DISTRIBUCIONES BIDIMENSIONALES

1. Variables estadísticas bidimensionales
2. Tablas bidimensionales de frecuencias.
3. Cálculo de parámetros. Covarianza.
4. Correlación.
5. Regresión lineal.

- Ejemplos.
- Problemas

1 Variables estadísticas bidimensionales

Hasta ahora, hemos descrito el estudio que puede realizarse de una población o muestra respecto a una sola variable estadística.

Lo que pretendemos ahora es abordar el estudio de un fenómeno respecto a dos variables unidimensionales simultáneamente, se obtiene así el concepto de variable estadística bidimensional en la que cada elemento de la misma vendrá representado por un par ordenado (x_i, y_i).

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

- Los pesos y las estaturas de un conjunto de personas.
- El número de encuentros ganados por un equipo de fútbol y el lugar que ocupa en la clasificación.
- Las notas obtenidas por cada alumno de una clase en dos asignaturas de similares características.
- Las velocidades a las que circulan un conjunto de vehículos y su consumo de combustible.
- Extensión en km^2 y número de habitantes de los distintos países de Europa.
- Ingresos y gastos de cada una de las familias de los trabajadores de una empresa.
- Renta nacional y número de universitarios de los distintos países de África.
- Edad y número de días que faltan al trabajo los empleados de una fábrica.
- Número de horas que dedican los estudiantes a ver la televisión 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 toma los valores $x_1, x_2, x_3, ..., x_k$ e 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:

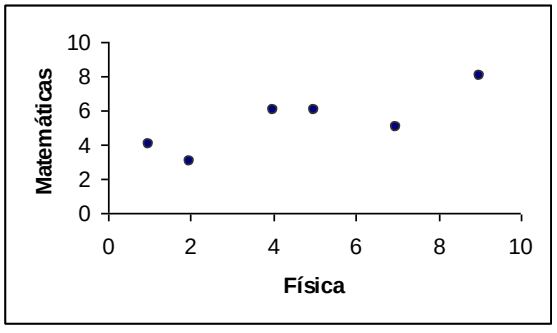
$(x_1, y_1), (x_2, y_2), (x_3, y_3), ..., (x_k, y_k)$ o también $(x_i, y_i),\ 1 \le i \le k$.

Si representamos los valores de ambas variables en una tabla de dos filas o columnas obtendremos una especie de tabla de valores similar a las que nos encontramos en la representación gráfica de una función. Ello nos sugiere representarlos sobre dos ejes de coordenadas poniendo $\{x_i\}$ en abcisas e $\{y_i\}$ en ordenadas, obteniendo lo que llamaremos una **nube de puntos** o **diagrama de dispersión**.

Ejemplo: Estudiamos las notas obtenidas por 6 alumnos en las asignaturas de Física y Matemáticas, viniendo los resultados recogidos en la siguiente tabla:

Física (x_i)	2	4	7	1	5	9
Matemáticas (y_i)	3	6	5	4	6	8

La nube de puntos que se obtiene es:



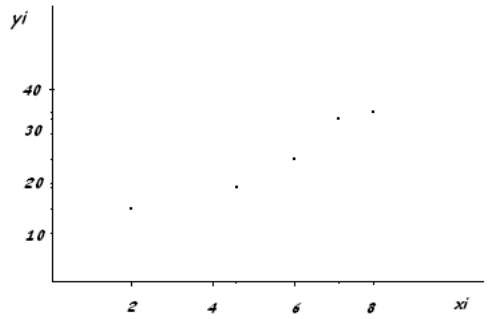
2 Tablas bidimensionales de frecuencias

Veamos algunos ejemplos:

Ejemplo1: Se observaron las edades de cinco niños y sus pesos respectivos, y se consiguieron los resultados siguientes:

Edad (Años)	2	4'5	6	7'2	8
Peso (Kg)	15	19	25	33	34

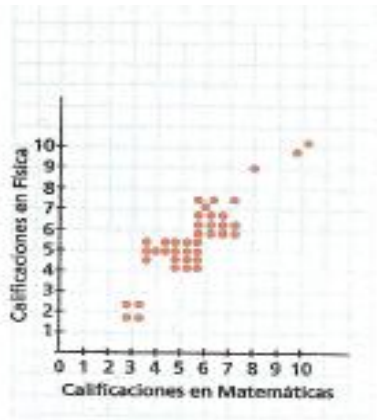
Su diagrama de dispersión es:



Ejemplo 2: Las calificaciones obtenidas por 40 alumnos en Matemáticas y Física son:

X= Notas Matemáticas	3	4	5	6	6	7	7	8	10
Y= Notas Física	2	5	5	6	7	6	7	9	10
Nº de alumnos (f_{ij})	4	6	12	4	5	4	2	1	2

Esto significa que, por ejemplo, hay 4 alumnos en total que han sacado un 3 en Matemáticas y un dos en Física.
A estos tipos de tabla se le denomina tablas simples. Su diagrama de dispersión es:



Las tablas de frecuencias para una variable estadística bidimensional pueden ser **simples o de doble entrada**
Veamos, ampliando nuestro ejemplo anterior sobre las notas de los alumnos cómo sería una tabla de doble entrada. De la tabla simple anterior se puede pasar a una de doble entrada y viceversa.

