DISTRIBUCIÓN BIDIMENSIONAL

En este tema se estudian fenómenos bidimensionales de carácter aleatorio. El objetivo es doble:

1. Determinar si existe relación entre las variables consideradas(Correlación).
2. Si esa relación existe, indicar el procedimiento para estimar el valor de una variable a partir de otra(Regresión).

Distribuciones bidimensionales

Una distribución de dos variables (bidimensional) es un conjunto de parejas de valores (x_i, y_i), que pueden presentarse mediante una tabla.

f_i(frecuencia)	f_1	f_2	...	f_i
x_i	x_1	x_2	...	x_i
y_i	y_1	y_2	...	y_i

Genéricamente, las variables se llaman x(variable independiente) e y(variable dependiente).

Correlación

Al estudiar distribuciones bidimensionales, el objetivo perseguido es determinar si existe relación estadística entre las dos variables consideradas; es decir, ver si los cambios en una de las variables influyen en los cambios de la otra. Cuando sucede esto, se dice que ambas variables están correlacionadas o que hay **correlación** entre ellas.

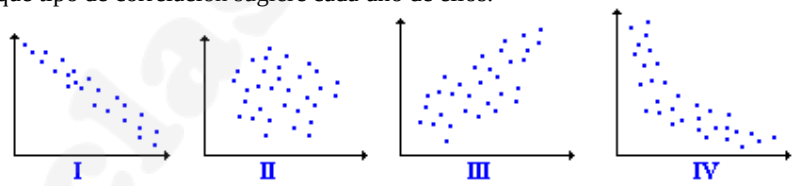
Si las variables crecen conjuntamente, la correlación es **directa**. Si, por el contrario, al aumentar una de ellas disminuye la otra, la correlación será **inversa.**

La correlación puede calificarse como fuerte cuando el grado de dependencia es alto; y como débil en caso contrario.

Diagramas de dispersión

El primer paso para determinar el sentido y el grado de la correlación entre dos variables consiste en representar gráficamente, en el plano cartesiano, los pares de valores conocidos. Estos gráficos, que reciben el nombre de **diagramas de dispersión**, permiten visualizar la posición de los datos en el plano. La forma de la **nube de puntos** asociada a cada diagrama permitirá establecer conjeturas sobre la correlación existente entre las variables estudiadas.

En la siguiente figura se dan algunos diagramas de dispersión. Por la forma de la nube de puntos, se puede intuir que tipo de correlación sugiere cada uno de ellos.



I. Esta nube, estrecha y decreciente, indica correlación lineal inversa y fuerte.
II. En este caso, la nube no adopta una forma definida: no hay correlación (o es muy débil).
III. Esta nube, ancha y con tendencia a crecer, sugiere una correlación lineal directa y débil.
IV. La nube presenta una forma clara, pero no rectilínea. La correlación no es lineal, podría ser exponencial o parabólica.

En general, dependiendo de la forma de la nube de puntos, puede asegurarse:

- Una nube de puntos alargada indica correlación lineal: los puntos se distribuyen entorno a una línea recta. La estrechez de la nube expresa que la correlación es fuerte.
- Si la recta que se ajusta a la nube tiene pendiente positiva, la **correlación será directa:** al aumentar x, aumenta y, ó viceversa
- Una recta con pendiente negativa indica que la **correlación es inversa,** al aumentar x, disminuye y, ó viceversa.

En la siguiente figura se puede observar todo lo anterior

El estudio cuantitativo de estos conceptos se realiza mediante los parámetros de correlación y de regresión

Parámetros de una distribución bidimensional

Los datos de una distribución bidimensional suele darse en forma de tabla. Por ejemplo:

x_i	x_1	x_2	...	x_n
y_i	y_1	y_2	...	y_n

en el caso que las frecuencias de cada pareja sean uno, en otro caso:

x_i	x_1	x_2	...	x_n
y_i	y_1	y_2	...	y_n
f_i	f_1	f_2	...	f_n

También se pueden presentar en cuadros de doble entrada

x / y	x_1	x_2	... ...	x_n
y_1	$f_{1.1}$	$f_{1.2}$	... ...	$f_{1.n}$
y_2	$f_{2.1}$	$f_{2.2}$	... ...	$f_{2.n}$
...	... ...	... ...	... ...	... ...
...	... ...	... ...	... ...	... ...
y_m	$f_{m.1}$	$f_{m.2}$	... ...	$f_{m.n}$

Lo datos correspondientes a cada una de las variables se llaman **datos marginales**. (En el caso de tablas de doble entrada puede hablarse de **frecuencias marginales**). Estos datos permiten el cálculo de los **parámetros marginales** de cada una de las variables.

