Cuaderno de actividades 1º

1. *INTRODUCCIÓN: Variables estadísticas bidimensionales.*

En numerosas ocasiones interesa estudiar simultáneamente dos (o más) caracteres de una población. En el caso de dos (o más) variables estudiadas conjuntamente se habla de **variable bidimensional** (multidimensional); si se trata de dos caracteres cualitativos, de **par de atributos.**

Si de un cierta población se estudian dos caracteres simultáneamente se obtienen dos series de datos.

Individuos	**A**	**B**	**C**	**.....**
Carácter X	x_1	x_2	x_3	.....
Carácter Y	y_1	y_2	y_3	.....

La lista de pares de datos correspondientes a cada individuo de la población (repetidos o no), es lo que llamamos variable estadística bidimensional.

Ejemplo 1. A cada uno de los reclutas de un reemplazo se les talla y pesa. Se trata de dos variables cuantitativas.

(tallas en m)	**1,70**	**1,70**	**1,69**	**1,68**	**.....**
y_i (peso en kg)	**67**	**75**	**70**	**66**	**.....**

Ejemplo 2. Entre los empleados de una empresa se ha realizado una encuesta sobre el consumo del tabaco, que ha arrojado los siguientes resultados:

Hábito / Sexo	**Fumadores**	**No fumadores**	**Totales de filas**
Varones	49	64	113
Mujeres	43	37	80
Totales de columnas	92	101	Total general **193**

Nota. En este tema nos limitaremos al estudio de **caracteres cuantitativos discretos**, puesto que si el carácter es continuo o discreto agrupado en intervalos, se trabajará con las marcas de clase.

2. *Distribuciones de frecuencias.*

Se disponen las frecuencias en una tabla de doble entrada donde las x_i y la y_j están ordenadas en forma creciente. Recibe el nombre de tabla de frecuencias o **tabla de correlación**.

Si hay pares que se repiten se agrupan siendo n_{ij} la frecuencia absoluta del par (x_i, y_j).

Las sumas:

$\sum_j n_{ij} = n_i$, frecuencia absoluta de x_i.

$\sum_i n_{ij} = n'_j$, frecuencia absoluta de y_j

se llaman frecuencias absolutas marginales de las variables X e Y respectivamente.

$\sum_j \sum_i n_{ij} = N$ = número total de pares.

X / Y	x_1	x_2	.....	x_k	Frec. absolutas marginales de **Y**
y_1	n_{11}	n_{21}	.....	n_{k1}	n'_1
y_2	n_{12}	n_{22}	.....	n_{k2}	n'_2
.....	.....	....	....	...	....
y_r	n_{1r}	n_{2r}	...	n_{kr}	n'_r
Frec. absolutas marginales de **X**	n_1	n_2	..	n_k	$\sum_i \sum_j n_{ij} = N$

En la práctica algunas de las n_{ij} pueden ser cero. En tal caso la casilla correspondiente se dejará en blanco.

Ejemplo 3. Dada la distribución bidimensional:

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

la tabla correspondiente es:

X / Y	1	2	3	Frec. absolutas marginales de y
2	1			1
3	2	2		4
4		1		1
5		2	2	4
Frec. absolutas marginales de X	3	5	2	N=10

◆ Al estudiar una variable bidimensional se obtienen varias distribuciones unidimensionales, según se consideren las filas o las columnas de la tabla en estudio.

Las distribuciones unidimensionales del total de los individuos de la población, respecto a cada una de las características reciben el nombre de **distribuciones marginales.**

Distribución marginal de la Y:

Y	Frec. absolutas marginales de Y
y_1	n'_1
y_2	n'_2
.	.
.	.
y_r	n'_r

Análogamente la distribución marginal de la X

Ejemplo 4.

Obtener la distribución marginal de la variable X.

X	Frec. absolutas marginal de X
1	**3**
2	**5**
3	**2**

◆ Si en la tabla de correlación consideramos la primera columna y una columna intermedia, la correspondiente a y_j, se obtiene una distribución unidimensional que llamaremos **distribución condicionada de la variable X** por la modalidad y_j de la variable Y.