En nuestro caso tenemos:
4 alumnos con un 3 en Matemáticas. y un 2 en Física.
6 alumnos con 4 en Matemáticas. y 5 en Física.
12 alumnos con 5 en Matemáticas. y 5 en Física.
4 alumnos con 6 en Matemáticas. y 6 en Física.
5 alumnos con 6 en Matemáticas y 7 en Física.
4 alumnos con un 7 en Matemáticas y un 6 en Física.
2 alumnos con 7 en Matemáticas. y 7 en Física.
1 alumno con 8 en Matemáticas y 9 en Física.
2 alumnos con 10 en Matemáticas y 10 en Física.
A este tipo de tabla se le denomina **tabla de doble entrada**.
Estas tablas se utilizan cuando se trata de muchos datos o bien los valores se encuentran agrupados en intervalos.
Donde f_{ij} es la frecuencia absoluta conjunta de ambas variables.

La tabla de doble entrada sería:

Y / X	2	5	6	7	9	10	f_i
3	4						4
4		6					6
5		12					12
6			4	5			9
7			4	2			6
8					1		1
10						2	2
f_i	4	18	8	7	1	2	40

Cada valor de una casilla interna es la frecuencia absoluta conjunta f_{ij} (número de veces que aparece el par dado por el número indicado en su fila y el indicado en su columna en los encabezamientos); cada valor de la última fila o columna es la frecuencia absoluta simple de cada una de las variables f_i ó f_j (número de veces que aparece el valor indicado en la cabecera de la fila o la columna aisladamente). La última casilla de la tabla indica el número total de casos en estudio (N).

Las distribuciones unidimensionales obtenidas de la tabla anterior:

x_i	f_i
3	4
4	6
5	12
6	9
7	6
8	1
10	2
	40

y

y_i	f_i
2	4
5	18
6	8
7	7
9	1
10	2
	40

Se llaman **distribuciones marginales**.

En resumen, las tablas de frecuencias bidimensionales pueden ser: simples o de doble entrada.

Tablas simples: Una tabla de frecuencias simple es la que recoge en filas o columnas las frecuencias de los valores $(x_i, y_i),\ 1 \le i \le k$, de la variable.

X	x_1	x_2	x_3	....	x_k
Y	y_1	y_2	y_3	....	y_k
f_i	f_1	f_2	f_3	....	f_k

Tablas de doble entrada: Una tabla de doble entrada es la que recoge las frecuencias de los $(x_i, y_j),\ 1 \le i \le k,\ 1 \le j \le p$, de las variables.

Y / X	y_1	y_2	y_3	....	y_p
x_1	f_{11}	f_{12}	f_{13}	....	f_{1p}
x_2	f_{21}	f_{22}	f_{23}	....	f_{2p}
x_3	f_{31}	f_{32}	f_{33}	....	f_{3p}
....	....	....	....	....	....
x_k	f_{k1}	f_{k2}	f_{k3}	....	f_{kp}

Frecuencias marginales:

Si en una tabla de doble entrada sumamos las frecuencias absolutas por filas y por columnas, obtenemos una nueva fila y una nueva columna: son las frecuencias marginales.

X \ Y	y_1	y_2	y_3	....	y_p	
x_1	f_{11}	f_{12}	f_{13}	....	f_{1p}	$f_{1.}$
x_2	f_{21}	f_{22}	f_{23}	....	f_{2p}	$f_{2.}$
x_3	f_{31}	f_{32}	f_{33}	....	f_{3p}	$f_{3.}$
....	....	....	....	....	....	....
x_k	f_{k1}	f_{k2}	f_{k3}	....	f_{kp}	$f_{k.}$
	$f_{.1}$	$f_{.2}$	$f_{.3}$	....	f.p	$\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij} = N$

Estas frecuencias obtenidas tienen en cuenta una sola variable y se puede construir con ellas dos distribuciones unidimensionales y obtener los parámetros representativos.

DISTRIBUCIÓN MARGINAL DE X.

