varianza residual* $= s_y^2$ (1-r^2) = *26,61x(1-0,7921)=5,53 unidades de varianza que representan el 20,79% de la variabilidad total del producto.*

- d) Construir un box-plot para el número de cetano e interpretarlo

  i. La mitad de ellos tiene un índice de cetano comprendido entre 87 y 114
  ii. La mitad de los combustibles de la muestra tienen un índice de cetano menor o igual que 100
  iii. Los índices de cetano mínimo y máximo obtenidos en la muestra de combustibles han resultado de 51 y 173
  iv. El índice de cetano más frecuente ha sido de 92.

i. Sabemos que el 50% central de la distribución está comprendido entre Q_1=87 y Q_3 =114
ii. La mitad de los combustibles de la muestra tienen un índice de cetano menor o igual que Me=100
iii. Los índices de cetano mínimo y máximo obtenidos en la muestra de combustibles han resultado de 51 y 173
iv. El índice de cetano más frecuente es la Md=92.



RI= Q_3- Q_1=114-87=27

- Calculamos Lím. Inferior = Q_1 – 1.5 RI=87-1,5x27 = 46,5 y Lím. Sup = Q_3 + 1.5 RI=114+1,5x27=154,5
- No hay valores atípicos por la izquierda: trazamos línea hasta el valor más pequeño que es > Lím. Inferior
- El valor máximo es atípico porque supera el lím. Superior: trazamos línea hasta el límite y marcamos el valor atípico. No es un valor extremo porque no supera Q_3 + 3 RI=114+3x27=195

- e) ¿Qué valor medio es más representativo, el del número de cetano o el del índice de yodo?

  Explicar por qué.

  $CV_X = \frac{s_x}{\bar{x}} = \frac{22,056}{93,39} = 0,2362$ > $CV_y = \frac{s_y}{\bar{y}} = \frac{5,16}{55,66} = 0,0927$

  - Es más representativa la media en el número de cetano puesto que la dispersión de la muestra es menor.

U-tad
CENTRO UNIVERSITARIO
DE TECNOLOGÍA Y ARTE DIGITAL

Problema 7

Un artículo especializado analizó cómo la intensidad del enfado en los berrinches de los niños podría estar relacionada con la duración de la rabieta. Se analizó para un grupo de niños un indicador Y (en puntos) medidor de la intensidad (gritar, empujar o tirar objetos) en función de la duración del berrinche medida en minutos (X).

I_i	[0,2)	[2,4)	[4,11)	[11,20)	[20,30)	[30,40)
y_i	136	92	71	26	7	3

a) Obtener una medida de la correlación entre ambas variables e interpretarla
b) Hallar una predicción del tiempo que duró la rabieta en un niño en el que se obtuvo un indicador de 81 indicando la fiabilidad de dicha predicción
c) ¿Cuál es la variabilidad en la intensidad de las rabietas de ese grupo de niños que se explica por el tiempo que duran? ¿Qué porcentaje representa esa variabilidad?
d) ¿Cuánto aumenta/disminuye el indicador de intensidad por cada 2 minutos más de berrinche?
e) ¿Cuál es el indicador medio de intensidad del enfado? Razonar si es o no un valor representativo de los datos
f) Calcular e interpretar en el contexto del problema el primer cuartil de X y el 90 Percentil de Y.

(xi) Duración del berrinche (minutos)	[0,2)	[2,4)	[4,11)	[11,20)	[20,30)	[30,40)	**22**
(yi) Intensidad del berrinche (puntos)	136	92	71	26	7	3	**17**
x_i^2	1	9	56,25	240,25	625	1225	**2.156,25**
y_i^2	18.496	8.464	5.041	676	49	9	**32.735**
xi.yi	136	276	532,5	403	175	105	**1.627,5**

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{87}{6} = 14{,}5\ minutos \qquad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{779{,}2}{14} = 55{,}83\ \ ptos\ de\ intensidad$$

$s_x^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2$ = 2.156,25/6-$(14{,}5)^2$=149,125→ s_x=12,21 minutos

$s_y^2 = \frac{\sum_{i=1}^{n} y_i^2}{n} - \bar{y}^2$ = 32.735/6-$(55{,}83)^2$=2.338,84 → s_y=48,36 puntos

