UNIDAD 14.- **Distribuciones bidimensionales. Correlación y regresión**

(tema 14 del libro)

1. VARIABLES ESTADÍSTICAS BIDIMENSIONALES

Vamos a trabajar sobre una serie de fenómenos en los que para cada observación se obtiene un par de medidas.
Por ejemplo:

- Extensión en km^2 y número de habitantes de los distintos países de Europa.
- Ingresos y gastos de cada una de las familias de los trabajadores de una empresa.
- Puntuaciones obtenidas en un test de fluidez verbal por un grupo de alumnos de 4º de E.S.O. e ingresos anuales de sus padres.
- Producción y ventas de una fábrica.
- Número de años afiliados a un partido político y nivel de satisfacción con dicho partido de un militante.
- Edad y grado de psicomotricidad de un grupo formado por 40 minusválidos mentales.
- Número de horas que dedican los escolares a ver televisión y posición económica de sus padres.
- Renta nacional y número de universitarios de los distintos países de África.
- Edad y número de días que faltan al trabajo los empleados de una fábrica.

A estas variables estadísticas resultantes de la observación de un fenómeno respecto de dos modalidades se las llama **variables estadísticas bidimensionales**.

Las variables estadísticas bidimensionales las representaremos por el par (X,Y) donde X es una variable estadística unidimensional que toma los valores $x_1, x_2, x_3, ..., x_k$ e Y es otra variable estadística unidimensional que toma los valores $y_1, y_2, y_3, ..., y_k$. Por tanto, la variable estadística bidimensional (X,Y) toma estos valores:

$(x_1,y_1), (x_2,y_2), (x_3,y_3)$, ..., (x_k,y_k) o también (x_i,y_i), $1 \le i \le k$.

Las tablas de frecuencias bidimensionales pueden ser: simples o de doble entrada.

a) *Tablas de doble entrada*: Para construirla se procede así:

En la 1ª fila se colocan todos los valores que toma la variable X y en la 1ª columna se colocan todos los valores que toma la variable Y.
A continuación, en el resto de casillas se van colocando las frecuencias absolutas o relativas. Este tipo de tabla se suele utilizar cuando tenemos una gran cantidad de datos, o bien, cuando estos datos se encuentran agrupados en clases.

X \ Y	y_1	y_2	y_3	....	y_p
x_1	n_{11}	n_{12}	n_{13}	....	n_{1p}
x_2	n_{21}	n_{22}	n_{23}	....	n_{2p}
x_3	n_{31}	n_{32}	n_{33}	....	n_{3p}
....	....	....	....	....	....
x_k	n_{k1}	n_{k2}	n_{k3}	....	n_{kp}

Donde n_{ij} es la frecuencia absoluta del dato (x_i, y_j)

Ejemplo: Consideremos los siguientes datos de una distribución bidimensional.
(8,0), (8,1), (9,2), (9,4), (10,2), (10,2), (13,2), (15,3).

La tabla de doble entrada correspondiente es la siguiente:

Y/X	8	9	10	13	15	Totales
0	1	-	-	-	-	1
1	1	-	-	-	-	1
2	-	1	2	1		4
3	-	-	-	-	1	1
4	-	-	-	-	-	1
Totales	2	2	2	1	1	8

b) *Tablas simples* Para construirlas se consideran 3 columnas. En la 1ª se colocan los valores que toma la variable X, en la segunda las de Y, y en la tercera las frecuencias absolutas o las relativas.

X	x_1	x_2	x_3	....	x_k
Y	y_1	y_2	y_3	....	y_k
n_i	n_1	n_2	n_3	....	n_k

Ejemplo: Para el ejemplo anterior, tenemos:
En el ejemplo, hemos usado las frecuencias relativas en la columna 4

x_i	y_i	n_{ij}	f_{ij}
8	0	1	1/8
8	1	1	1/8
9	2	1	1/8
9	4	1	1/8
10	2	2	2/8
13	2	1	1/8
15	3	1	1/8

Si en una tabla de doble entrada sumamos las frecuencias absolutas por filas y por columnas, obtenemos una nueva fila y una nueva columna: son las **frecuencias marginales.**