X	$f_{i.}$
x_1	$f_{1.}$
x_2	$f_{2.}$
x_3	$f_{3.}$
....	....
x_k	$f_{k.}$
	N

DISTRIBUCIÓN MARGINAL DE Y.

Y	y_1	y_2	y_3	....	y_p	
$f_{.i}$	$f_{.1}$	$f_{.2}$	$f_{.3}$	....	$f_{.p}$	N

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.

Frecuencias condicionadas.

Si en una tabla de doble entrada nos centramos en los valores de una variable con la condición de que el correspondiente valor de la otra sea fijo, tenemos una distribución unidimensional, llamada distribución de la variable en cuestión condicionada al valor tomado como referencia en la otra variable.

DISTRIBUCIÓN DE X CONDICIONADA A $Y = y_j$

X \ Y=yj	y_j
x_1	f_{1j}
x_2	f_{2j}
x_3	f_{3j}
....	....
x_k	f_{kj}

DISTRIBUCIÓN DE Y CONDICIONADA A $X = x_i$

Y / X=xi	y_1	y_2	y_3	....	y_p
x_i	f_{i1}	f_{i2}	f_{i3}	....	f_{ip}

Como es lógico, entre la frecuencia absoluta conjunta y las frecuencias de las marginales se ha de cumplir la relación:

$$\sum_i \sum_j f_{ij} = \sum_i f_i = \sum_j f_j = N$$

3 Cálculo de parámetros. Covarianza.

Las medias de las distribuciones de frecuencia marginales se calculan del modo habitual ya conocido, esto es:

	Variable X	Variable Y
Media	$\bar{x} = \frac{\sum_{i=1}^{n} x_{i.} \cdot f_i}{N}$	$\bar{y} = \frac{\sum_{i=1}^{n} y_{i.} \cdot f_i}{N}$
Varianza	$\sigma^2_x = S_x^2 = \frac{\sum_{i=1}^{n} f_i \cdot (x_i - \bar{x})^2}{N}$	$\sigma^2_y = S_y^2 = \frac{\sum_{i=1}^{n} f_i \cdot (y_i - \bar{y})^2}{N}$
	$\sigma^2_x = S_x^2 = \frac{\sum_{i=1}^{n} f_i \cdot x_i^2}{N} - \bar{x}^2$	$\sigma^2_y = S_y^2 = \frac{\sum_{i=1}^{n} f_i \cdot y_i^2}{N} - \bar{y}^2$

Al par $(\bar{x}, \bar{y})$ se le llama centro de gravedad de la distribución.

Existe, no obstante en las distribuciones bidimensionales un nuevo parámetro que no existía en las unidimensionales se trata de la **covarianza s_{xy} o σ_{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.

$$\sigma_{xy} = S_{xy} = \frac{\sum_{i=1}^{n} f_i \cdot (x_i - \bar{x})(y_i - \bar{y})}{N} = \frac{\sum_{i=1}^{n} f_i \cdot x_i y_i}{N} - \bar{x} \cdot \bar{y}$$

Vamos a demostrar ésta última fórmula que es más cómoda de utilizar, si operamos:

$$s_{xy} = \frac{\sum_i (x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x}\bar{y}) f_i}{N} =$$

$$\frac{\sum x_i y_i f_i}{N} - \bar{y} \cdot \frac{\sum x_i f_i}{N} - \bar{x} \cdot \frac{\sum y_i f_i}{N} + \bar{x}\bar{y} \frac{\sum f_i}{N} = \frac{\sum x_i y_i f_i}{N} - \bar{x}\bar{y} - \bar{x}\bar{y} + \bar{x}\bar{y} =$$

$$\frac{\sum x_i y_i f_i}{N} - \bar{x}\bar{y}$$

Ejemplo: en el caso de las notas que nos ocupa se tendrá, usando la tabla de frecuencias simple colocada en columnas:

X= Notas Matemáticas	3	4	5	6	6	7	7	8	10
Y= Notas Física	2	5	5	6	7	6	7	9	10
Nº de alumnos (f_{ij})	4	6	12	4	5	4	2	1	2

x	y	f_i	$x_i.f_i$	$y_i.f_i$	$x_i.y_i$ .fi	$x_i^2.f_i$	$y_i^2.f_i$	
3	2	4	12	8	24	36	16	
4	5	6	24	30	120	96	150	
5	5	12	60	60	300	300	300	
6	6	4	24	24	144	144	144	
6	7	5	30	35	210	180	245	
7	6	4	28	24	168	196	144	
7	7	2	14	14	98	98	98	
8	9	1	8	9	72	64	81	
10	10	2	20	20	200	200	200	
		40	220	224	1336	1314	1378	

Y los parámetros serán:

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

$$s_x^2 = \frac{\sum f_i x_i^2}{N} - \bar{x}^2 = \frac{1314}{40} - (5{,}6)^2 = 32{,}85 - 30{,}25 = 2{,}6 \qquad s_x = \sqrt{s_x^2} = \sqrt{2{,}6} = 1{,}61$$

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

Por último la covarianza será:

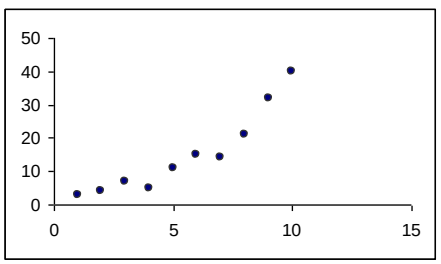
$$s_{xy} = \frac{\sum f_i x_i y_i}{N} - \bar{x}.\bar{y} = \frac{1336}{40} - (5{,}3).(5{,}6) = 33{,}4 - 30{,}8 = 2{,}6$$

La covarianza en un parámetro que tiene la siguiente interpretación:

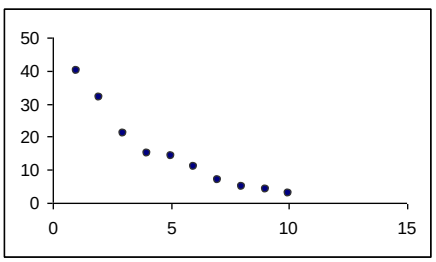
- Si es positiva: las variables X e Y tienen relación directa (al aumentar X aumenta Y)
- Si es negativa: las variables X e Y tienen relación inversa (al aumentar X disminuye Y).

Las nubes de puntos para una relación directa tienen el aspecto siguiente:

Relación directa

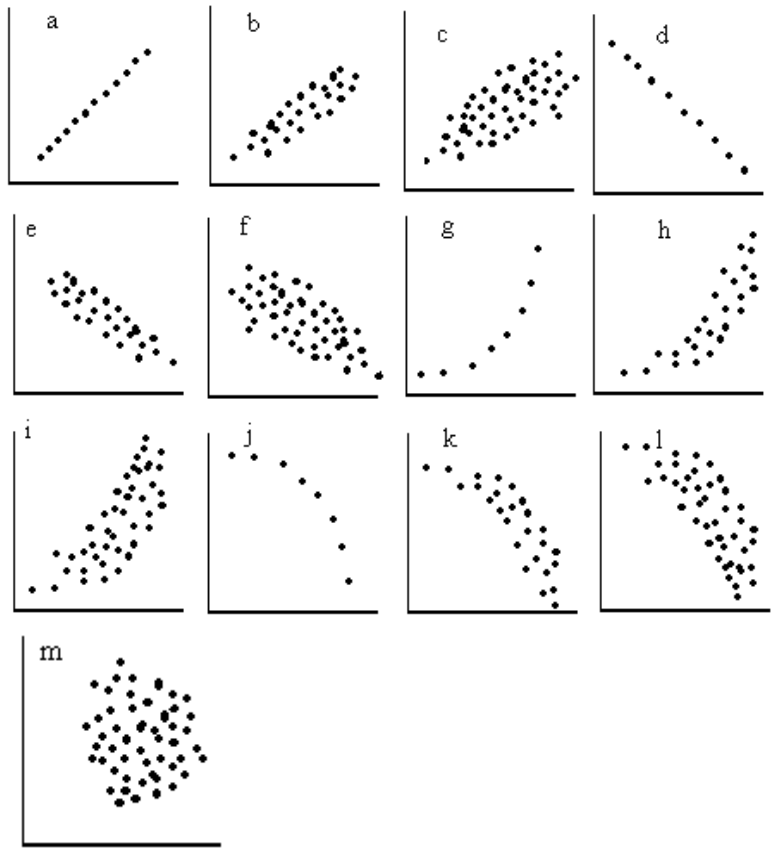


Relación inversa



4 Correlación.

Representamos a continuación varios diagramas de dispersión:



A la vista de estos diagramas, podemos hacernos las siguientes preguntas:

¿Existe alguna relación entre las variables X e Y?
Si existe, ¿es lineal o curvilínea?
Será lineal si los puntos se condensan en torno a una recta.
Será curvilínea si los puntos se condensan en torno a una curva.
¿Al crecer una variable crece la otra? (relación directa o positiva), o ¿al crecer una variable la otra disminuye? (relación inversa o negativa).
¿Es la relación funcional?
Será funcional cuando ambas variables estén relacionadas por una función. En caso contrario, será tanto más fuerte o más débil dependiendo de la mayor o menor tendencia de los puntos del diagrama a acercarse a la representación de una función.