$s_{xy} = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \bar{x}\bar{y}$ = 1.627,5/6-$(14{,}5x55{,}83)$ = -538,285

- a) Obtener una medida de la correlación entre ambas variables e interpretarla

$$r=\frac{s_{xy}}{s_x.s_y}=\frac{-538,285}{12,21x48,36}=-0,9116$$

- Interdependencia fuerte e inversa: la intensidad del berrinche disminuye cuando la duración del mismo es mayor.
- Los valores observados están cerca de los teóricos y la recta de regresión pasará por tanto cerca de la nube de puntos
- El ajuste es muy bueno y las predicciones tendrán una fiabilidad del 83,1% (R^2)
- Las dos rectas de regresión forman un ángulo próximo a 0º.

- a) Hallar una predicción del tiempo que duró la rabieta en un niño en el que se obtuvo un indicador de 81 indicando la fiabilidad de dicha predicción

- Queremos predecir el valor de X cuando la intensidad del berrinche(Y) es de 81: sustituimos en la rectas de regresión de x sobre y

$$x-\bar{x}=\frac{s_{xy}}{s_y^2}(y-\bar{y}) \qquad x-14,5=\frac{-538,285}{2.338,84}(y-55,83) \qquad x=14,5-0,23(81-93,39)=17,35 \text{ minutos}$$

- c) ¿Cuál es la variabilidad en la intensidad de las rabietas de ese grupo de niños que se explica por el tiempo que duran? ¿Qué porcentaje representa esa variabilidad?
  - La variabilidad en la intensidad de las rabietas que se explica por el tiempo que duran mediante el ajuste lineal es la varianza explicada
  - *Varianza explicada* $= s_y^2\ r^2$ = *2.338,84x0,831=1.943,576 unidades de varianza que representan el* 83,1 *% de la variabilidad total de la intensidad.*

- d) ¿Cuánto aumenta/disminuye el indicador de intensidad por cada 2 minutos más de berrinche?
  - $b_{yx}=\frac{s_{xy}}{s_x^2}=\frac{-538,285}{149,125}$ =-3,61
  - Por cada minuto más de berrinche el indicador de intensidad disminuye en 3,61 ptos de intensidad

- e) ¿Cuál es el indicador medio de intensidad del enfado? Razonar si es o no un valor representativo de los datos

  - $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{779,2}{14} = 55,83 \ \ ptos\ de\ intensidad$

  - $CV_y = \frac{s_y}{\bar{y}} = \frac{48,36}{55,83} = 0,8662 < 1$ La media es representativa de los datos porque la dispersión no es grande

- f) Calcular e interpretar en el contexto del problema el primer cuartil de X y el 90 Percentil de Y.

(xi) Duración del berrinche (minutos)	[0,2)	[2,4)	[4,11)	[11,20)	[20,30)	[30,40)
N_i	1	2	3	4	5	6

- Calculamos el Q_1 de la variable X: $\frac{n}{4} = 1{,}5 \qquad Q_1 \in [2,4)$
- $Q_1 = L_{i-1} + \frac{n/4 - N_{i-1}}{n_i} |L_i - L_{i-1}| = 2 + \frac{1.5-1}{1} x2 = 5$

Los berrinches de la cuarta parte de los niños de la muestra duran como mucho 5 minutos

(yi) Intensidad del berrinche (puntos)	3	7	26	71	92	136
N_i	1	2	3	4	5	6

- Calculamos el P_{90} de la variable Y: $\frac{90n}{100} = 5{,}4 \qquad P_{90} = 136$

El 90 por ciento de los niños de la muestra tienen berrinches de intensidad máxima 136 puntos.

Problema 8

Clase /Tipo tren	Via estrecha	Convencional	Alta velocidad	
Turista	400	600	800	1800
Preferente	100	400	700	1200
	500	1000	1500	3000

- 800 =n_{13} es una frecuencia conjunta, es el nº de viajeros que van en clase turista y en tren de alta velocidad
- La proporción de billetes de clase turista es la frecuencia relativa marginal $f_{1.}$=$\frac{1800}{3000} = 60\%$
- n.2=1000 y es el número total de viajeros que utilizan un tren convencional
- **Proporción** de los billetes **vendidos en preferente** que son de alta velocidad.