X	Frec. absolutas condicionadas por y_j
x_1	n_{1j}
x_2	n_{2j}
.	.
.	.
x_k	n_{kj}

Análogamente se define la **distribución condicionada de la variable Y** por la modalidad x_i de la variable X.

Ejemplo 5.

Obtener la tabla de la distribución condicionada de la variable Y por la modalidad x_2.

Y	Frec. absolutas condicionadas por x_2
2	0
3	2
4	1
5	2

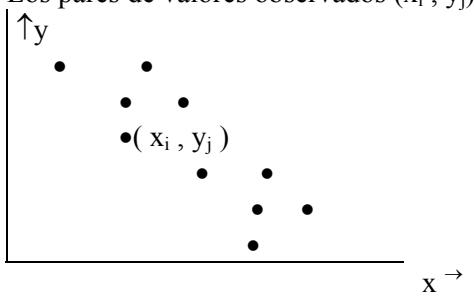
3. *Representaciones gráficas.*

Consideremos la distribución:

x_1	x_2	.....	x_N
y_1	y_2	.....	y_N

(Los pares pueden estar repetidos)

Los pares de valores observados (x_i , y_j) se pueden representar en unos ejes de coordenadas,.



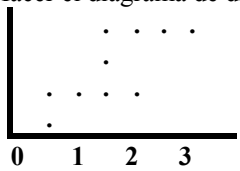
El conjunto de puntos que resulta se llama diagrama de dispersión o **nube de puntos** de la distribución bidimensional.

Cuando el número de datos es grande (se usa una tabla de doble entrada) los datos se representan con un diagrama de dispersión reticulado de tal manera que la visión de la nube de puntos indique realmente cómo es la distribución.

En estos casos también se suele usar un diagrama de barras sobre un sistema cartesiano de tres dimensiones (**estereogramas**).

Ejemplo 6.

Hacer el diagrama de dispersión de la distribución del ejemplo 3.



Ejercicio 1. Dibuja el estereograma correspondiente .

4. *Parámetros de la variable estadística bidimensional.*

Considerando las distribuciones marginales, como son unidimensionales es posible calcular los siguiente parámetros:

Llamadas *medias marginales*.

a) Medias

$$\bar{x} = \frac{\sum x_i n_i}{N} \qquad \bar{y} = \frac{\sum y_j n'_j}{N}$$

Donde $N = \sum n_i = \sum n'_j$ es el numero total de pares.

Nota. En una distribución bidimensional al punto (**x, y**) se le llama centro de gravedad de la distribución.

b) **Varianzas**

Se define:

$$S_x^2 = \frac{\sum (x_i - \bar{x})^2 n_i}{N} = \frac{\sum x_i^2 n_i}{N} - \bar{x}^2$$

Varianza marginal de la variable X

(Es decir la "media del cuadrado menos el cuadrado de la media")

Análogamente la varianza marginal de la variable Y. De ellas (extrayendo la raíz cuadrada) se obtienen las correspondientes **desviaciones típicas**.

Ejemplo 7. Calcula las medias marginales y las Varianzas de la v.e. del ejemplo 3.

Solución **x** = 19/10=1,9 ; **y** =38/10= 3,8 ; S_x^2= 4,1-(1,9)^2= 0,49 ; S_y^2= 15,6 - 14,44=1.16.

c) **Covarianza**

Para las variables estadísticas bidimensionales se define la " *covarianza* '' como la media aritmética de los productos de las desviaciones respecto de la media de cada una de las variables componentes. Es decir :

$$S_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y}) n_{ij}}{N}$$

Se demuestra que

$$S_{xy} = \frac{\sum x_i y_j n_{ij}}{N} - \overline{xy}$$

propiedad que facilita el cálculo de la covarianza. (**Ver problema resuelto 2**)

Ejemplo 8. Calcula la covarianza de la distribución del ejemplo 3.