De una manera general, llamaremos **correlación** a la teoría que trata de estudiar la relación o dependencia que existe entre las dos variables que intervienen en una distribución bidimensional.

La correlación es lineal o curvilínea según que el diagrama de puntos se condense en torno a una línea recta o una curva, respectivamente.
La correlación es positiva o directa cuando a medida que crece una variable la otra también crece.
La correlación es negativa o inversa cuando a medida que crece una variable la otra decrece.
La correlación es nula cuando no existe ninguna relación entre ambas variables. En este caso los puntos del diagrama están esparcidos al azar sin formar ninguna línea, y se dice que las variables están incorreladas.
La correlación es de tipo funcional si existe una función tal que todos los valores de la distribución la satisfacen.
En caso contrario, será tanto más fuerte o más débil dependiendo de la mayor o menor tendencia de los valores de la distribución a satisfacer una determinada función.

Coeficiente de correlación lineal.-

El procedimiento más frecuentemente utilizado para asignar valores a las posibles correlaciones entre las variables es el **coeficiente de correlación lineal de Pearson**.

El coeficiente de correlación de Pearson se define mediante la siguiente expresión:

$$r = \frac{s_{XY}}{s_X \cdot s_Y}$$

Observaciones:

- El cálculo práctico del coeficiente de correlación lineal "r" resulta muy sencillo una vez que se sabe calcular la covarianza de la variable (X,Y), así como las desviaciones típicas de las variables X e Y.
- El signo del coeficiente "r" viene dado por el signo de la covarianza, ya que las desviaciones típicas son siempre positivas. Así pues, el signo de la covarianza decide el comportamiento de la correlación:
    - Si la covarianza es positiva la correlación es directa.
    - Si la covarianza es negativa la correlación es inversa.
    - Si la covarianza es nula no existe correlación.

- El coeficiente de correlación lineal es un número real comprendido entre -1 y 1.
- Si $r = -1$ se puede demostrar que todos los valores de la variable bidimensional (X,Y) se encuentran situados sobre una recta; en consecuencia satisfacen la ecuación de una recta. Entonces se dice que entre las variables X e Y existe una dependencia funcional.
- Si $-1 < r < 0$ la correlación es negativa y será tanto más fuerte a medida que r se aproxima más a -1 y tanto más débil a medida que se aproxima a 0. En este caso se dice que las variables X e Y están en dependencia aleatoria.
- Si $r = 0$ entonces no existe ningún tipo de relación entre las dos variables. En este caso se dice que las variables X e Y son aleatoriamente independientes.
- Si $0 < r < 1$ la correlación es positiva y será tanto más fuerte a medida que r se aproxima a 1 y tanto más débil a medida que se aproxima a 0. En este caso se dice que las variables X e Y están en dependencia aleatoria.
- Si $r = 1$ se puede demostrar que todos los valores de la variable bidimensional (X,Y) se encuentran situados sobre una recta; en consecuencia satisfacen la ecuación de una recta. En este caso se dice que entre las variables X e Y existe una dependencia funcional.

Podemos estimar un poco más la correlación según los valores de r según la siguiente tabla:

VALORES DE r	Intensidad de la correlación
$\|r\| = 1$	Correlación perfecta(dependencia funcional)
$0,8 < \|r\| < 1$	Correlación muy alta
$0,6 < \|r\| < 0,8$	Correlación alta
$0,4 < \|r\| < 0,6$	Correlación moderada
$0,2 < \|r\| < 0,4$	Correlación baja
$0 < \|r\| < 0,2$	Correlación muy baja
$r = 0$	Correlación nula

En el ejemplo anterior, el coeficiente de correlación es:

$$r = \frac{2,6}{(1,61).(1,75)} = 0,92$$

Por lo que podemos decir que la correlación entre las notas de física y matemáticas es directa y relativamente fuerte.

5 Regresión lineal

En las distribuciones bidimensionales suele considerarse necesario ser capaz de expresar mediante una relación matemática (como si de una relación funcional se tratase), la relación que existe entre las dos variables.

De esta forma se podrán hacer estimaciones de los valores que adopta una de ellas conocidos algunos de la otra. Estas estimaciones serán más o menos fiables dependiendo de cuánto se aproxime a la unidad el coeficiente de correlación (a más aproximación a la unidad más fiabilidad en la estimación).

