MATEMATICAS (MAS I)

1º Bachillerato

ESTADÍSTICA BIDIMENSIONAL

Departamento de Matemáticas

Ies Dionisio Aguado

En este capítulo veremos el estudio de un fenómeno respecto a dos variables unidimensionales simultáneamente, se obtiene así el concepto de variable estadística bidimensional.

Una distribución bidimensional o de dos variables es aquella en la que para cada elemento de la población o muestra se consideran dos caracteres cuantitativos distintos, (X,Y). Así, a cada individuo le corresponden los valores de dos variables que representamos mediante el par ordenado (x_i, y_i);, siendo x_i; el valor del primer carácter e y_i; el del segundo.

Parece lógico pensar que las siguientes parejas de variables deben guardar alguna relación entre sí:

- Los pesos (X) y las estaturas (Y) de un conjunto de personas.
- El número de encuentros ganados por un equipo de fútbol (X) y el lugar que ocupa en la clasificación (Y).
- Las notas obtenidas por cada alumno de una clase en dos asignaturas de similares características.(X)=nota1 e $(Y) = nota2$
- Las velocidades a las que circulan un conjunto de vehículos (X) y su consumo de combustible(Y).
- Ingresos y gastos de cada una de las familias de los trabajadores de una empresa.
- Edad y número de días de absentismo de los empleados de una fábrica.
- Número de horas que dedican los estudiantes jugar on-line y resultados académicos.

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 tomalos valores $x_1, x_2, x_3, , ..., x_k$

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:

$(x_1, y_1), (x_2, y_2), (x_3, y_3), ..., (x_k, x_k)$

Si representamos esto valores en un plano coordenado X,Y obtenemos el llamado diagrama de dispersión o la nube de puntos.

Ejemplo

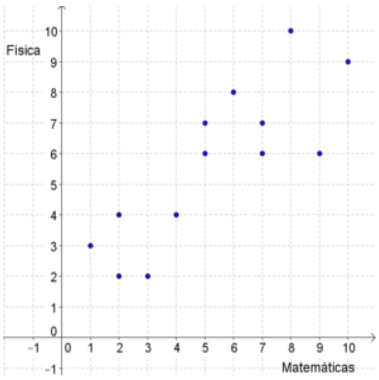
Notas obtenidas en las asignaturas de Matemáticas y Física por los alumnos de una clase.

(X,Y)=(Matemáticas,Física)={(3,2),(2,2),(5,6),(1,3),(7,6),(6,8),
(2,4),(4,4),(8,10),(9,6),(5,7),(10,9),(7,7)}

En el anterior ejemplo sobre las notas de Matemáticas y Física, teníamos la siguiente tabla de observaciones:

Matemáticas (X)	3	2	5	1	7	6	2	4	8	9	5	10	7
Física (Y)	2	2	6	3	6	8	4	4	10	6	7	9	7

Su correspondiente nube de puntos será:



A partir de la nube de puntos podemos observar como se relacionan las dos variables. Si al aumentar o disminuir una de ellas, aumenta o disminuye la otra de forma sistemática, diremos que existe correlación entre las dos variables.

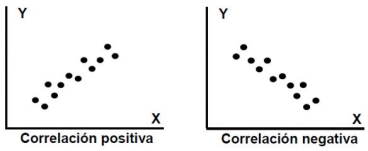
La correlación puede ser:

- Fuerte: Si hay mayor alineamiento de los puntos de la nube.
- Débil: Si el alineamiento es menor, hay mas "dispersión" de los puntos.

La correlación (fuerte o débil), además puede ser:

- Positiva: Si al aumentar X aumenta Y
- Negativa: Si al aumentar X, disminuye Y

La correlación más o menos fuerte viene, por tanto, determinada por lo apretados que estén los puntos de la nube en torno a una recta que llamaremos recta de regresión. El signo de la pendiente de esta recta determina si la correlación es positiva (pendiente positiva) o negativa (pendiente negativa).