Y / X	y_1	y_2	y_3	....	y_p	
x_1	n_{11}	n_{12}	n_{13}	....	n_{1p}	$n_{1\cdot}$
x_2	n_{21}	n_{22}	n_{23}	....	n_{2p}	$n_{2\cdot}$
x_3	n_{31}	n_{32}	n_{33}	....	n_{3p}	$n_{3\cdot}$
....	....	....	....	....	....	....
x_k	n_{k1}	n_{k2}	n_{k3}	....	n_{kp}	$n_{k\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	....	$n_{\cdot p}$	$\sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij} = N$

Estas frecuencias obtenidas tienen en cuenta una sola variable y se puede construir con ellas dos distribuciones unidimensionales y obtener los parámetros representativos.

DISTRIBUCIÓN MARGINAL DE X.

X	$n_{i\cdot}$
x_1	$n_{1\cdot}$
x_2	$n_{2\cdot}$
x_3	$n_{3\cdot}$
....	....
x_k	$n_{k\cdot}$
	N

DISTRIBUCIÓN MARGINAL DE Y.

Y	y_1	y_2	y_3	....	y_p	
$n_{\cdot j}$	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	....	$n_{\cdot p}$	N

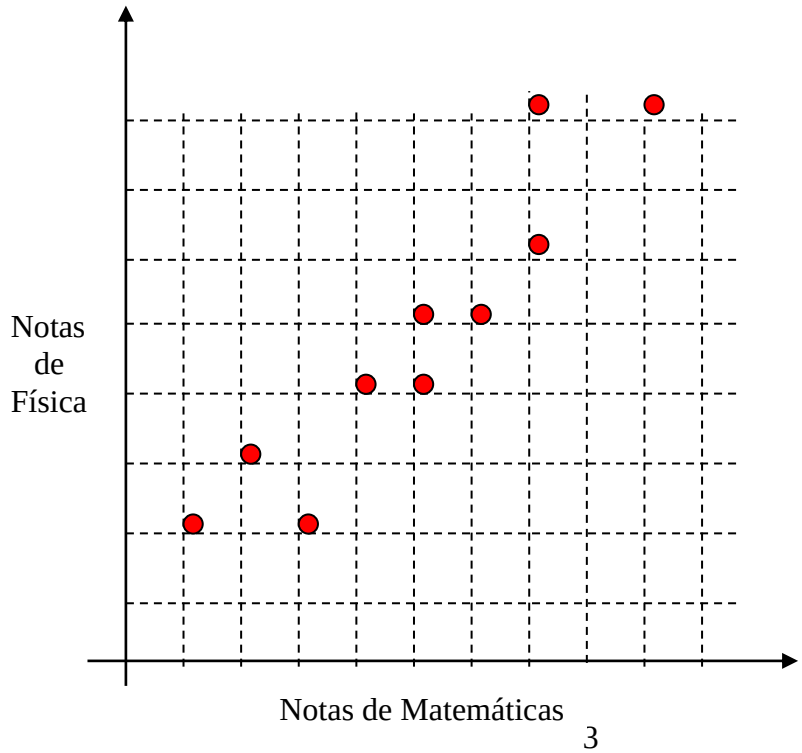
La suma de las frecuencias absolutas marginales coincide con la suma de las frecuencias bidimensionales de la tabla de doble entrada.
Las frecuencias marginales tienen en cuenta una sola variable. Cuando las variables son cuantitativas se pueden obtener parámetros representativos como media, mediana, desviación típica,...., pero cuando son variables cualitativas determinamos sólo el porcentaje.

2. DIAGRAMAS DE DISPERSIÓN O NUBE DE PUNTOS

Los valores de una variable estadística bidimensional son pares de números reales de la forma (x_i, y_j), $1 \le i \le k$, $1 \le j \le p$. Si representamos estos pares en un sistema de ejes cartesianos se obtiene un conjunto de puntos sobre el plano. A este conjunto de puntos se le denomina **diagrama de dispersión o nube de puntos**.

