ESTADÍSTICA BIDIMENSIONAL

ÍNDICE GENERAL

1.-Variable Estadística Bidimensional. Tablas de frecuencia

1.1.- Concepto de variable estadística bidimensional. Ejemplos.

En muchas ocasiones no basta con estudiar la descripción de un fenómeno y sus variaciones, es conveniente conocer a qué son debidas esas variaciones. Puede resultar interesante e incluso necesario estudiar los cambios producidos en una variable en relación con otras, o cómo influyen unas variables para que otra cambie. Cuando se estudian conjuntamente varias variables se entra en el campo de la estadística multivariable (muchas variables). Si el estudio se reduce a dos variables, como en este tema, se llama estadística bidimensional.

La estadística bidimensional estudia fenómenos en los que intervienen dos variables conjuntamente, buscando la relación que existe entre ambas. Así, por ejemplo, se puede estudiar la influencia que tienen los ingresos de una determinada familia en los gastos que tiene, o cómo influye la velocidad de un cierto automóvil en su consumo de combustible, o qué relación existe entre los pesos y las estaturas de un grupo de personas. Una variable bidimensional se representa por un par (X, Y), donde X es la primera variable y toma los valores $x_1, x_2, x_3, ...,x_n$ e Y la segunda y toma los valores, $y_1, y_2, y_3, ...,y_n$.

Sin embargo, al considerar dos variables de una población o muestra, no podemos afirmar que se trata de una variable bidimensional porque la relación entre las variables puede no ser estadística. Así, entre dos variables puede existir:

Dependencia Funcional.

Cuando es posible predecir con exactitud los valores de una variable a partir de los de la otra, se dice que ambas variables están en relación funcional. Dada la variable (X,Y) existirá una función f(x) tal que $y_i = f(x_i)$. Para cada valor de x se puede conocer el valor de y.

Ejemplo:

a) La altura desde la que cae un cuerpo y el tiempo que tarda en llegar al suelo está sujeto a la ley de la gravedad. Siempre tarda lo mismo en recorrer el mismo espacio.

b) El precio de una tela es función del coste del metro de tela y del número de metros.

Independencia o Incorrelación.

Cuando las dos variables no tienen ninguna relación entre ellas y podemos estudiarlas por separado.

Ejemplo:

a) La estatura y la nota de matemáticas.

b) La nota en selectividad y el número de letras del nombre.

Dependencia estadística o correlación.

Cuando no podemos establecer una relación funcional pero tampoco podemos afirmar que no existe interrelación, se dice que están en relación estadística.

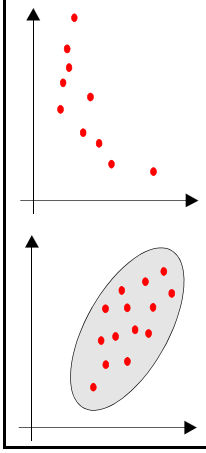
Lo que nos proponemos a lo largo de estos apartados es determinar el grado de correlación que existe entre las variables.

Ejemplo:

a) De un colectivo se anota el número de horas de sueño y la edad de cada individuo. Hay una tendencia general a afirmar que a mayor edad menos horas de sueño, pero no podemos afirmar con exactitud que a X años le corresponden Y horas de sueño.

b) Otro ejemplo de relación estadística es el nº de cigarrillos consumidos y el riesgo de fallo cardiaco.

Diagrama de dispersión. Nube de Puntos



Una manera sencilla y eficaz de estudiar la posible correlación entre variables es recurrir a los **diagramas de dispersión**, que son representaciones gráficas en un sistema de ejes cartesianos donde cada par (x, y) se representa por un punto. La representación gráfica resultante se denomina diagrama de dispersión o nube de puntos.

Podemos apuntar un par de ideas sobre la nube de puntos:

1.- En muchas ocasiones la nube de puntos sugiere la forma de la gráfica de alguna función conocida: una recta, una parábola, una función exponencial. Esto significa que puede existir alguna relación entre las variables. Si así ocurriese, se diría que las variables están correlacionadas.

