TEMA 2. ESTADÍSTICAS BIDIMENSIONALES

1. Definición. Objetivos

En el tema anterior hemos estudiado las distribuciones unidimensionales obtenidas al estudiar un único carácter en la población.

Cuando hacemos referencia a objetos colectivos muchas veces deseamos relacionar dos características, en principio relacionadas. Estas características (variables estadísticas) pueden estudiarse por separado, aunque muchas veces lo importante es estudiarlas de forma conjunta con el fin de obtener la relación entre ambas variables unidimensionales. Los objetivos principales de la estadística bidimensional son lo siguientes:

- Estudiar el grado de causas comunes entre ambos, lo que denominamos ***correlación***.
- Analizar una de las variables, condicionando su comportamiento a la otra. Esto es lo que denominamos ***regresión***.

Estas ideas matemáticas y sus desarrollos son debidas a los científicos ingleses Francis Galton y Kart Person (finales siglo XiX y principios del XX), en su búsqueda de relacionar la evolución y la herencia.

Las variables estadísticas bidimensionales se representan por pares de X e Y (cada una de las dos variables) sobre una población común a ambas. La notación seguida en estos apuntes es la siguiente (X,Y) y cada uno de los individuos de la población viene caracterizado por las parejas (x_i,y_j), donde x_i representa los datos (o marcas de clase) x_1, x_2,…, x_n de la variable X e y_j los datos (o marcas de clase) y_1, y_2,…,y_m de la variable Y.

Para centrarnos en el tema expongamos el siguiente ejemplo: las notas de un clase de 1º Bachillerato con 20 alumnos de Matemáticas (X) y Lengua(Y): (3,4), (4,4), (7,5), (6,6), (9,5), (8,10), (4,5), (6,7), (4,5), (4,4), (6,7), (7,7), (7,7), (5, 5), (4,3), (1,2), (2,3), (3,3), (1,1), (10,10)

Las tablas bidimensionales simples relacionan cada par (x_i, y_j) con su frecuencia absoluta f_{ij}.

Variable (X,Y)	f_{ij}
(1,1)	1
(1,2)	1
(2,3)	1
(3,3)	1
(3,4)	1
(4,3)	1
(4,4)	2
(4,5)	2
(5,5)	1
(6,6)	1
(6,7)	2
(7,5)	1
(7,7)	2
(8,10)	1
(9,5)	1
(10,10)	1
Total	20

Si bien la forma más útil de representar la información es mediante las tablas de correlación o de doble entrada, en donde una variable se expresa en las columnas y la otra en las filas, siendo las celdas intermedias las que nos expresen el valor de la frecuencia absoluta. Se suele añadir una columna y una fila de más, en las que se presentan las llamadas ***distribuciones marginales***, y que corresponden a las tablas estadísticas de correspondientes a las variables unidimensionales (las frecuencias en estas se obtienen sumando las frecuencias de todas las filas o columnas encima o la izquierda). Veamos en nuestro ejemplo:

Y / X	1	2	3	4	5	6	7	8	9	10	Marginal X
1	1	1	0	0	0	0	0	0	0	0	2
2	0	0	1	0	0	0	0	0	0	0	1
3	0	0	1	1	0	0	0	0	0	0	2
4	0	0	1	2	2	0	0	0	0	0	5
5	0	0	0	0	1	0	0	0	0	0	1
6	0	0	0	0	0	1	2	0	0	0	3
7	0	0	0	0	1	0	2	0	0	0	3
8	0	0	0	0	0	0	0	0	0	1	1
9	0	0	0	0	1	0	0	0	0	0	1
10	0	0	0	0	0	0	0	0	0	1	1
Marginal Y	1	1	3	3	5	1	4	0	0	2	N=20

La notación que utilizaremos para la frecuencia absoluta es la siguiente:

f_{ij} = "número de individuos con par (x_i,y_j)"

$f_{i\bullet}=\sum_{j=1}^{m} f_{ij} = f_{i1} + f_{i2} + ... + f_{im} =$ "número de individuos con x_i, independientemente de Y