Tablas de frecuencias

Las tablas de frecuencias para una variable estadística bidimensional pueden ser simples o de doble entrada

Tabla Simple

(X)	3	2	5	4	7	6	2	4	8	9	7	10	7
(Y)	2	2	6	3	6	8	4	4	10	6	7	9	7

Tablas de doble entrada

Y\X	1	2	3	4	5	6	7	8	9	10
2		1	1							
3				1						
4		1		1						
5										
6					1				1	
7							2			
8						1				
9										1
10								1		

De la tabla simple se puede pasar a una de doble entrada y viceversa.
Las distribuciones unidimensionales obtenidas de la tabla anterior:

(X)	n_i=frecuencia
2	2
3	1
4	2
5	1
6	1
7	2
8	1
9	1
10	1
n=	$\sum = 12$

(Y)	n_i=frecuencia
2	2
3	1
4	2
5	0
6	2
7	2
8	1
9	1
10	1
n=	$\sum = 12$

Se llaman distribuciones marginales

Las frecuencias marginales pues, tienen en cuenta una sola variable y dan lugar a dos distribuciones unidimensionales.

La suma de las frecuencias absolutas marginales coincide con la suma de las frecuencias bidimensionales de la tabla de doble entrada.

Cuando las variables son cuantitativas se pueden obtener parámetros representativos como media, mediana, desviación típica,..., pero cuando son variables cualitativas determinamos sólo el porcentaje.

Centro de gravedad de una distribución bidimensional

Llamaremos centro de gravedad de la distribución al punto $(\overline{x}, \overline{y})$ cuyas coordenadas son **las medias** de las distribuciones unidimensionales de X e Y:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

En nuestro ejemplo

$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{2\cdot 2+3+4\cdot 2+5+6+7\cdot 2+8+9+10}{12} = \frac{67}{12} = 5.58333$

$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{2\cdot 2+3+4\cdot 2+6\cdot 2+7\cdot 2+8+9+10}{12} = \frac{68}{12} = 5.666$

Luego el centro de gravedad de la distribución es $(\overline{x}, \overline{y}) = (5.58, 5.6)$

Cálculo de parámetros. Covarianza.

Las medias de las distribuciones de frecuencia marginales se calculan del modo habitual ya conocido, esto es $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ $\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}$.

Varianzas

Las varianzas las calculamos de la siguiente forma: :

$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{\sum x_i^2}{n} - \overline{x}^2$

$s_y^2 = \frac{1}{n}\sum(y_i - \overline{y})^2 = \frac{\sum y_i^2}{n} - \overline{y}^2$

Existe, no obstante en las distribuciones bidimensionales un nuevo parámetro queno existe en las unidimensionales, se trata de la Covarianza S_{xy} , que se define como la media aritmética de los productos de las desviaciones de los valores de cada una de las variables respecto de su media

$$S_{xy} = \frac{\sum x_i y_i}{n} - \overline{x} \cdot \overline{y}$$

3.-Cálculo del coeficiente de correlación lineal.

La correlación mide el grado de ajuste de la nube de puntos a la función matemática asignada.

Responde a la pregunta: ¿en qué medida una recta, describe de un modo adecuado la relación existente entre las variables?.

La relación entre dos variables puede ajustarse muy bien a una recta o cualquier otra función matemática. Para medir el grado de ajuste de la distribución a una recta, se emplea el coeficiente de correlación de Pearson, cuya expresión es:

$$r = \frac{S_{xy}}{S_x S_y}$$

Este coeficiente soluciona los problemas que presentaba la Covarianza por varias razones:

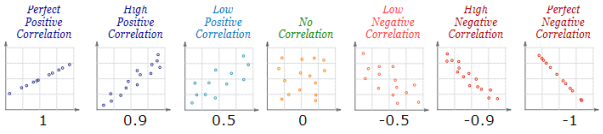
1.- Si el coeficiente de una variable (x, y) es D el de (aAx, bAy) también es D

2.- No tiene unidades, lo que nos permitirá estudiar la correlación con independencia de como tomemos las medidas.