2.- Si la forma de la nube es estirada y sus puntos se pueden encerrar en una elipse, la estrechez de esa elipse es un indicador de la fuerza de la correlación

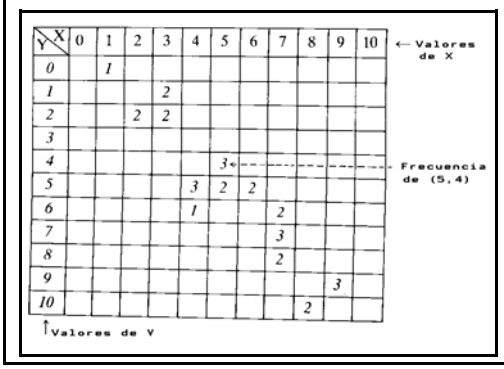
1.2.-Tablas bidimensionales de frecuencias. Tablas de doble entrada.

Lo primero que tenemos que resolver es como tabular los datos obtenidos en las observaciones. Podemos construir una tabla de frecuencias, pero son mucho más prácticas las tablas de doble entrada que reflejan los valores y frecuencias de las variables bidimensionales.

Ejemplo: Las notas en Lengua y en Idioma de los 30 alumnos de una clase en la última evaluación han sido:

Lengua: 3, 7, 8, 7, 5, 2, 5, 9, 5, 4, 3, 5, 3, 6, 3, 8, 5, 7, 7, 6, 2, 4, 9, 4, 9 , 7, 6, 7, 1, 7

Idioma: 2, 6, 10, 6, 4, 2, 5, 9, 5, 5, 2, 4, 1, 5, 1, 10, 4, 7, 8, 4, 2, 5, 9, 5, 9, 8, 5, 7,0,7



Para el estudio de los resultados podemos disponer las notas en una tabla de doble entrada, en la que junto a los distintos resultados aparezcan sus frecuencias. Obtenemos así la distribución de frecuencias de la variable estadística bidimensional (X,Y), donde X representa la nota en Lengua e Y la nota en Idioma de los alumnos de la clase.

Escribe la distribución de frecuencias de **X="nota en Lengua**". Calcula su media y su varianza.
Escribe la distribución de frecuencias de **Y="nota en Idioma**". Calcula su media y su varianza.

Observa esta nueva tabla en la que se ha añadido una fila y una columna más con los totales:

Y \ X	0	1	2	3	4	5	6	7	8	9	10	Total
0		1										1
1				2								2
2			2	2								4
3												0
4						3						3
5					3	2	2					7
6					1			2				3
7								3				3
8								2				2
9										3		3
10									2			2
Total	0	1	2	4	4	5	2	7	2	3	0	30

1.3.-Distribuciones marginales

Se denomina distribución marginal de una variable bidimensional a la distribución que se obtiene al estudiar independientemente cada variable.

Si tomamos la primera columna y la última columna en la tabla anterior, obtenemos la distribución de frecuencias marginales de la variable estadística **Y**:

Y	0	1	2	3	4	5	6	7	8	9	10
n_i	1	2	4	0	3	7	3	3	2	3	2

Si tomamos la primera fila y la última, obtenemos la distribución de frecuencias de **X**:

X	0	1	2	3	4	5	6	7	8	9	10
n'_i	0	1	2	4	4	5	2	7	2	3	0

Con estas distribuciones podemos calcular los mismos parámetros estadísticos que calculamos para las distribuciones unidimensionales.

Las medias marginales son: $\overline{x}=\dfrac{\sum_{i=1}^{p} x_i \cdot n_i}{N}$ $\overline{y}=\dfrac{\sum_{i=1}^{q} x_j \cdot n'_i}{N}$

Varianzas marginales son:

$$S_x^2 = \frac{\sum_i (x_i - \overline{x}) \cdot n_i}{N} = \overline{(x^2)} - (\overline{x})^2 \qquad S_y^2 = \frac{\sum_j (y_j - \overline{y}) \cdot n'_j}{N} = \overline{(y^2)} - (\overline{y})^2$$

1.4.-Vector de medias.

Sea (X,Y) una distribución estadística bidimensional. Al par $(\bar{x}, \bar{y})$ se le denomina vector de medias o centro de gravedad de la distribución.