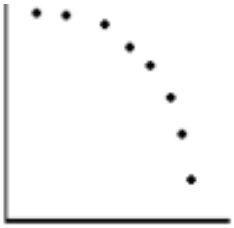
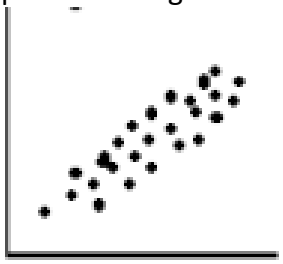
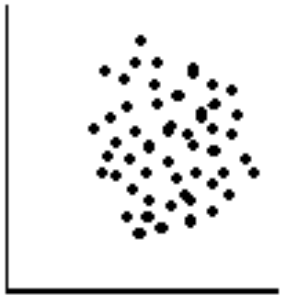
Ejemplo: Realizar la nube de puntos correspondiente a la variable estadística bidimensional que nos da la nota de Física y Matemáticas de 10 alumnos, cuyos datos son:

Matemáticas	5	6	2	9	4	5	1	3	7	7
Física	4	5	3	8	4	5	2	2	6	8

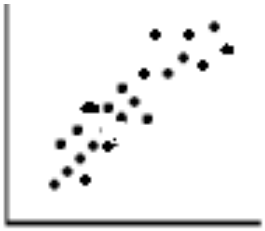
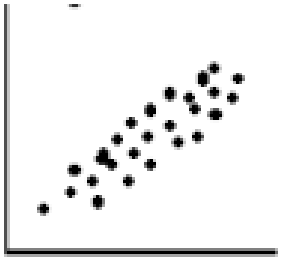
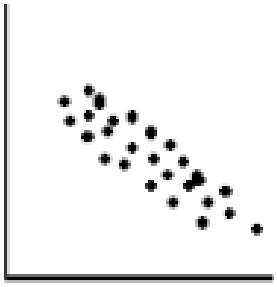


3. DEPENDENCIA O CORRELACIÓN

Según el estudio de la nube de puntos se puede apreciar, de forma cualitativa, el tipo y grado de relación o dependencia entre las dos variables de la distribución bidimensional. A esa dependencia la llamamos *correlación*.
Esta dependencia o correlación puede ser:

Dependencia funcional: Si la nube de puntos se sitúa en la gráfica de una función, excepto que esta sea constante 	*Dependencia lineal*: Si la nube de puntos se sitúa sobre una recta 
Correlación o dependencia aleatoria: Si la nube de puntos se sitúa próxima a la gráfica de una función 	*Independencia*: hay una ausencia de correlación 

A su vez la correlación se clasifica en grados, que pueden ser:

Correlación fuerte: Si la nube de puntos se aproxima mucho a una recta o curva 	*Correlación débil*: Si la nube de puntos se aproxima poco a una recta o a una curva 
Correlación positiva: Si a medida que crece una variable crece la otra 	*Correlación negativa*: Si a medida que crece una variable decrece la otra 

4. CORRELACIÓN LINEAL. COEFICIENTE DE PEARSON

Definición: La **covarianza** es una medida de la influencia del valor de una variable en el valor de la otra y se representa por σ_{xy}. Se obtiene según la fórmula:

$$\sigma_{xy} = \frac{\sum f_i x_i y_i}{N} - \overline{x}.\overline{y}$$

Propiedad: Si las variables X e Y son independientes, entonces $\sigma_{xy} = 0$

Definición: Se llama **coeficiente de correlación lineal de Pearson** a un valor que mide la correlación de tipo lineal que existe entre dos variables y se representa por *r* . Se obtiene de la siguiente fórmula:

$$r = \frac{\sigma_{xy}}{\sigma_x . \sigma_y}$$

donde:

σ_{xy} es la covarianza

σ_x es la desviación típica de la variable X

σ_y es la desviación típica de la variable Y

El coeficiente de correlación nos proporciona información acerca del tipo de correlación que existe entre las variables X e Y. Dicha información nos la proporciona dos hechos: el signo del coeficiente y su valor.