Solución : $S_{xy} = \frac{2+6+12+8+20+30}{10}$ - (1,9)·(3,8) = 0,58.

4. *Regresión lineal.*

Al considerar los dos caracteres de una variable bidimensional puede ocurrir.

♠ Que exista una **dependencia funcional** entre ellos, de tal manera que a cada valor le corresponda un único valor del otro. Ejemplo: la temperatura a la que calentamos una barra de hierro y la longitud alcanzada.

♠ Que haya una **dependencia estadística** o correlativa, de tal manera que los valores sigan unas pautas similares. Por ejemplo el *número de horas* de estudio y las *notas obtenidas.*

♠ Que se de una **independencia** entre los caracteres. Por ejemplo la estatura y las calificaciones en Matemáticas.

El estudio de la relación entre dos caracteres de una variable estadística bidimensional es el objeto de la **regresión lineal.**

La nube de puntos de una distribución bidimensional nos da una primera idea de la relación existente entre los datos de la misma.

Cuando la nube de puntos del diagrama de dispersión permita deducir algún tipo de dependencia entre las dos variables X, Y, concentrándose los puntos alrededor de una cierta línea (línea de regresión) se plantean dos cuestiones:

A) ***Definir la línea.***

B) ***Medir*** *el nivel de aproximación* de dicha línea.

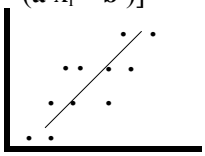
Sí la línea es una recta, el problema es un caso típico de regresión lineal.

A***) Rectas de regresión.***

Se llama **recta de regresión** a aquella que mejor se ajusta a la nube de puntos.
El procedimiento más usado, para hallar dicha recta, es el de los mínimos cuadrados.
Se calcula la recta:

$y = \mathbf{a}x + \mathbf{b}$, de tal manera que:

$S=\sum [y_i - (\mathbf{a}\, x_i + \mathbf{b})]^2$ sea mínima



El cálculo de **a** y **b** incluye conocimientos que no se dan en este nivel[1] por lo que sólo daremos el resultado:

Se verifica:

$$\mathbf{a} = \frac{S_{xy}}{S_x^2} \qquad \mathbf{b} = \bar{y} - \frac{S_{xy}}{S_x^2}\bar{x}$$

luego se puede escribir: $y = \frac{S_{xy}}{S_x^2}x + \bar{y} - \frac{S_{xy}}{S_x^2}\bar{x}$ o lo que es igual

$$y - \bar{y} = \frac{S_{xy}}{S_x^2}(x - \bar{x})$$

Esta es la ecuación de la **recta de regresión de Y sobre X.** Sirve para hacer estimaciones o predicciones de los valores de Y conocidos los de X.

Análogamente la recta de regresión de X sobre Y tiene por ecuación:

$$x - \bar{x} = \frac{S_{xy}}{S_y^2}(y - \bar{y})$$

A $m_{xy} = \frac{S_{xy}}{S_x^2}$ **y** $m_{xy} = \frac{S_{xy}}{S_y^2}$ se les llama **coeficientes de regresión**

Ejemplo 10. Hallar las rectas de regresión para la distribución del ejemplo 3.
Solución : recta de regresión de Y sobre X $y - 3{,}8 = 1{,}18\,(x - 1{,}9)$
recta de regresión de X sobre Y $x - 1{,}9 = 0{,}5\,(y - 3{,}8)$.

Nota. Daremos **sin demostración** algunas propiedades del coeficiente de regresión que facilitan los cálculos de estos, pues permiten hacer un cambio de variable.

Propiedades del coeficiente de regresión:

1) Si se suma o resta una constante a todos los valores de X o de Y el coeficiente de regresión $m_{y\,x}$ no varía.

2) Si se multiplican todos los valores de X por una constante, el coeficiente de regresión queda dividido por esa constante.

Si se multiplican todos los valores de y por una constante, el coeficiente de regresión $m_{y\,x}$ queda multiplicado por es constante.

Ejemplo 11. Consideramos la tabla:

1980	430000
1983	450000
1986	475000
1989	500000

[1] La derivación parcial.