3.2 Interpretación del coeficiente de correlación Lineal

- El signo del coeficiente de correlación de Pearson coincide con el signo de la covarianza , puesto que s_x y s_y son dos números positivos.
- Los valores que puede tomar el coeficiente de correlación de Pearson están comprendidos entre -1 y 1.
- Si $\mathbf{0 < r < 1}$, la correlación es positiva.

- La correlación es positiva o directa cuando al aumentar una variable, se produce un aumento en la otra, y al disminuir una, se produce una disminución en la otra. Esto ocurre cuando la covarianza es positiva.
- Si **-1<r** $<$ **0**, la correlación es negativa. La correlación es negativa, o inversa, cuando al aumentar una variable, se produce una disminución de la otra, y al disminuir una variable, se produce un aumento en la otra. Esto ocurre, cuando la covarianza es negativa.
- Si $r = +1$ ó $r = -1$ el ajuste es perfecto. Cuando se da este caso, las variables X e Y guardan una relación funcional lineal exacta, y = f(x).
- Si r = 1 la recta tiene pendiente positiva y si r = -1 la recta tiene pendiente negativa.
- Si r = 0 no hay recta de regresión, la nube de puntos no se ajusta a una recta



2.- Introducción a la regresión lineal

La regresión es el estudio de los métodos de ajuste de una curva conocida a una nube de puntos. La regresión calcula la expresión matemática de la curva que más se aproxima, o que mejor se ajusta, a la nube de puntos.

Trata, por lo tanto, de averiguar cuál es la función que refleja del modo más exacto la relación entre ambas variables.

Esto nos permitirá estimar y predecir valores para una de las variables a partir de los valores de la otra.

La regresión lineal estudia los distintos métodos, o técnicas, de ajustar **una recta** a una nube de puntos.

Recta de regresión: Significado y cálculo de la recta de regresión de y sobre x.

Cálculo de la recta de regresión de x sobre y. Dada una nube de puntos, la recta de regresión que mejor se ajuste a ella tendrá una ecuación de la forma y = Ax + B.

Para obtener los valores de A y B, se impondrán dos condiciones:

Punto de gravedad de la nube de puntos.

La recta que buscamos pasa por el punto $(\overline{x}, \overline{y})$ es decir su ecuación será

$$(y - \overline{y}) = m(x - \overline{x})$$

Sólo queda por determinar el valor de la pendiente de la recta m

El valor de m es $m = \frac{S_{xy}}{S_X^2}$. Por lo tanto la recta de regresión de y sobre x es:

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

Esta ecuación permite aproximar valores de y conocidos los de x .

Al valor $\frac{S_{xy}}{S_X^2}$ se le denomina Coeficiente de regresión de Y sobre X

Del mismo modo obtenemos la ecuación de la recta de regresión de x sobre y que será: x x que permite aproximar valores de x conociendo los de y.

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

Al valor $\frac{s_{xy}}{s_y^2}$ se le denomina Coeficiente de regresión de X sobre Y

El método de obtención de esta recta se denomina método de mínimos cuadrados y la recta de regresión se llama también recta de mínimos cuadrados.

Ejemplo

En una academia para aprender a conducir se han estudiado las semanas de asistencia a clase de sus alumnos y las semanas que tardan en aprobar el examen teórico (desde que se apuntaron a la autoescuela). Los datos correspondientes a seis alumnos son:

X: Asistencia	6	1	4	3	5	8
Y: Aprobado	6	5	5	6	5	10

a) Halla las dos rectas de regresión y represéntalas.

b) Observando el grado de proximidad entre las dos rectas, ¿cómo crees que será la correlación entre las dos variables?

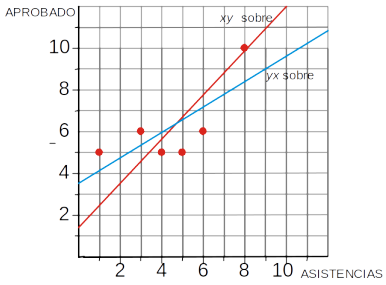
x_i	y_i	x_i^2	y_i^2	$x_i y_i$
6	6	36	36	36
1	5	1	25	5
4	5	16	25	20
3	6	9	36	18
5	5	25	25	25
8	10	64	100	80
27	37	151	247	184