$f_{\bullet j}=\sum_{i=1}^{n} f_{ij} = f_{1j} + f_{2j} + ... + f_{nj} =$ "número de individuos con y_j, independientemente de X

Se cumple: $\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij} = \sum_{j=1}^{m} f_{\bullet j} = \sum_{i=1}^{n} f_{i\bullet} = N$

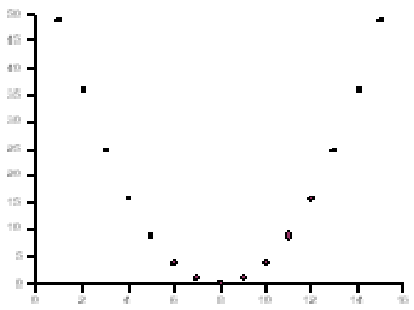
2. Coeficiente de Correlación. Lineal

En una distribución bidimensional puede ocurrir que las dos variables guarden algún tipo de relación entre si.

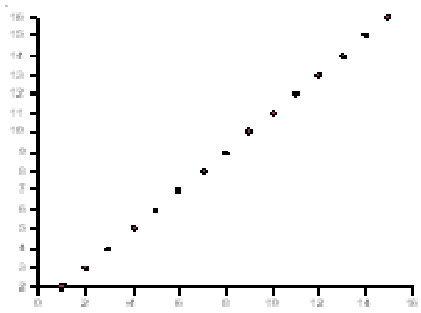
Por ejemplo, si se analiza la estatura y el peso de los alumnos de una clase es muy posible que exista relación entre ambas variables: mientras más alto sea el alumno, mayor será su peso. Lo mismo con las notas de dos asignaturas.

El coeficiente de correlación lineal mide el grado de intensidad de esta posible relación entre las variables. Este coeficiente se aplica cuando la relación que puede existir entre las variables es lineal (es decir, si representáramos en un gráfico los pares de valores de las dos variables la nube de puntos se aproximaría a una recta).

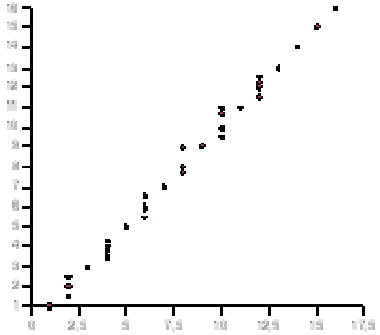
Tipos de dependencia:



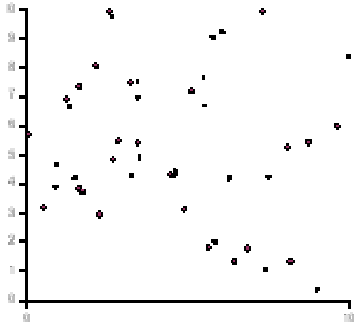
Dependencia funcional (parabólico)



Dependencia lineal

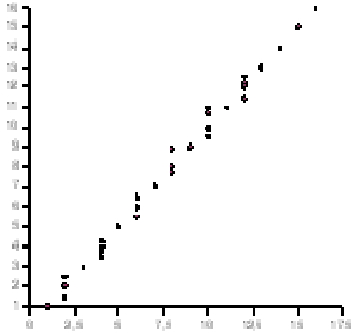


Correlación lineal

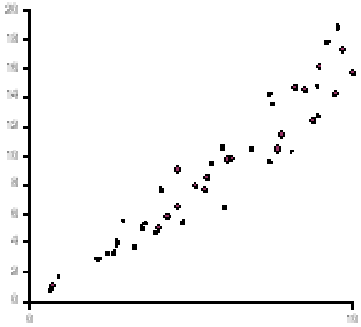


Independiente

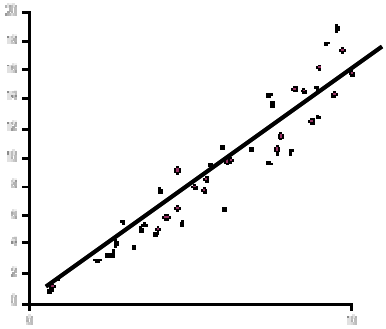
Si nos centramos en la dependencia lineal podemos distinguir entre:



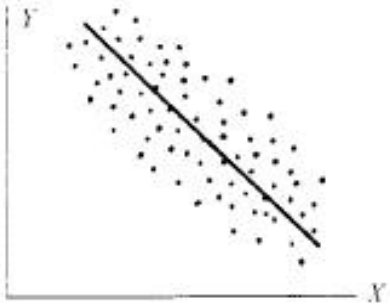
Correlación fuerte



Correlación débil



Correlación positiva



Correlación negativa

- Correlación fuerte o débil, si la nube de puntos se aproxima mucho o no a una recta. El grado de correlación se calculará a partir del coeficiente de correlación.
- Correlación positiva, cuando la recta es creciente (a mayor valor de la variable X más valor de Y) y negativa cuando la recta es decreciente (a mayor valor de la variable X menor valor de Y).

No obstante, puede que exista una relación que no sea lineal, sino exponencial, parabólica, etc. En estos casos, el coeficiente de correlación lineal mediría mal la intensidad de la relación las variables, por lo que convendría utilizar otro tipo de coeficiente más apropiado.

El nivel de correlación de tipo lineal se mide a partir del ***coeficiente de correlación lineal de Pearson***, cuyo valor se calcula como:

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \text{ donde:}$$

$$\sigma_x = \sqrt{\left(\overline{x^2}\right) - \left(\overline{x}\right)^2}\ , \quad \sigma_y = \sqrt{\left(\overline{y^2}\right) - \left(\overline{y}\right)^2}$$

$$\sigma_{xy} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}(x_i - \overline{x})\cdot(y_j - \overline{y})}{N} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij} x_i \cdot y_j}{N} - \overline{x}\cdot\overline{y}$$

Siendo:

- σ_x y σ_y las desviaciones típicas de las variables marginales X e Y, y que en el tema anterior vimos como se calculaban.
- σ_{xy} la ***covarianza*** conjunta de X e Y.

Para calcular la covarianza se puede añadir una columna a mayores en las tablas bidimensionales:

Variable (X,Y)	f_{ij}	$f_{ij}\cdot x_i \cdot y_j$
(1,1)	1	1
(1,2)	1	2
(2,3)	1	6
(3,3)	1	9
(3,4)	1	12
(4,3)	1	12
(4,4)	2	32
(4,5)	2	40
(5,5)	1	25
(6,6)	1	36
(6,7)	2	84
(7,5)	1	35
(7,7)	2	98
(8,10)	1	80
(9,5)	1	45
(10,10)	1	100
Total	20	617

Interpretación del coeficiente de correlación:

- Correlación positiva o negativa:
  - Si r>0 la correlación es positiva.
  - Si r<1 la correlación es negativa.
- Correlación fuerte o débil:
  - Si |r| próximo a 1 (|r|≈1) correlación fuerte.
  - Si |r| próximo a 0 (|r|≈0) correlación débil.

En nuestro ejemplo:

$$\bar{x} = 5{,}05 \quad \sigma_x = 2{,}4$$

$$\bar{y} = 5{,}15 \quad \sigma_y = 2{,}3 \qquad\qquad r = \frac{4{,}8}{2{,}3 \cdot 2{,}4} = 0{,}86$$

$$\sigma_{xy} = \frac{617}{20} - 5{,}05 \cdot 5{,}15 = 4{,}8$$

El resultado era esperado, r>0 y cerca de 1, es decir correlación lineal fuerte y positiva.

3. Rectas de regresión.

En numerosas situaciones el diagrama de dispersión, o nube de puntos, sugiere la búsqueda de una curva o recta que nos relacione las dos variables de forma funcional. Esta curva se llama línea de regresión.