Ejercicios:

1) Un vendedor de helados anota durante doce días la temperatura (T) a las doce de la mañana y el número de bloques vendidos (V) en ese día, obteniendo los siguientes valores: $(30^0,10)$, $(27^0,8)$, $(28^0,9)$, $(27^0,8)$, $(30^0,10)$, $(31^0,11)$, $(27^0,9)$, $(28^0,10)$, $(29^0,11)$, $(30^0,11)$, $(29^0,12)$, $(30^0,10)$.

Escribe la distribución de frecuencias de la variable bidimensional (T,V) en forma de tabla de doble entrada. Escribe las distribuciones marginales de la distribución anterior y calcula la temperatura media y el número medio de bloques vendidos. Calcula en las dos distribuciones marginales la varianza y la desviación típica.

2) Hemos preguntado a los 20 alumnos de una clase el número de horas semanales que dedican al estudio (E) y el número de horas semanales que ven televisión (T):

E	2	5	6	5	3	1	4	0	2	1	3	4	3	2	1	1	2	4	0	1
T	1	7	2	7	6	9	5	5	9	6	7	5	6	8	5	5	9	5	5	8

Construye una tabla de doble entrada. escribe las distribuciones marginales de ambas variables y calcula sus medias y varianza.

3) Las alturas (X) y los pesos (Y) de 25 personas son los siguientes:

X (Kgr.)	[60-65)	[60-65)	[65-70)	[65-70)	[65-70)
Y (Cm.)	[165-170)	[170-175)	[165-170)	[170-175)	[175-180)
Frecuencia	**1**	**3**	**2**	**4**	**2**
X (Kgr.)	[70-75)	[70-75)	[70-75)	[75-80)	[80-85)
Y (Cm.)	[165-170)	[170-175)	[175-180)	[170-175)	[170-175)
Frecuencia	**1**	**4**	**3**	**3**	**2**

Expresa estos resultados mediante una tabla de doble entrada. Escribe las distribuciones marginales y calcula sus medias y varianzas (toma como valor de cada intervalo el punto medio, Marca de clase).

1.5.-Distribuciones condicionadas

Son las distribuciones que se obtienen al fijar un valor en una de las variables y estudiar las frecuencias correspondientes a la otra.

Por ejemplo la distribución de la variable Y para el valor $X=x_i$.

La distribución que se obtiene es unidimensional.

1.6.-Covarianza: Concepto y cálculo. Matriz de covarianza.

Se llama covarianza de la variable (X,Y) a la media aritmética de los productos de las desviaciones de cada variable respecto de la media.. También se le denomina varianza conjunta o sincronizada de las variables X e Y.

La covarianza es la medida más simple de la relación lineal entre dos variables. Viene dada por: $\sigma_{xy} = \sum_i (x_i - \bar{x}) \cdot \sum_j (y_j - \bar{y}) \cdot f_r(x_i, y_j) = \dfrac{\sum_{i,j} (x_i - \bar{x}) \cdot (y_j - \bar{y}) \cdot n_{ij}}{N}$

Se demuestra que: $\sigma_{xy} = \left(\frac{\sum_{i,j} x_i \bullet y_j \bullet n_{ij}}{N} \right) - \overline{x} \bullet \overline{y}$

Interpretación de la covarianza

Una **covarianza positiva y alta** indica que ambas variables crecen o decrecen simultáneamente, es decir, presentan una fuerte correlación. Cuando mayor sea la covarianza, más estrecha es la relación entre las variables.

Una **covarianza alta y negativa** indica que cuando una variable crece, la otra decrece y viceversa, es decir, presentan una fuerte correlación inversa. Cuanto menor sea la covarianza, puesto que es negativa, más estrecha es esta relación entre las variables.