Si hacemos $X'=\frac{X-1980}{3}$, $Y'=\frac{Y-450000}{1000}$

Se obtiene :

0	-20
1	0
2	25
3	50

Para la variable X' ,Y' es más fácil el cálculo del coeficiente de regresión y la relación entre éste y el de XY es:

$$m'_{yx} = \frac{3m_{xy}}{1000}$$

B) ***Correlación lineal.***

Se entiende por **correlación** la dependencia que existe entre las variables de una distribución., cuando ésta es, en cierta forma, lineal se habla de *correlación lineal*. Cuando no existe tal dependencia se dice que las variables están *incorreladas***.**

Para medir, de una forma cuantitativa, dicha dependencia se utiliza el llamado **coeficiente de correlación lineal, o de Pearson,** que se define así:

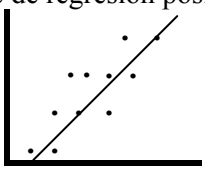
$$r = \frac{S_{xy}}{S_x . S_y} = \sqrt{m_{yx} . m_{xy}}$$

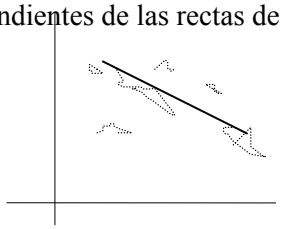
El signo es + si la covarianza es positiva y - si es negativa..

Propiedades de ***r***

a) $-1 \le r \le 1$

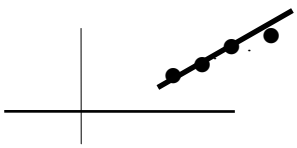
b) Si **r** es positivo la correlación **es directa**, es decir, al aumentar una variable también aumenta la otra (coeficiente de regresión positivo). En este caso las pendientes de las rectas de regresión son positivas.





Si **r** es negativo la correlación es **inversa,** es decir**,** al aumentar una variable disminuye la otra. En este caso las pendientes de la rectas de regresión son negativas.

c) Si $\mathbf{r}^2 = 1$, es decir, **r** igual a 1 o a -1, las dos rectas de regresión coinciden y la nube de puntos está contenida en la recta (correlación perfecta). Hay dependencia funcional entre las variables.



d) Si **r** = 0 las rectas de regresión son perpendiculares entre sí y paralelas a los ejes. Las variables son incorreladas.

Para los demás valores de **r** la dependencia es tanto más fuerte cuanto más próximo esté a 1 o a -1. Será más débil cuando se aproxime a 0:

Para la correlación directa:

Si $0{,}75 \le \mathbf{r} \le 1$ correlación muy alta.

Si $0{,}40 \le \mathbf{r} \le 0{,}75$ correlación baja

Si **r** < 0,40 la correlación es casi despreciable .

Ejemplo 12. Hallar el coeficiente de correlación lineal para la distribución del ejemplo 3.

Solución : **r** =+ $\sqrt{(1{,}18).(0{,}5)}$ = 0,76. Se trata de una correlación directa alta.

Problemas resueltos

1. Una asociación dedicada a la protección de la infancia decide estudiar la relación entre la mortalidad infantil en cada país y el número de camas de hospitales por cada mil habitantes.. Datos

x	50	100	70	60	120	180	200	250	30	90
y	5	2	2,5	3,75	4	1	1,25	0,75	7	3

Donde **x** es el nº de camas por mil habitantes e **y** el tanto por ciento de mortalidad.

Se pide calcular las rectas de regresión y el coeficiente de correlación lineal.

¿ Si se dispusiese de 175 camas por mil habitantes que tanto por ciento de mortalidad cabria esperar?. ¿La estimación es fiable? Razona la respuesta.