En el caso que nos ocupa de la regresión lineal, vamos a tratar de encontrar la ecuación de una recta que "se aproxime lo más posible a todos los puntos de la nube de la variable bidimensional". Pero, ¿Qué significa que se aproxime lo más posible? Consideraremos que se aproxima lo más posible cuando la suma de los cuadrados de las diferencias entre cada valor y_i de la variable y el valor y que predice la recta buscada sea lo menor posible (ajuste por mínimos cuadrados). Las rectas de regresión que se obtienen con esta condición son:

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

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

Vamos a demostrar que las rectas buscadas son las escritas anteriormente:

Supongamos que la recta:

$y=a+bx$ satisface la condición de ajuste por mínimos cuadrados, hemos de determinar los coeficientes a y b con esa condición. Los valores de y en esta recta correspondientes a $\{x_1, x_2,...x_N\}$ son $\{a+bx_1, a+bx_2,...,a+bx_N\}$, mientras que los valores reales son $\{y_1, y_2, ...,y_N\}$ Se tendrá entonces que:

$\sum(a+bx_i-y_i)^2$ ha de ser mínima. Sustituyendo el valor de y dado por la recta:

Esta expresión será mínima para aquellos valores de a y b que anulen la primera derivada, esto es:

$\sum 2(a+bx_i-y_i)=0$ derivando con respecto a "a"

$\sum 2x_i(a+bx_i-y_i)=0$ derivando con respecto a "b"

Desarrollando el sumatorio, tenemos:

$$2aN+2b\sum x_i-2\sum y_i=0\Rightarrow aN+b\sum x_i-\sum y_i=0$$
$$2a\sum x_i+2b\sum x_i^2-2\sum x_iy_i=0\Rightarrow a\sum x_i+b\sum x_i^2-\sum x_iy_i=0$$

Estas dos últimas igualdades forman el sistema de ecuaciones (con incógnitas "a" y "b"):

$$\left.\begin{aligned} aN+b\sum x_i&=\sum y_i\\ a\sum x_i+b\sum x_i^2&=\sum x_iy_i\end{aligned}\right\}$$ llamadas ecuaciones normales de la recta de regresión.

Multiplicando la primera $\sum x_i$ por y la segunda por N para resolver por reducción, tenemos:

$$\left.\begin{aligned} aN\sum x_i+b\left(\sum x_i\right)^2&=\sum x_i\sum y_i\\ aN\sum x_i+bN\sum x_i^2&=N\sum x_iy_i\end{aligned}\right\}$$

Y restando miembro a miembro:

$$b\left[\left(\sum x_i\right)^2 - N\sum x_i^2\right] = \sum x_i \sum y_i - N\sum x_i y_i \Rightarrow$$

$$\Rightarrow b = \frac{\sum x_i \sum y_i - N\sum x_i y_i}{\left[\left(\sum x_i\right)^2 - N\sum x_i^2\right]} = \frac{N^2\overline{xy} - N^2(s_{xy} + \overline{xy})}{N^2\overline{x}^2 - N^2(s_x^2 + \overline{x}^2)} = \frac{-s_{xy}}{-s_x^2} = \frac{s_{xy}}{s_x^2}$$

Pues, supuesta la frecuencia de cada valor de la variable igual a la unidad, se tiene que:

$$\sum x_i = N\overline{x}$$
$$\sum y_i = N\overline{y}$$
$$\sum x_i y_i = N(s_{xy} + \overline{xy})$$
$$\sum x_i^2 = N(s_x^2 + \overline{x}^2)$$

La recta de regresión, que tiene por pendiente b será, pues:

$y = a + \frac{s_{xy}}{s_x^2}x$ y determinamos "a" con la condición de que la recta pasa por $(\overline{x},\overline{y})$,o sea, el centro de gravedad ha de satisfacer la ecuación de la recta con lo que:

$$\overline{y} = a + \frac{s_{xy}}{s_x^2}\overline{x} \Rightarrow a = \overline{y} - \frac{s_{xy}}{s_x^2}\overline{x}$$

Adoptando finalmente la recta la ecuación:

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

Que es la **ecuación de la recta de regresión de Y sobre X**.
Intercambiando ahora X e Y podríamos haber obtenido otras recta de regresión llamada de X sobre Y que será de la siguiente forma:

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

A los valores de $\frac{s_{xy}}{s_x^2}$ y $\frac{s_{xy}}{s_y^2}$ se les llama coeficientes de regresión (no confundir con el coeficiente de correlación)

Observaciones:

- Hay que tener absolutamente clara la notación de cuál es la variable independiente (x) y cuál la variable dependiente (y), pues no son intercambiables en un problema concreto: una cosecha puede depender de la cantidad de lluvia caída pero la lluvia no depende en absoluto de la cosecha.
- La recta de regresión sólo sirve para predecir la variable dependiente. Es decir la de Y sobre X, estimamos Y a partir de X, y la de X sobre Y, predecimos X a partir de Y.
- Al usar la recta de regresión para predecir un resultado se comete un error que es mayor a medida que nos alejamos del valor medio (media) y a medida que el coeficiente de correlación se aleja del valor 1 ó -1.
- El coeficiente de regresión tiene el mismo signo siempre que el coeficiente de correlación, pero el hecho de que el primero sea más o menos grande no indica que la correlación sea más o menos fuerte.