En este tema sólo nos vamos a centrar en aquellas que podemos ajustar a una recta, estas rectas se denominan ***rectas de regresión***. Podemos calcular dos rectas de regresión según la información que deseemos:

- La recta de regresión de Y sobre X: Es la recta que más se ajusta a la nube de puntos de forma que cumple: y=f(x)=mx+n
  - ✓ r pasa por el "centro de gravedad" de la nube de puntos $(\bar{x}, \bar{y})$
  - ✓ Es la recta que menor es la suma de las diferencias de los cuadrados entre el valor de la recta en x_i ($f(x_i)$), con $y_i \rightarrow \min \sum_i (y_i - f(x_i))^2$
  - ✓ Se utiliza cuando queremos conocer el valor de la variable Y conocida la X

- La recta de regresión de X sobre Y: Es la recta que más se ajusta a la nube de puntos de forma que cumple: x=f(y)=m'y+n'
  - ✓ r pasa por el "centro de gravedad" de la nube de puntos $(\bar{x}, \bar{y})$
  - ✓ Es la recta que menor es la suma de las diferencias de los cuadrados entre el valor de la recta en y_i ($f(y_i)$), con $x_i \rightarrow \min \sum_i (x_i - f(y_i))^2$
  - ✓ Se utiliza cuando queremos conocer el valor de la variable X conocida la Y

Calculo de las rectas de regresión:

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

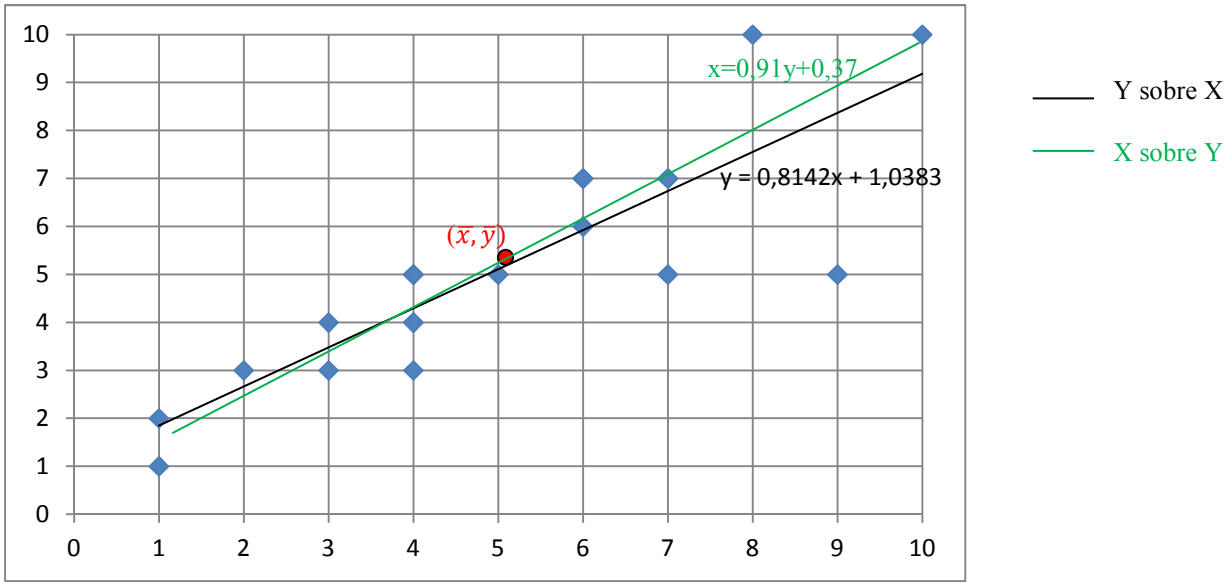
b) La recta de regresión de X sobre Y: $x = \bar{x} + \frac{\sigma_{xy}}{\sigma_y^2}(y - \bar{y})$