Solución :

Para facilitar los cálculos de los parámetros se utiliza la siguiente tabla:

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
50	5	2500	25	250
100	2	10000	4	200
70	2,5	4900	6,25	170
60	3,75	3600	14,0625	225
120	4	14400	16	480
180	1	32400	1	180
200	1,25	40000	1,5625	250
250	0,75	62500	0,5625	187,5
30	7	900	49	210
90	3	8100	9	270
∑ =	1150	30,25	179300	126,4375	2422,5

x =115; **y** = 3,025%; $S_x = \sqrt{17930 - 13225} =$ 68,59; $S_y = \sqrt{12{,}64375 - 9{,}150625} =$ 1,87 ; $S_{xy} = 242{,}25 - (115)(3{,}025) =$ -105,625

Las rectas de regresión serán por tanto:

y - 3,025 = -0,022449 (x - 115)

x - 115 = -30,2053 (y - 3,025)

El coeficiente de correlación lineal:

$$\mathbf{r} = \frac{-105{,}625}{(68{,}59)(1{,}87)} = -0{,}8235$$

es una correlación inversa alta .

Para la estimación que nos piden utilizaremos la recta de regresión de Y sobre X.

y= 3,025 - 0,022449(175- 115) = 1,6783 que sería fiable por ser alto el coeficiente de correlación.

2. Dada la distribución bidimensional:

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

Encuentra el valor del coeficiente de correlación lineal usando una tabla de correlación.

Solución

Se usa la siguiente tabla de doble entrada que facilita los cálculos:

X / Y	1	2	3	n'_j	$n_j'y_j$	$n'_jy_j^2$	$n_{ij}x_iy_j$
2	1			1	2	4	2
3	2	2		4	12	36	18
4		1		1	4	16	8
5		2	2	4	20	100	50
n_i	3	5	2	10	∑=38	∑=15	∑=78
n_ix_i	3	10	6	∑=19			
$n_ix_i^2$	3	20	18	∑=41			
$n_{ij}x_iy_j$	7	40	30	∑=78			

De aquí se tiene:

x = 19/10= 1,9; **y** = 38/10= 3,8; S_x^2= 4,1 - $(1,9)^2$= 0,49, S_x =0,7 ; S_y^2=15,6 - $(3,8)^2$= 1,16,

S_y = 1,077; S_{xy} = 7,8 - (1,9)(3,8)= 0,58.

Luego $\mathbf{r} = \dfrac{0,58}{(0,7)(1,077)} = 0,769$

3. En la tabla siguiente se dan los valores y algunas frecuencias absolutas de un par de variables tratadas conjuntamente. Los valores de la primera fila corresponden a la variable Y, y los de la primera columna a la variable X. La última columna es la marginal de X y la última fila es la marginal de Y.

	1	2	4	7	9	11	
1	1	2		1	0	0	5
3	0			1	1	0	4
4	1	0	2	1	1	3	
5	1	1	3	2	4	0	
6		1	1		1	0	4
7	0	0	0	1	3	1	
	4	5	8	6	10	4	

a) Completar la tabla.

b) Calcular el coeficiente de correlación y las rectas de regresión.

c) ¿Sirven las rectas de regresión para hacer predicciones de una variable en función de la otra? ¿Por qué?

Solución

x \ y	1	2	4	7	9	11	
1	1	2	**1**	1	0	0	5
3	0	**1**	**1**	1	1	0	4
4	1	0	2	1	1	3	**8**
5	1	1	3	2	4	0	**11**
6	**1**	1	1	**0**	1	0	4
7	0	0	0	1	3	1	**5**
	4	5	8	6	10	4	**37**

b) $\mathbf{x} = \dfrac{1.5 + 3.4 + 4.8 + 5.11 + 6.4 + 7.5}{37} = 4,405$; $\mathbf{y} = \dfrac{1.4 + 2.5 + 4.8 + 7.6 + 9.10 + 11.4}{37} = 6$

M_{xy} [2] $= \dfrac{\sum_{ij} x_i y_j n_{ij}}{N} = 28,378$, luego $S_{xy} = M_{xy}$ - **x.y** = 1,948

[2] $M_{xy} = \frac{1+4+4+7+6+12+21+27+4+32+28+36+132+5+10+60+180+6+12+24+54+49+189+77}{37}$