Ejemplos

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

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

$$s_y^2 = \frac{\sum y_i^2}{N} - \bar{y}^2 = \frac{1113}{5} - 14{,}2^2 = 20{,}96\,; \; s_{xy} = \frac{\sum x_i y_i}{N} - \bar{x}.\bar{y} = \frac{409}{5} - 4{,}8.14{,}2 = 13{,}64$$

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

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

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

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 notas psicología	**Y notas estadística**	**Nº 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	$x_i f_i$	$y_i f_i$	$x_i^2.f_i$	$y_i^2.f_i$	$x_i y_i f_i$
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

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

$$s_{xy} = \frac{\sum f_i x_i y_i}{N} - \bar{x}.\bar{y} = \frac{1336}{40} - (5{,}3).(5{,}6) = 33{,}4 - 30{,}8 = 2{,}6$$

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

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

$y - 5{,}6 = \frac{2{,}6}{2{,}6}(x - 5{,}5)$, 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: $r = \frac{s_{xy}}{s_x.s_y}$

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

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

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

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.

3.- La siguiente tabla de doble entrada muestra las estaturas (x_i en cm.) y pesos (y_i en kg.) de 120 personas. Determinar el coeficiente de correlación, la recta de regresión de Y sobre X y estimar la estatura de una persona cuyo peso sea de 100 kg. Dibujar la nube de puntos y la citada recta de regresión:

x_i / y_i	60	62	73	90	F_{yi}	$y_i f_{yi}$	$y_i^2 f_{yi}$	$x_i y_j f_{ij}$
150	1	12			13	1950	292500	120600
171			20	10	30	5130	877230	403560
180	5	30	10	9	54	9720	1749600	666000
200			3	20	23	4600	920000	403800
f_{xi}	6	42	33	39	N=120	21400	3839330	1593960
$x_i f_{xi}$	360	2604	2409	3510	8883			
$x_i^2 f_{xi}$	21600	161448	175857	315900	674805			
$x_i y_i f_{ij}$	63000	446400	424860	659700	1593960			

La parte de la tabla doblemente recuadrada constituye los datos del problema. Las filas y columnas restantes se han añadido como cálculos posteriores para determinar los parámetros necesarios. Se tiene:

Medias:

$$\overline{x} = \frac{8883}{120} = 74 \text{ kg.} \qquad \overline{y} = \frac{21400}{120} = 178 \text{ cm.}$$

Desviaciones típicas:

$$s_x = \sqrt{\frac{674805}{120} - 74^2} = 12{,}1 \text{ kg.} \qquad s_y = \sqrt{\frac{3839330}{120} - 178^2} = 17{,}6 \text{ cm.}$$

Covarianza:

$$s_{xy} = \frac{1593969}{120} - 74 \cdot 178 = 111{,}1$$

Coeficiente de correlación de Pearson:

$$r = \frac{111{,}1}{12{,}1 \cdot 17{,}6} = 0{,}5$$ correlación directa (r>0) pero baja en magnitud.

Recta de regresión de Y sobre X:

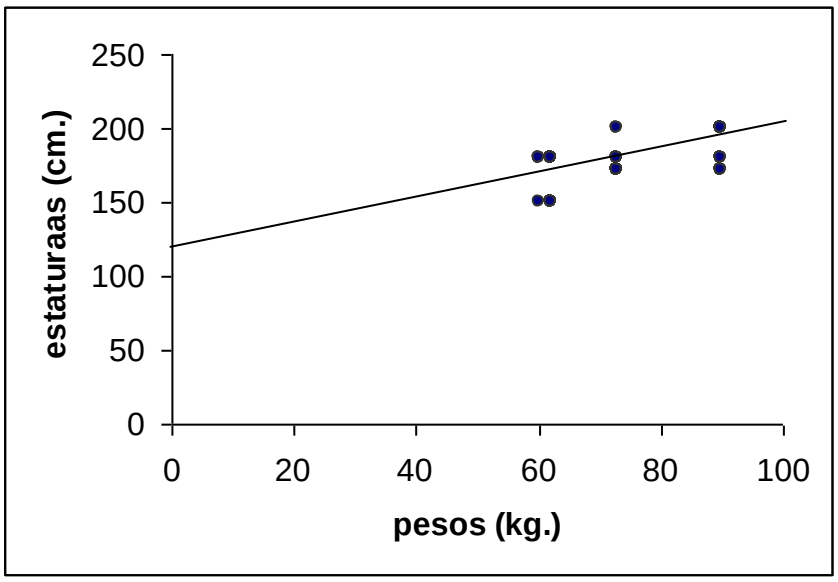
$$y - 178 = \frac{111{,}1}{12{,}1^2}(x - 74) \Rightarrow y = 0{,}76x + 121{,}76$$

Estimación para x= 100 kg. :

$$y = 0{,}76 \cdot 100 + 121{,}76 = 197{,}76 \text{ cm.}$$

Esta estimación es poco significativa dado el valor muy bajo obtenido para el coeficiente de correlación.