Calculemos las rectas de regresión para nuestro ejemplo:

a) La recta de regresión de Y sobre X: $y = 5{,}15 + \frac{4{,}8}{2{,}4^2}(x - 5{,}05) = 0{,}814x + 1{,}04$

b) La recta de regresión de X sobre Y: $x = 5{,}05 + \frac{4{,}8}{2{,}3^2}(y - 5{,}15) = 0{,}91y + 0{,}37$

Veamos la grafica del ejemplo:



Las rectas nos permiten hacer estimaciones o previsiones de lo que puede a una de las dos variables conocida la otra, pero deben de tenerse en cuenta las siguientes consideraciones:

- Las estimaciones serán fiables siempre que el coeficiente de correlación lineal, r, sea próximo a 1 o a -1.
- Las estimaciones además tienen especialmente sentido cuando queremos conocer el valor para valores próximos a los datos.

Ejercicios finales

Ejercicio 1. Una empresa de venta y elaboración de ropa para jóvenes ha realizado el siguiente gasto en publicidad y obtenido los siguientes resultados en ventas en los últimos 10 años (los datos vienen expresados en miles de euros)

Gasto Publicidad	7,5	8	8,5	10	10,5	12	13	14	15	18
Beneficios ventas	200	205	230	240	250	270	280	300	310	325

Si denotamos X al gasto en publicidad e Y a los beneficios en las ventas, halla:

a) El coeficiente de correlación lineal. Analiza las dependencias de ambas variables
b) La recta de regresión que nos relaciona el gasto en publicidad con el beneficio.
c) La empresa decide invertir el próximo año 25.000€ en publicidad. Suponiendo que se mantiene la misma tendencia que años atrás, ¿Cuál será el volumen en venta esperado?
d) Si la empresa desea lograr 350.000€ en beneficios, ¿Cuánto estimas que debe invertir en publicidad?
e) ¿Sería lógico decir que podemos estimar de forma aproximada los beneficios con un gasto en publicidad de 200.000€? ¿Por qué?

Ejercicio 2. La siguiente tabla muestra las notas que 5 amigos de primer curso de bachillerato han obtenido en la asignatura de Matemáticas en la primera y segunda evaluación:

1ª Evaluación (X)	5	6,5	8	4	3
2ª Evaluación (Y)	4,5	7	7,5	5	3,5

a) Calcular el coeficiente de correlación e interpreta el valor obtenido.
b) Determina las rectas de regresión de Y sobre X y de X sobre Y.
c) Halla el punto donde se cortan las rectas de regresión y comprueba que es "el centro de gravedad"
d) Si otro amigo saco un 9 en la primera evaluación ¿qué estimación le daría la recta de regresión calculada de la nota de la segunda evaluación?

Ejercicio 3. En la tabla siguiente se presenta una relación entre las horas de entrenamiento de un atleta a la semana y el tiempo que tarda en recorrer los 400m.

Tiempo entrenamiento (X)	10	12	15	17	20
Tiempo 400m (Y)	48	46	45	44,2	43,8

a) Representa la nube de puntos y estima sin calcularlo el coeficiente de correlación. ¿Qué signo tiene?
b) Calcular el coeficiente de correlación y la recta de regresión X frente Y.
c) Si entrenara 40 horas estima el valor del tiempo obtenido. El record del mundo es de 43,18, ¿es posible que alguien lo baje por tanto duplicando el entrenamiento? ¿Qué ha fallado en nuestra estimación?

Ejercicio 4. De dos variables X e Y se sabe que la desviación típica de X es $\sqrt{3}$, la media y la desviación típica de Y es 1 y 2, respectivamente, y la ecuación de la recta de regresión de Y sobre X es 2x+3y=6. Hallar

a) La media de X
b) La covarianza de X e Y
c) El coeficiente de correlación
d) La recta de regresión de X frente a Y .

Ejercicio 5. Las rectas de regresión para una muestra de las variables X e Y son:

y= 0,52x+21,7 ; x=0,85y+40,97

Determinar las medias muestrales de X e Y y el coeficiente de correlación lineal de Pearson