

DISTRIBUCIONES BIDIMENSIONALES

RESULTAN DE ESTUDIAR FENÓMENOS EN LOS QUE PARA CADA OBSERVACIÓN SE OBTIENE UN PAR DE MEDIDAS Y, EN CONSECUENCIA, DOS VARIABLES.

Ejemplos.

- Talla y peso de los soldados de un regimiento.
- Calificaciones en Física y Matemáticas de los alumnos de una clase.
- Gastos de publicidad y ventas de una fábrica.
- Etc.

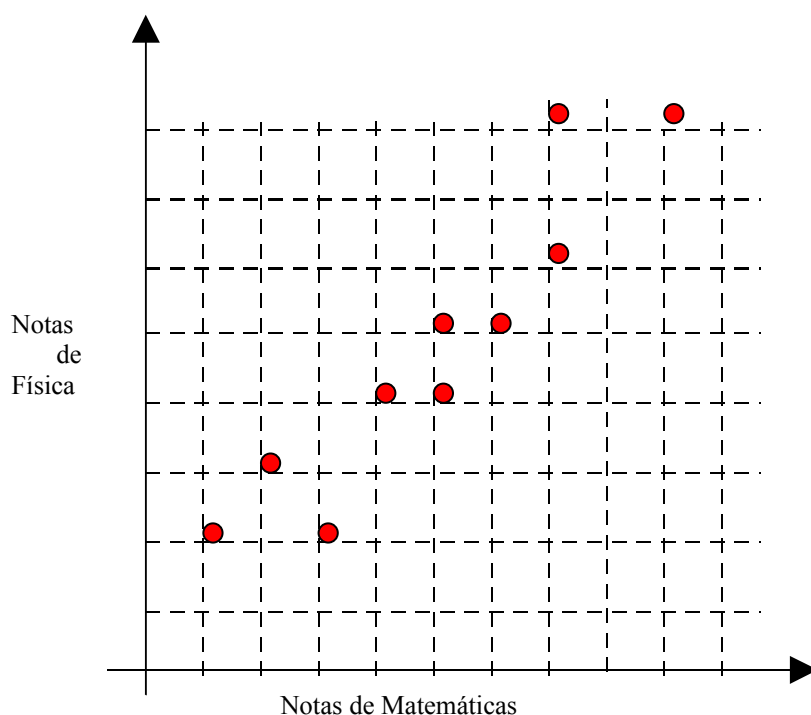
Estas variables resultantes de la observación de un fenómeno respecto de dos modalidades se llaman variables estadísticas bidimensionales.

Los valores de una variable estadística bidimensional son pares de números reales de la forma (x_i, y_i) . Representados en un sistema de ejes cartesianos se obtiene un conjunto de puntos llamado diagrama de dispersión o nube de puntos.

Ejemplo: Nube de puntos de la distribución dada por la tabla siguiente:

Notas de Matemáticas y Física de 10 alumnos

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



Parámetros estadísticos.

Media de la variable X: $\bar{x} = \frac{\sum f_i x_i}{N}$

Media de la variable Y: $\bar{y} = \frac{\sum f_i y_i}{N}$

Varianza de la variable X: $s_x^2 = \frac{\sum f_i x_i^2}{N} - \bar{x}^2$

Varianza de la variable Y: $s_y^2 = \frac{\sum f_i y_i^2}{N} - \bar{y}^2$

Covarianza: $s_{xy} = \frac{\sum f_i x_i y_i}{N} - \bar{x} \cdot \bar{y}$

Correlación.

Estudia la relación o dependencia que existe entre dos variables que intervienen en una distribución bidimensional.

Coefficiente de correlación lineal.

Es un número que mide el grado de dependencia entre las variables X e Y.

Se mide mediante la siguiente fórmula: $r = \frac{s_{xy}}{s_x \cdot s_y}$

Su valor está comprendido entre -1 y 1.

- Si $r = -1$ ó $r = 1$ todos los valores de la variable bidimensional se encuentran situados sobre una recta.
- Si $-1 < r < 0$ se dice que las variables X e Y están también en dependencia aleatoria. La correlación es negativa.
- Si $0 < r < 1$ la correlación es positiva. Las variables X e Y están también en dependencia aleatoria.

La correlación es tanto más fuerte a medida que r se aproxima a -1 ó 1 y es tanto más débil a medida que se aproxima a 0.

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

Recta de regresión de x sobre y : $x - \bar{x} = \frac{s_{xy}}{s_y^2} (y - \bar{y})$

A partir de esta recta podemos calcular los valores de x conocidos los de y . 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.

Ejercicios resueltos.

1.- Una compañía de seguros considera que el número de vehículos (Y) que circulan por una determinada autopista a más de 120 kms/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 kms/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

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

$$s_y^2 = \frac{\sum y_i^2}{N} - \bar{y}^2 = \frac{1113}{5} - 14,2^2 = 20,96; \quad s_{xy} = \frac{\sum x_i y_i}{N} - \bar{x} \cdot \bar{y} = \frac{409}{5} - 4,8 \cdot 14,2 = 13,64$$

$$a) \quad r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{13,64}{\sqrt{8,96} \cdot \sqrt{20,96}} = 0,996$$

$$b) \text{ Recta de regresión de } y \text{ sobre } x: \quad y - \bar{y} = \frac{s_{xy}}{s_x^2} (x - \bar{x})$$

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

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

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

2.- 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 :

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

Disponemos los datos de la siguiente forma:

x_i	y_i	f_i	$f_i x_i$	$f_i y_i$	$f_i x_i^2$	$f_i y_i^2$	$f_i x_i y_i$
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

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

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

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

Sustituyendo en la ecuación de la recta de regresión, resulta:

$$y - 5,6 = \frac{2,6}{2,6}(x - 5,5), \text{ es decir, } y = x + 0,1$$

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

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

Se sustituye en la recta de regresión.

La fiabilidad viene dada por el coeficiente de correlación: $r = \frac{s_{xy}}{s_x \cdot s_y}$

$$s_{xy} = 2,6; \quad s_x = \sqrt{s_x^2} = \sqrt{2,6} = 1,61$$

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

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

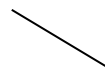
La correlación es positiva, es decir, a medida que aumenta la nota de estadística aumenta también la nota en psicología. Su valor está próximo a 1 lo que indica que se trata de una correlación fuerte, las estimaciones realizadas están cerca de los valores reales.

Tablas de doble entrada.

En las distribuciones bidimensionales, cuando hay pocos pares de valores, se procede como hemos hecho, es decir, enumerándolos. Si algún par está repetido se pone dos veces, pero cuando el número de datos es grande, se recurre a las tablas de doble entrada.

En cada casilla se pone la frecuencia correspondiente al par de valores que definen esa casilla.

Ejemplo:



x	0	1	2
y			
0	2	1	0
1	3	4	1
2	0	5	3

Lo que indica el número de veces que está cada par. El par (0, 1) está 3 veces.

El par (1, 2) está 5 veces. Etc.