Clase /Tipo tren	Via estrecha	Convencional	Alta velocidad	
Turista	400	600	800	1800
Preferente	100	400	700	1200
	500	1000	1500	3000

	Clase condicionada a tren vía estrecha	
	ni/j=1=1	fi/j=1 (%)
Turista	400	80%
Preferente	100	20%
500		

Son 700 de un total de 1200, luego la proporción es de $f_{3/i=2}$ =700/1200. Es una frecuencia relativa condicionada

- Proporción de billetes vendidos que son de clase turista y de tipo convencional.

Es una frecuencia relativa conjunta: f_{12}=$\frac{600}{3000} = 20\%$

- Distribución de frecuencias relativas marginales para la variable clase
- Distribución de frecuencias relativas de la variable clase condicionada a tren de vía estrecha

CENTRO UNIVERSITARIO
DE TECNOLOGÍA Y ARTE DIGITAL

Problema 9

Arreglo Y→ Entrada X↓	Arreglos de chapa	Arreglos eléctricos	Arreglos mecánicos	$n_{i.}$
mañana	37	81	100	218
tarde	123	95	47	265
$n_{.j}$	160	176	147	n=483

- Obtener la distribución marginal de la variable "entrada":Y
- ¿Cuál es la frecuencia conjunta de "tarde" y "arreglos eléctricos"? n_{22}=95
- ¿y cuál es la proporción de clientes que llegan de "tarde" y requieren "arreglos eléctricos"? $f_{22}=\frac{95}{483}*100=19{,}67\%$
- ¿Cuál es el número total de "arreglos de chapa" y qué proporción representan sobre el total de vehículos atendidos este mes en el taller? Indicar cuál es el concepto estadístico al que corresponde cada uno de ellos.
- Se trata de la frecuencia absoluta y relativa marginal de "chapa": son 160 vehículos y representan el 33,13% del total. son las frecuencias marginales ($n_{.1}$ y $f_{.1}$) absoluta y relativa, respectivamente.

Problema 9

- Obtener la frecuencia absoluta marginal de "tarde" e indicar su significado

$n_{2.} = 265$ vehículos han entrado por la tarde

- Frecuencia relativa marginal de "tarde" y expresar su significado
- Son $f_{2.} = \frac{265}{483}$ *100=54,87% de los vehículos han entrado por la tarde
- Obtener la distribución de frecuencias de arreglos mecánicos condicionada a "mañana":

	Tipo de arreglo condicionado a turno de mañana $n_{j/i=1}$	$f_{j/i=1}$
Arreglos de chapa	37	37/218=16,97%
Arreglos eléctricos	81	81/218=37,16%
Arreglos mecánicos	100	100/218=45,87%

- ¿Qué proporción de coches que entraron por la mañana, requirieron de arreglos mecánicos? Un 45,87%

- Obtener el valor de f.3 e indicar su significado en el contexto del problema

Es la proporción de vehículos que han requerido arreglos mecánicos y se obtiene $f_{.3}$=$\frac{147}{483}$ *100=30,43%

- Calcular el valor de n2. e indicar qué significa en el contexto del problema

Es el número de vehículos que han entrado de tarde y se obtiene $n_{2.}$=$265\ vehículos$

Problema 9

- Proporción de vehículos que han entrado en el taller por la tarde y requieren arreglos de chapa. Calcular el valor e indicar el concepto estadístico correspondiente

Es la frecuencia relativa conjunta $f_{21} = \frac{123}{483} * 100 = 25{,}47\%$

- Proporción de vehículos que, habiendo entrado en el taller por la tarde, necesitan reparación de chapa. Calcularla e indicar el concepto estadístico

Es la frecuencia relativa condicionada $f_{2/i=1} = \frac{123}{265} * 100 = 46{,}42\%$

CENTRO UNIVERSITARIO
DE TECNOLOGÍA Y ARTE DIGITAL

Problema 10

Disponemos de los siguientes datos del PIB municipal per cápita X (en miles de euros) y la tasa de paro (en %) (Y) correspondientes a 15 municipios