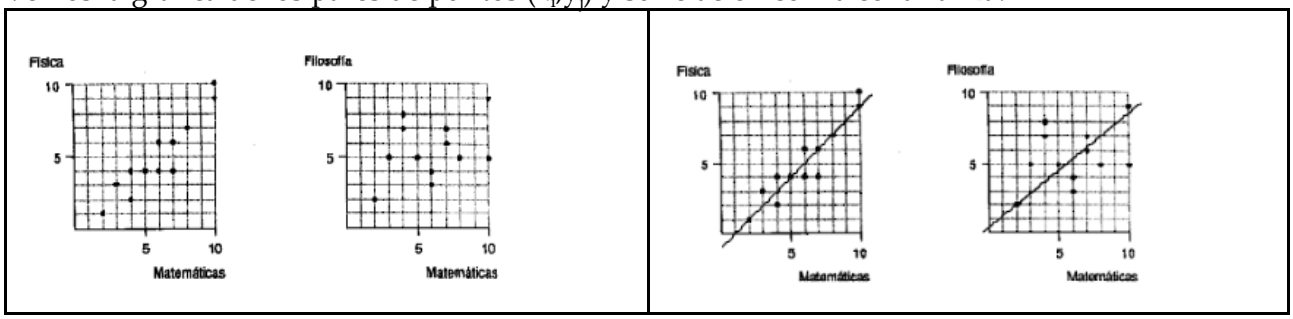
La **covarianza cero** o próxima a cero indica que no existe relación entre las variables.

Ejemplo:

A 12 alumnos de un colegio se les toma las notas de los últimos exámenes de matemáticas, física y filosofía. Observa que hay una relación fuerte entre las notas de matemáticas y las de física.

ALUMNO	MATEMÁTICAS.SSS	FÍSICA	FILOSOFÍA
A	2	1	2
B	3	3	5
C	4	2	7
D	4	4	8
E	5	4	5
F	6	4	3
G	6	6	4
H	7	4	6
I	7	6	7
J	8	7	5
K	10	9	5
L	10	10	9

Vemos la gráfica de los pares de puntos (x_i,y_j) y su relación con la covarianza.



En la primera gráfica, los puntos están más alineados y por tanto la relación entre las variables (**CORRELACIÓN**) es más fuerte. Por el contrario, en la segunda gráfica es más débil.

Pero la covarianza presenta algún inconveniente:

1- Los puntos más alejados de la nube influyen más en su valor y signo que los centrales.

2- Las escalas influyen en el valor de la covarianza. Así, al cambiar la escala cambia el valor de la covarianza y sin embargo, la relación entre las variables es la misma.

Matriz de covarianzas

Podemos agrupar los parámetros anteriores en la siguiente matriz:

$$M = \begin{pmatrix} S_x^2 & S_{xy} \\ S_{xy} & S_y^2 \end{pmatrix}$$

2.- Introducción a la regresión lineal

2.1.- Idea intuitiva del ajuste de una linea a un diagrama de dispersión

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.

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

2.2 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:

1.- Gravedad de la nube de puntos. Esta condición implica que la recta de regresión pasa por el punto $(\bar{x}, \bar{y})$ es decir su ecuación será $y - \bar{y} = A \bullet (x - \bar{x})$. Sólo queda por determinar el valor de la pendiente de la recta, A.

2.- A cada punto P_i, de coordenadas (x_i, y_i), perteneciente a la nube de puntos, le corresponde, en la recta, el punto P_i' de coordenadas (x_i, y'_i). Si se llamamos Di a la diferencia yi - y'i, se impondrá la condición de que la suma de los cuadrados de estas diferencias sea mínima.

Puesto que el punto (x_i, y'_i) pertenece a la recta se verifica que $y'_i = \bar{y} + A \bullet (x_i - \bar{x})$

Como D_i^2 tiene que ser mínimo, para cometer el menor error. Entonces la derivada de $\sum_i D_i^2 = \sum_i (y'_i - y_i)^2 = \sum_i \left(\bar{y} + A \bullet (x_i - \bar{x}) - y_i \right)^2$ con respecto a "A" debe de ser 0. De esta condición, y mediante un tratamiento matemático, se deduce que el valor de A debe ser $A = S_{xy} / S_x^2$. Por lo tanto la recta de regresión de y sobre x es:

$y - \bar{y} = \frac{S_{xy}}{S_x^2}(x - \bar{x})$ 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:

$x - \bar{x} = \frac{S_{xy}}{S_y^2}(y - \bar{y})$ que **permite aproximar valores de x conociendo los de 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, minimizando la suma de los cuadrados de las diferencias yi - y'i, se denomina método de mínimos cuadrados y la recta de regresión se llama también recta de mínimos cuadrados.

Los valores de la aproximación serán mejores si el coeficiente de correlación se acerca a 1 o -1.

Interpolación y extrapolación .