Medias

Las medias marginales para cada una de las variables X e Y valen, respectivamente:

$$\bar{x} = \frac{\sum x_i \cdot f_i}{n} \qquad \bar{y} = \frac{\sum y_i \cdot f_i}{n}$$

El punto $(\bar{x}, \bar{y})$ se llama centro medio de la distribución. Es el centro de gravedad (o centro de masas) de la nube de puntos. Si se considera las medidas ponderadas se llamaría centro medio ponderado.

Varianzas marginales

Las varianzas marginales, son:

Para x: $$s_x^2 = \sigma_x^2 = \frac{\sum (x_i - \bar{x})^2 \cdot f_i}{n} = \frac{\sum x_i^2 \cdot f_i}{n} - \bar{x}^2$$

Para y: $$s_y^2 = \sigma_y^2 = \frac{\sum (y_i - \bar{y})^2 \cdot f_i}{n} = \frac{\sum y_i^2 \cdot f_i}{n} - \bar{y}^2$$

Desviaciones típicas marginales

Para x: $$s_x = \sigma_x = \sqrt{\frac{\sum (x_i - \bar{x})^2 \cdot f_i}{n}} = \sqrt{\frac{\sum x_i^2 \cdot f_i}{n} - \bar{x}^2}$$

Para y: $$s_y = \sigma_y = \sqrt{\frac{\sum (y_i - \bar{y})^2 \cdot f_i}{n}} = \sqrt{\frac{\sum y_i^2 \cdot f_i}{n} - \bar{y}^2}$$

Covarianza

La covarianza es un parámetro estadístico conjunto ya que en su cálculo intervienen las dos variables a la vez. Se define como la media aritmética de los productos de las diferencias de cada variable respecto de su media marginal. Por tanto, vale:

$$s_{xy} = \sigma_{xy} = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y}) \cdot f_i}{n} \quad \text{ó} \quad s_{xy} = \sigma_{xy} = \frac{\sum x_i \cdot y_i \cdot f_i}{n} - \bar{x} \cdot \bar{y}$$

La covarianza permite estimar conceptos relativos a la correlación entre las dos variables

I. Su signo indica el sentido de la correlación entre las variables.
   - Si $s_{xy} > 0$, la correlación es directa.
   - Si $s_{xy} < 0$, la correlación es inversa.

II. Un valor grande de s_{xy} advierte que la correlación entre las variables puede ser fuerte, pero no lo asegura, no siendo interesante la comparación de dos distribuciones por la covarianza.

La covarianza sólo da el sentido de la correlación: directa si es positiva e inversa si es negativo.

Coeficiente de correlación lineal

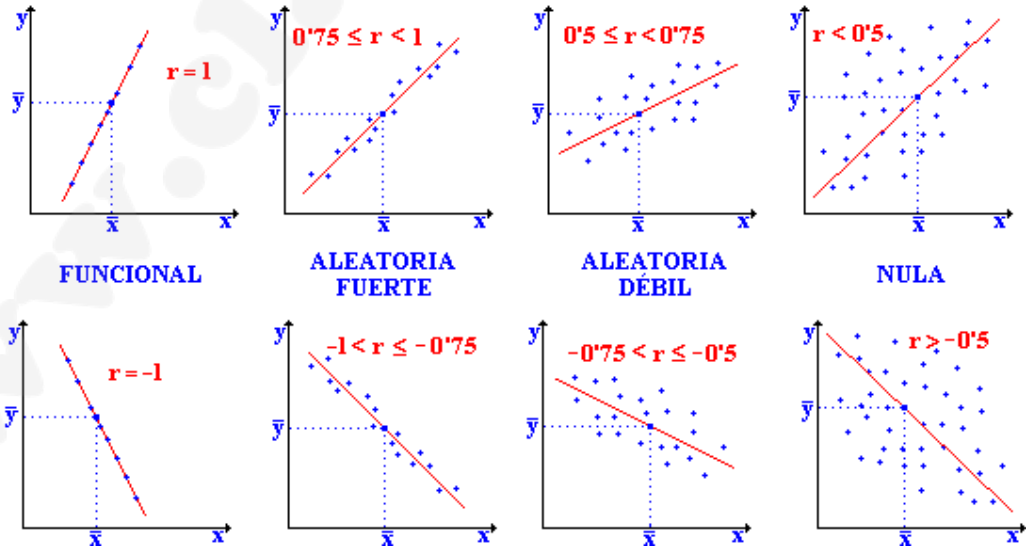
El coeficiente de correlación lineal(r) es el criterio que se utiliza para medir la fuerza de la correlación lineal entre dos variables, se define como:

$$r = \frac{s_{xy}}{s_x \cdot s_y}$$

Es la razón entre la covarianza de las variables x e y, y el producto de sus desviaciones típicas marginales. Sus propiedades fundamentales son:
Las propiedades fundamentales del coeficiente de correlación son:

I. El valor de r no es función de la escala de medida.
II. El signo de r es el mismo que el de la covarianza, pues las desviaciones siempre son positivas. Luego:
   - Si r > 0, la correlación es directa;
   - Si r < 0, la correlación es inversa.
III. El valor de r está comprendido entre –1 y +1: $-1 \le r \le 1$
IV. Si r toma valores próximos a –1, la correlación es fuerte e inversa.
V. Si r toma valores próximos a +1, la correlación es fuerte y directa.
VI. Si $|r| = 1$, la correlación es perfecta denominándose correlación funcional. Hay dependencia lineal entre las variables X e Y.
VII. Si r toma valores cercanos a 0, la correlación prácticamente no existe.

En función del valor numérico del coeficiente de correlación lineal, se puede clasificar la correlación en diferentes tipos:



El coeficiente de correlación (r) mide exclusivamente la correlación lineal entre dos variables, no siendo capaz de detectar correlaciones de otro tipo(Exponencial, Cuadrática, ... etc).

A **r^2** se le denomina **coeficiente de determinación**, y da una medida de la fiabilidad de las estimaciones de Y a partir de X. El valor del coeficiente de determinación indica la proporción de la variación en la variable Y que puede ser explicada en la variable X

Recta de regresión

La recta de regresión es la que mejor se ajusta a la nube de puntos, haciendo mínima la suma de las distancias de todos los puntos de la nube a ella. Debe pasar por el punto $(\bar{x}, \bar{y})$, centro de gravedad de la distribución bidimensional.

La recta que mejor se ajusta a estos propósitos es la recta de regresión mínimo cuadrática. Con estas condiciones, los valores de la pendiente a y de la ordenada en el origen b de esa recta valen:

$$a = \frac{s_{xy}}{s_x^2} \qquad b = \bar{y} - \frac{s_{xy}}{s_x^2} \cdot \bar{x}$$

Luego, la ecuación de la recta de regresión es:

$$y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x})$$

Siendo $\bar{x}$ e $\bar{y}$ las medidas de las variables X e Y, s_x^2 la varianza de X y s_{xy} la covarianza.

Esta recta de regresión se llama de **Y sobre X**, pues se utiliza para predecir (estimar) los valores de Y a partir de los de X. Si lo que se desea es estimar los valores de X partiendo de los de Y, se empleará la ecuación de la **recta de regresión de X sobre Y**, que es:

$$x - \bar{x} = \frac{s_{xy}}{s_y^2}(y - \bar{y})$$

A $\frac{S_{xy}}{S_y^2}$ se le denomina **coeficiente de regresión de X sobre Y**. No es la pendiente de la recta, sino su inversa.

Las rectas de regresión de Y sobre X y de X sobre Y se cortan en el centro de gravedad de la distribución $(\bar{x}, \bar{y})$. Su posición relativa es función del coeficiente de correlación, oscilando desde perpendiculares cuando r = 0, hasta coincidentes cuando r = 1.

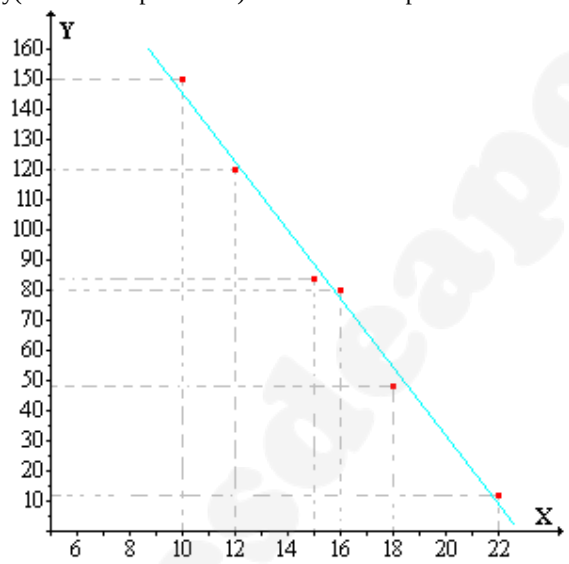
Ejemplo 1. La temperatura media anual, en ºC, de varias ciudades, y el gasto medio anual en calefacción por habitante en € fue:

Temperatura ºC	10	12	15	16	18	22
Gasto €	150	120	84	60	48	12

a. Representar la nube de puntos asociada. ¿Qué tipo de correlación se observa?
b. Hallar el coeficiente de correlación y la recta de regresión del gasto sobre la temperatura.
c. Interpretar el coeficiente de determinación
d. Que gasto cabe esperar en ciudades con temperatura media de 8, 17, 26 ºC.
e. Que temperatura media hubo en una ciudad cuyo gasto media por habitante fue de 98 €.
f. Representar las dos rectas de regresión.

a.

x(Variable independiente) ≡ Temperatura media en ºC
y(Variable dependiente) ≡ Gasto medio por habitante en €



Se puede observar que los puntos se ajustan bien a una recta de pendiente negativa, por lo tanto entre las dos variables cabe esperar una correlación aleatoria fuerte inversamente proporcional.

b. Para el cálculo de los parámetros de la distribución es necesario el siguiente cuadro de frecuencias:

x_i	y_i	x_i^2	y_i^2	$x_i \cdot y_i$
10	150	100	22500	1500
12	120	144	14400	1440
15	84	225	7056	1260
16	60	256	3600	960
18	48	324	2304	864
22	12	484	144	264
$\sum x_1 = 93$	$\sum y_i = 474$	$\sum x_i^2 = 1533$	$\sum y_i^2 = 50004$	$\sum x_i \cdot y_i = 6288$

Parámetros de la distribución:

- Medias: $\begin{cases} \bar{x} = \dfrac{\sum x_i}{N} = \dfrac{93}{6} = 15'5 \\ \bar{y} = \dfrac{\sum y_i}{N} = \dfrac{474}{6} = 79 \end{cases}$

- Varianzas: $\begin{cases} s_x^2 = \dfrac{\sum x_i^2}{N} - \bar{x}^2 = \dfrac{1533}{6} - 15'5^2 = 15'25 \\ s_y^2 = \dfrac{\sum y_i^2}{N} - \bar{y}^2 = \dfrac{50004}{6} - 79^2 = 2093 \end{cases}$

- Desviaciones: $\begin{cases} s_x = \sqrt{\dfrac{\sum x_i^2}{N} - \bar{x}^2} = \sqrt{\dfrac{1533}{6} - 15'5^2} = 3'91 \\ s_y = \sqrt{\dfrac{\sum y_i^2}{N} - \bar{y}^2} = \sqrt{\dfrac{50004}{6} - 79^2} = 45'75 \end{cases}$

- Covarianza: $S_{xy} = \dfrac{\sum x_i \cdot y_i}{N} - \bar{x} \cdot \bar{y} = \dfrac{6288}{6} - 15'5 \cdot 79 = -176'5$

- Coeficiente de correlación: $r = \dfrac{S_{xy}}{s_x \cdot s_y} = \dfrac{-176'5}{3'91 \cdot 45'75} = 0'988$

- Coeficiente de determinación: $r^2(\%) = 0'988^2 \cdot 100 = 97'6$

- Recta de regresión de Y sobre X: $y - \bar{y} = \dfrac{s_{xy}}{s_x^2}(x - \bar{x})$

$$y - 79 = \frac{-176'5}{15'25}(x - 15'5)$$

ordenando

$$y = 258'3 - 11'6x$$

c. En las variaciones producidas en el gasto medio por habitante, el 97'6% es función de las variaciones en la temperatura media, el 2'8% restante se debe a otros conceptos.

d. Para estimar el gasto medio conocida la temperatura media se usa la recta de regresión

$$\left.\begin{array}{l} x_1 = 8^o\,C \\ x_2 = 17^o\,C \\ x_3 = 26^o\,C \end{array}\right\} y = 258'3 - 11'6x : \begin{cases} y_1 = 258'3 - 11'6 \cdot 8 = 165'5\,€ \\ y_2 = 258'3 - 11'6 \cdot 17 = 61'1\,€ \\ y_3 = 258'3 - 11'6 \cdot 26 = -43'3\,€ \end{cases}$$

el valor y_3, no tiene sentido, pero tampoco es lógico usar calefacción para una temperatura media de 26 ºC

e. Conocido el gasto medio en calefacción también se puede estimar le temperatura media, mediante la recta de regresión.

$$y_4 = 98\,€ \xrightarrow{y = 258'3 - 11'6x} 98 = 258'3 - 11x \Rightarrow x = \frac{98 - 258'3}{-11'6} = 13'8\,^o\,C$$

d. X sobre Y: $x - \bar{x} = \dfrac{s_{xy}}{s_y^2}(y - \bar{y})$, $x - 15'5 = \dfrac{-176'5}{2093}(y - 79)$, ordenando: y = 262'8– 11'9x