Calculamos el centro de gravedad $\overline{x} = \frac{27}{6} = 4,5 \qquad \overline{y} = \frac{37}{6} = 6,17$

Calculamos las desviaciones típicas marginales

$S_x = \sqrt{\frac{151}{6} - 4,5^2} = \sqrt{4,92} = 2,22 \qquad S_y = \sqrt{\frac{247}{6} - 6,17^2} = \sqrt{3,1} = 1,76$

Covarianza:
$S_{xy} = \frac{\sum x_i y_i}{n} - \overline{x} \cdot \overline{y} \Rightarrow S_{xy} = \frac{184}{6} - 4,5 \cdot 6,17 = 2,9$
Coeficientes de regresión:
$y \quad sobre \quad x \rightarrow \ m_{yx} = \frac{2,9}{4,92} = 0,59$
Rectas de regresión:
$y \quad sobre \quad x \Rightarrow \ y - 6,17 = +0,59(x - 4,5) \ \Rightarrow y = 0,59x + 3,52$
$x \quad sobre \quad y \Rightarrow \ x = 4,5 + 0,94(y - 6,17)$
$x = 4,5 + 0,94y - 5,80$
$x = 0,94y - 1,3$
$x + 1,3 = 0,94y$
$y = \frac{x+1,3}{0,94} \ \Rightarrow y = 1,06x + 1,38$
Representación:



b) La correlación entre las variables no es demasiado fuerte, pues las dos rectas no están muy próximas. Con los datos obtenidos comprobamos que el coeficiente de correlación es: r =0,74

Ejemplo

Una compañía de seguros considera que el número de vehículos (Y) que circulan por una autopista, puede ponerse en función del número de accidentes (X) que ocurren en ella. Durante cinco días se obtuvo los siguiente resultado.

x	5	7	2	1	9
y	15	18	10	8	20

a) Calcula el coeficiente de correlación.
b) Si ayer se produjeron 6 accidentes. ¿ Cuantos vehículos podemos suponer que circulaban por la autopista?
c) ¿Es buena esta predicción?
SOLUCIÓN
Construir la tabla de frecuencias.

	X_i	Y_i	f_i	$Xi \cdot f_i$	$Y_i \cdot f_i$	$fi \cdot X_i{}^2$	$fi \cdot Y_i{}^2$	$fi \cdot X_i \cdot Y_i$
	5	15	1	5	15	25	225	75
	7	18	1	7	18	49	324	126
	2	10	1	2	10	4	100	20
	1	8	1	1	8	1	64	8
	9	20	1	9	20	81	400	180
Σ	24	71	5	24	71	160	1113	409

Calcula el coeficiente de correlación

Hallamos las medias de X y de Y

$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{24}{5} = 4,8$

$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{71}{5} = 14,2.$

Calculamos la covarianza. (La covarianza indica el sentido de la correlación entre las variables):

$$S_{xy} = \frac{\sum x_i y_i}{n} - \overline{x} \cdot \overline{y} = \frac{409}{5} - 4,8 \cdot 14,2 = 13,64$$

Calculamos las desviaciones típicas.

$S_x = \sqrt{\frac{160}{5} - 4,8^2} = 2,99 \qquad S_y = \sqrt{\frac{113}{5} - 14,2^2} = 4,57$

Calculamos el coeficiente de correlación.

$$r = \frac{S_{xy}}{S_x S_y} = \frac{13,64}{2,99 \cdot 4,57} = 0,995$$

) Si ayer se produjeron 6 accidentes. ¿ Cuantos vehículos podemos suponer que circulaban por la autopista? En este apartado lo que nos piden es calcular la recta de regresión de y (Número de vehículos) sobre x (Número de accidentes).

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

$$(y - 14,2) = \frac{13,64}{2,99^2}(x - 4,8)$$

$$(y - 14,2) = \frac{13,64}{2,99^2}(6 - 4,8) \Longrightarrow y = 16,03$$

Es buena esta predicción pues el coeficiente de correlación entre el número de accidentes y la cantidad de vehículos circulando es muy alto casi pegado a 1. (r=0,995)