- *Signo*: A la vista de como está definido el coeficiente de correlación, al ser las desviaciones típicas siempre positivas, el signo del coeficiente viene determinado por el signo de la covarianza. Así, se tiene:
  - Si $r > 0$ → la correlación es directa
  - Si $r < 0$ → la correlación es inversa
  - Si $r = 0$ → No existe correlación entre las variables
- *Valor*: Se cumple que $-1 \leq r \leq 1$. Teniendo esto en cuenta, y dependiendo de cuál sea su valor, obtendremos cierta información:
  - Si $r = -1$ ó $r = 1$ →la correlación lineal es perfecta; esto es la nube de puntos está sobre una recta. (correlación funcional)
  - Si $-1 < r < 0$ →la correlación es lineal negativa y será más fuerte cuanto más cerca de -1 esté r.
  - Si $0 < r < 1$ →la correlación es lineal positiva y será más fuerte cuanto más cerca de 1 esté r
  - Si $r = 0$ → No existe correlación lineal. Esto no excluye que haya correlación funcional

Ejemplo: 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:

X	5	7	2	1	9
Y	15	18	10	8	20

Calcula el coeficiente de correlación lineal.

Solución:

Disponemos los cálculos de la siguiente forma:

(Accidentes) x_i	Vehículos y_i	x_i^2	y_i^2	$x_i y_i$
5	15	25	225	75
7	18	49	324	126
2	10	4	100	20
1	8	1	64	8
9	20	81	400	180
24	71	160	1113	409

$$\bar{x} = \frac{\sum x_i}{N} = \frac{24}{5} = 4{,}8\,; \quad \bar{y} = \frac{\sum y_i}{N} = \frac{71}{5} = 14{,}2\,; \quad \sigma_x^2 = \frac{\sum x_i^2}{N} - \bar{x}^2 = \frac{160}{5} - 4{,}8^2 = 8{,}96$$

$$\sigma_y^2 = \frac{\sum y_i^2}{N} - \bar{y}^2 = \frac{1113}{5} - 14{,}2^2 = 20{,}96\,; \; \sigma_{xy} = \frac{\sum x_i y_i}{N} - \bar{x}.\bar{y} = \frac{409}{5} - 4{,}8.14{,}2 = 13{,}64$$

Por tanto

$$r = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = \frac{13{,}64}{\sqrt{8{,}96}.\sqrt{20{,}96}} = 0{,}996$$

Observamos que el coeficiente de Pearson está próximo a 1 y, por tanto, existe una correlación lineal positiva fuerte entre las dos variables del problema

5. REGRESIÓN. RECTAS DE REGRESIÓN

Tenemos una distribución bidimensional y representamos la nube de puntos correspondiente. La recta que mejor se ajusta a esa nube de puntos recibe el nombre de recta de regresión. Su ecuación es la siguiente:

Recta de regresión de Y sobre X: $$y - \bar{y} = \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x})$$

Recta de regresión de X sobre Y: $$x - \bar{x} = \frac{\sigma_{xy}}{\sigma_y^2}(y - \bar{y})$$

A partir de esta recta podemos calcular los valores de X conocidos los de Y o los de Y conocidos los de X. La fiabilidad que podemos conceder a los cálculos obtenidos viene dada por el coeficiente de correlación:

- Si r es muy pequeño no tiene sentido realizar ningún tipo de estimaciones.
- Si r es próximo a – 1 ó 1, las estimaciones realizadas estarán cerca de los valores reales.
- Si r = 1 o r = -1, las estimaciones realizadas coincidirán con los valores reales.

Las rectas de regresión siempre pasan por el punto $(\bar{x}, \bar{y})$

Ejemplo: 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:

X	5	7	2	1	9
Y	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?

Solución:

Disponemos los cálculos de la siguiente forma:

(Accidentes) x_i	Vehículos y_i	x_i^2	y_i^2	$x_i y_i$
5	15	25	225	75
7	18	49	324	126
2	10	4	100	20
1	8	1	64	8
9	20	81	400	180
24	71	160	1113	409

$$\bar{x} = \frac{\sum x_i}{N} = \frac{24}{5} = 4{,}8\,; \quad \bar{y} = \frac{\sum y_i}{N} = \frac{71}{5} = 14{,}2\,; \quad \sigma_x^2 = \frac{\sum x_i^2}{N} - \bar{x}^2 = \frac{160}{5} - 4{,}8^2 = 8{,}96$$

$$\sigma_y^2 = \frac{\sum y_i^2}{N} - \bar{y}^2 = \frac{1113}{5} - 14{,}2^2 = 20{,}96\,; \; \sigma_{xy} = \frac{\sum x_i y_i}{N} - \bar{x}.\bar{y} = \frac{409}{5} - 4{,}8.14{,}2 = 13{,}64$$

a) $r = \dfrac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = \dfrac{13{,}64}{\sqrt{8{,}96}.\sqrt{20{,}96}} = 0{,}996$

b) Recta de regresión de ***Y*** sobre ***X***: $y - \bar{y} = \dfrac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x})$

$$y - 14{,}2 = \frac{13{,}64}{8{,}96}(x - 4{,}8)\,; \quad y - 14{,}2 = 1{,}53(x - 4{,}8)$$