$S_x^2 = \frac{1.5+3^2.4+4^2.8+5^2.11+6^2.4+7^2.5}{37} - (4{,}405)^2 = 3{,}11$; $S_x = 1{,}764$

$S_y^2 = 47{,}027 - 36 = 11{,}027$; $S_y = 3{,}321$

El coeficiente de correlación lineal $\mathbf{r} = \frac{1{,}948}{(1{,}764)(3{,}321)}$ **= 0,3325 <0,40, correlación baja.**

$m_{yx} = 1{,}948/3{,}11 = 0{,}626$ y $m_{xy} = 1{,}948/11{,}027 = 0{,}177$ son los coeficientes de regresión.

Las rectas de regresión son:

y - 6 = 0,626 (**x** - 4,405) de Y sobre X, y x - 4,405 = 0,177(**y** - 6) de X sobre y

c) Las rectas de regresión no sirven para hacer predicciones, **fiables,** de una variable respecto de la otra ya que la correlación es baja. (*El módulo del coeficiente de correlación lineal está muy alejado de la unidad)*

Problemas propuestos

1. Las tallas y los pesos de 10 personas vienen recogidos en la siguiente tabla:

talla (cm)	160	165	170	180	185	190	192	175	182	172
pesos (kg)	58	61	65	73	80	85	83	68	74	67

Estimar el peso medio de una persona que mida 168 cm.

2. El número de licencias de caza, en miles, y el número de votantes a un determinado partido en 6 comunidades autónomas, en decenas de miles, está expresado en la siguiente tabla.:

Nº de licencias (**X**)	103	26	3	7	26	5
Nº de votantes (**Y**)	206	26	27	14	24	12

Determinar:

1) Media y varianza de las variables X e Y.

2) Coeficiente de correlación, interpretando su valor.

3) En el caso de que exista correlación: si en una determinada comunidad existen 50 decenas de millar de votantes, ¿cuántas licencias de caza, en miles, se puede estimar que existen.

3. Las distancias medias de los 19 planetas al Sol son:

1. Merc	2. Ven.	3. Tie.	4. Ma.	5. Ast.	6. Jup.	7. Sat.	8. Ur.	9. Nep.	10. Plu
0,39	0,72	1	1,52	2,65	5,2	9,54	19,19	30,07	39,52

(Se ha tomado como unidad la distancia entre la Tierra y el Sol, a lo que se llama unidad astronómica (u.a.). El quinto lugar está ocupado por los asteroides que, para estos efectos, son considerados como un planeta más.)

Representa la nube de puntos correspondiente, traza la recta de regresión y calcula el coeficiente de correlación. Si hubiera un nuevo planeta más allá de Plutón, ¿a qué distancia en u.a. estaría del Sol?. ¿Sería "fiable" esta medida?

4. Observaciones realizadas con estudiantes de Matemáticas, sobre el efecto del paso del tiempo en los conocimientos adquiridos, arrojan los siguientes resultados:

1 día ..... 90 % de permanencia de conocimientos.
2 días ..... 75 % " "
3 días ..... 42 % " "
4 días ..... 30 % " "
5 días ..... 21 % " "

Tomando los días transcurridos (X) y el tanto por ciento (Y) como variables de una distribución bidimensional, halla la recta de regresión de Y sobre X y estima, si existe una correlación fuerte, el tanto por ciento de conocimientos que permanecerán a los ocho días. Organiza los cálculos y explica el resultado.

5. Las horas de estudio y las calificaciones en Matemáticas de siete alumnos han sido:

	1º	*2º*	*3º*	*4º*	*5º*	*6º*	*7º*
H. estudio	17	17,5	13	17	17,5	15	4
Matemáticas	8	9	6	7	8	6	2

a) Halla el coeficiente de correlación entre las Matemáticas y las horas de estudio de esos alumnos.

b) Explica el significado de coeficiente de correlación.

c) Explica razonadamente como se estima la calificación en Matemáticas que obtendría un alumno al estudiar 20 horas.