La recta de regresión puede utilizarse para predecir el valor de Y que corresponde a un determinado valor de X conocido. Se llama interpolación a la estimación de un valor de la variable Y para un cierto valor de X, dentro de su recorrido. Se llama extrapolación a la estimación de un valor de Y, para un cierto valor de X fuera de su recorrido.

Ejemplo:

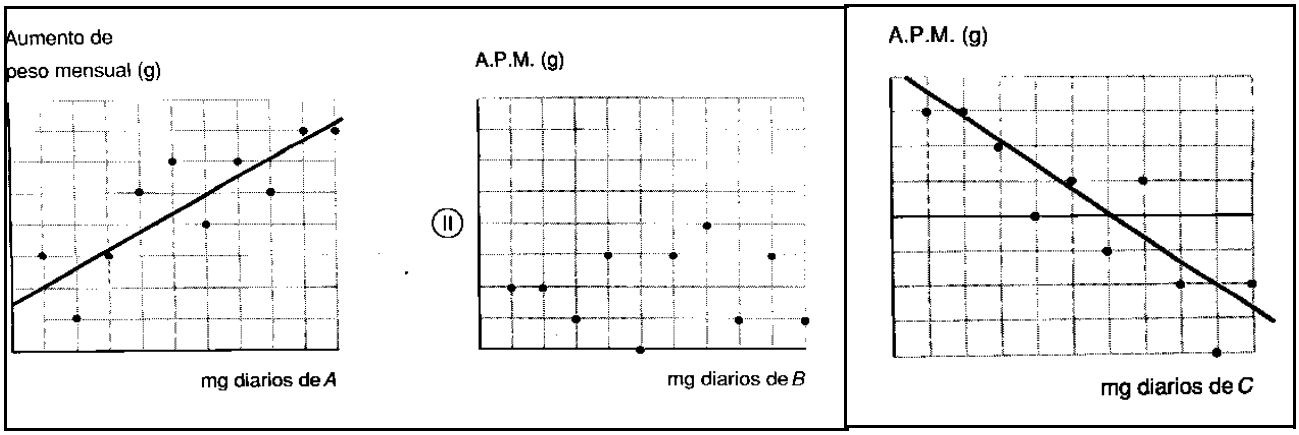
Realizamos un experimento que consiste en suministrar a cada una de 10 ratas una dosis diaria de 1 mg, 2 mg, 3 mg, ..., 10 mg, respectivamente, de un cierto fármaco A, y calculamos el aumento de peso de cada rata después de un mes.

Realizamos el mismo experimento con otras 10 ratas y otro fármaco B. Y por último un tercer experimento con otras 10 ratas y otro fármaco C.

Los resultados gráficamente son:

A la vista de las tres gráficas, nos inclinamos a pensar que A favorece el engorde de las ratas, B no influye y C es perjudicial.

La correlación de la gráfica 1 es positiva y la de la 3 es negativa, igual que las pendientes de las rectas de regresión correspondientes. En la segunda gráfica se observa que la nube de puntos es amorfa y no sugiere ninguna recta. No hay correlación entre las variables. Se dice que son *Incorreladas*.



3.-Significado y cálculo del coeficiente de correlación.

3.1 Coeficiente de correlación lineal: Definición y cálculo.

La correlación mide el grado de ajuste de la nube de puntos a la función matemática asignada. Responde por tanto a la pregunta: ¿en qué medida una recta, u otra función matemática, 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:

$$\rho_{xy} = \frac{S_{xy}}{S_x \cdot S_y} = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$$

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

1.- Si el coeficiente de una varia ble (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 σ_{xy}, puesto que *$_x$ y *$_y$ son dos números positivos.

Los valores que puede tomar el coeficiente de correlación de Pearson están comprendidos entre -1 y 1.

A **Si 0<D$_{xy}$ <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.

A **Si -1<D$_{xy}$ < 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 D$_{xy}$ = $\pm$1 el ajuste es perfecto. Cuando se da este caso, las variables X e Y guardan una relación funcional lineal exacta, y = f(x). Si **D$_{xy}$ = 1** la recta tiene pendiente positiva y si **D$_{xy}$ = -1** la recta tiene pendiente negativa.

Si D$_{xy}$ = 0 no hay recta de regresión, la nube de puntos no se ajusta a una recta