Para x = 6, tenemos que $y - 14{,}2 = 1{,}53(6 - 4{,}8)$, es decir, y = 16,04. Podemos suponer que ayer circulaban 16 vehículos por la autopista a más de 120 km/h.

c) La predicción hecha es buena ya que el coeficiente de correlación está muy próximo a 1.

Ejemplo: Las calificaciones de 40 alumnos en psicología evolutiva y en estadística han sido las siguientes:

X calif. en psicol.	**Y** calif. en estad.	**Número** de alumnos.
3	2	4
4	5	6
5	5	12
6	6	4

6	7	5
7	6	4
7	7	2
8	9	1
10	10	2

Obtener la ecuación de la recta de regresión de calificaciones de estadística respecto de las calificaciones de psicología.

¿Cuál será la nota esperada en estadística para un alumno que obtuvo un 4,5 en psicología?

Solución:

Se pide la recta de regresión de ***y*** sobre ***x***:

$$y - \bar{y} = \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x})$$

Disponemos los datos de la siguiente forma:

x_i	y_i	n_i	f_ix_i	f_iy_i	$f_ix_i^2$	$f_iy_i^2$	$f_ix_iy_i$
3	2	4	12	8	36	16	24
4	5	6	24	30	96	150	120
5	5	12	60	60	300	300	300
6	6	4	24	24	144	144	144
6	7	5	30	35	180	245	210
7	6	4	28	24	196	144	168
7	7	2	14	14	98	98	98
8	9	1	8	9	64	81	72
10	10	2	20	20	200	200	200
		40	220	224	1314	1378	1336

$$\bar{x} = \frac{\sum f_i x_i}{N} = \frac{220}{40} = 5{,}5\,; \quad \bar{y} = \frac{\sum f_i y_i}{N} = \frac{224}{40} = 5{,}6$$

$$\sigma_{xy} = \frac{\sum f_i x_i y_i}{N} - \bar{x}.\bar{y} = \frac{1336}{40} - (5{,}3).(5{,}6) = 33{,}4 - 30{,}8 = 2{,}6$$

$$\sigma_x^2 = \frac{\sum f_i x_i^2}{N} - \bar{x}^2 = \frac{1314}{40} - (5{,}6)^2 = 32{,}85 - 30{,}25 = 2{,}6$$

Sustituyendo en la ecuación de la recta de regresión, resulta: $y - 5{,}6 = \frac{2{,}6}{2{,}6}(x - 5{,}5)$, es decir, $y = x + 0{,}1$

Si un alumno que tiene una nota de 4,5 en psicología, la nota esperada en estadística será:

y(4,5) = 4,5 + 0,1 = 4,6

La fiabilidad viene dada por el coeficiente de correlación: $r = \frac{\sigma_{xy}}{\sigma_x.\sigma_y}$

$$\sigma_{xy} = 2{,}6\,; \quad \sigma_x = \sqrt{\sigma_x^2} = \sqrt{2{,}6} = 1{,}61$$

$$\sigma_y^2 = \frac{\sum n_i y_i^2}{N} - \bar{y}^2 = \frac{1378}{40} - (5{,}6)^2 = 3{,}09\,; \quad \sigma_y = \sqrt{3{,}09} = 1{,}75$$

y resulta $r = \frac{2{,}6}{(1{,}61).(1{,}75)} = 0{,}92$

La correlación es positiva, es decir, a medida que aumenta la nota de estadística aumenta también la nota en psicología. Su valor está próximo a 1 lo que indica que se trata de una correlación fuerte, las estimaciones realizadas están cerca de los valores reales.