$\bar{x} = 21{,}2$ $\bar{y} = 10{,}5$ $\sum_{i=1}^{n} x_i^2$=7.033 $\sum_{i=1}^{n} y_i^2$=2.158 $\sum_{i=1}^{n} x_i\, y_i$ =2.964,12

a) Obtener una medida de la interdependencia entre ambas variables e interpretarla
b) Obtener una predicción de la tasa de paro de un municipio con un PIB municipal per cápita de 23.300 euros
c) ¿Cuál es la variabilidad en la tasa de paro que no se explica por el PIB per cápita de los municipios? ¿Qué porcentaje representa esa variabilidad?
d) Calcular la pendiente de la recta de regresión de y sobre x e interpretarla en el contexto del problema

Si nos facilitan además las tasas de paro de los 15 municipios que son:

18,6 6,7 11,2 9,3 11,4 10,6 11,2 6,7 11,2 9,3 10,7 9,3 11,2 9,3 5,7

a) Dibujar un boxplot y extraer conclusiones acerca de la tasa de paro en esos municipios.

CENTRO UNIVERSITARIO
DE TECNOLOGÍA Y ARTE DIGITAL

Problema 10

$s_x^2 = \frac{\sum_{i=1}^{n} x_i^2}{n}$ - $\bar{x}^2$= 7.033/15-$(21{,}2)^2$=19,43 → s_x=4,41 m. euros

$s_y^2 = \frac{\sum_{i=1}^{n} y_i^2}{n}$ - $\bar{y}^2$= 2.158/15-$(10{,}5)^2$=33,62 → s_y=5,8 %

$s_{xy} = \frac{\sum_{i=1}^{n} x_i y_i}{n}$ - $\bar{x}\bar{y}$= 2.964,12/15-$(21{,}2x10{,}5)$ = -24,99

- a) Obtener una medida de la correlación entre ambas variables e interpretarla

$$r=\frac{s_{xy}}{s_x.s_y}=\frac{-24{,}99}{4{,}41x5{,}8}=-0{,}977$$

- Interdependencia muy fuerte e inversa: la tasa de paro disminuye a medida que aumenta el PIB.
- Los valores observados están muy cerca de los teóricos y la recta de regresión pasará por tanto muy cerca de la nube de puntos
- El ajuste es muy bueno y las predicciones tendrán una fiabilidad del 95,46% (R^2)
- Las dos rectas de regresión forman un ángulo muy próximo a 0º.

- a) Obtener una predicción de la tasa de paro de un municipio con un PIB municipal per cápita de 23.300 euros
- Queremos predecir el valor de Y cuando el PIB (X) es de 23.300: sustituimos en la recta de regresión de y sobre x

$$y-\bar{y}=\frac{s_{xy}}{s_x^2}(x-\bar{x}) \qquad y-10{,}5=\frac{-24{,}99}{19{,}43}(x-21{,}2) \qquad y-10{,}5=-1{,}286\,(23{,}3-21{,}2)=7{,}8\%$$

Pendiente de la recta=-1,286: Por cada 1000 euros más de PIB per cápita la tasa de paro disminuye un 1,286%.

Problema 10

c) ¿Cuál es la variabilidad en la tasa de paro que no se explica por el PIB per cápita de los municipios? ¿Qué porcentaje representa esa variabilidad?

- *Varianza explicada* $= s_y^2$ (1-r^2) = *33,62x(1-0,9543)=1,536 unidades de varianza que representan el 4,57 % de la variabilidad total de la tasa de paro.*

d) Datos: 18,6 6,7 11,2 9,3 11,4 10,6 11,2 6,7 11,2 9,3 10,7 9,3 11,2 9,3 5,7

Ordenamos de menor a mayor: 5,7 6,7 6,7 9,3 9,3 9,3 9,3 10,6 10,7 11,2 11,2 11,2 11,2 11,4 18,6

- Calculamos los tres cuartiles: Q_1 , Me, Q_3

$\frac{n}{4}$=3,75 ➡ Q_1=9,3 $\frac{n}{2}$=7,5 ➡ Me=10,6 $\frac{3n}{4}$=11,25 ➡ Q_3=11,2

- Lím inf =Q_1 -1,5RI=9,3-1,5x1,9=6,45 Lím sup =Q_3 +1,5RI=11,2+1,5x1,9=14,05