Nube de puntos:



Donde en cada punto de la nube hay superpuestos tantos puntos como indique la frecuencia absoluta conjunta de ambas variables. Contando la totalidad de puntos visibles y superpuestos deberían haber 120.

Ejercicios propuestos.

1 Las notas obtenidas por 10 alumnos en Matemáticas y en Música son:

Alumnos	Mat.	Mús.
1	6	6,5
2	4	4,5
3	8	7
4	5	5
5	3,5	4
6	7	8
7	5	7
8	10	10
9	5	6
10	4	5

a) Calcula la covarianza, las varianzas y el coeficiente de correlación.
b) ¿Existe correlación entre las dos variables?
c) Calcula la recta de regresión. ¿Cuál será la nota esperada en Música para un alumno que hubiese obtenido un 8,3 en Matemáticas?

(Sol. 3,075; 3,76; 2,96; 0,92; y = 1,6 + 0,817x; 8,38)

2.- Cinco niñas de 2, 3, 5, 7 y 8 años de edad pesan respectivamente 14, 20, 30, 42 y 44 Kg. Halla la ecuación de la recta de regresión de la edad sobre el peso. ¿Cuál sería el peso aproximado de una niña de 6 años?.
(Sol. x = 0,192y-0,76; 35,2 Kg.)

3.- La tabla adjunta da el índice de mortalidad de una muestra de población en función del consumo diario de cigarrillos:

Número de cigarrillos	*x*	*3*	*5*	*6*	*15*	*20*
Índice de mortalidad	*y*	*0,2*	*0,3*	*0,4*	*0,5*	*0,7*

a) Determina el coeficiente de correlación e interpreta el resultado.
b) Halla la recta de regresión de y sobre x
c) ¿Cuál será el índice de mortalidad para un consumidor de 40 cigarrillos diarios?

4 Dadas las siguientes variables estadísticas bidimensionales, razona si entre ellas existe una correlación positiva, negativa o nula.

a) La clasificación de un equipo de fútbol en la liga y la estatura media de sus jugadores.
b) Las temperaturas diarias de una ciudad durante el mes de febrero y el consumo de energía eléctrica por habitante.
c) El nº de coches por habitante y el nº de accidentes de tráfico en una ciudad.
d) Las notas de Matemáticas y de Física y Química de un grupo de alumnos.

5 Para cada una de las variables bidimensionales siguientes, se ha hecho un estudio para investigar la correlación existente entre los datos recogidos. Los coeficientes de correlación obtenidos han sido: $r_1 = 0,9$, $r_2 = -0,83$, $r_3 = 1$, $r_4 = 0,6$ y $r_5 = 0$.
Asigna a cada par de variables el correspondiente coeficiente:

a) Horas diarias que ve la televisión un alumno y asignaturas aprobadas en una evaluación.
b) Peso de un recién nacido y color de sus ojos.
c) Número de partidos ganados y número de canastas conseguidas por un equipo de baloncesto.
d) Nota final de Matemáticas y nota final de Lengua en 3.º de ESO.
e) Espacio recorrido por un coche en un tiempo determinado y velocidad del mismo en dicho tiempo.

6 En una ciudad se han celebrado en una semana 8 matrimonios. Las edades de los novios están reflejadas la tabla.

Edad del novio (x_i)	23	25	26	27	28	28	29	30
Edad de la novia (y_i)	23	22	25	24	26	27	26	27

a) Representa la nube de puntos y calcula las medias aritméticas de las edades de los novios ($\bar{x}$) y de las novias ($\bar{y}$).

b) ¿Cuál es el coeficiente de correlación?

c) Escribe la ecuación de la recta de regresión.

d) ¿Qué edad cabe esperar para una novia cuando el novio tenga 34 años?

7 En una empresa seleccionan 6 trabajadores, y se anotan sus años de servicio y el tiempo de servicio en horas solicitado el último mes. Los resultados son:

X: años en la empresa	1	3	2	4	5	4
Y: nª de horas de permiso	1	1	3	4	6	5

a) Representar gráficamente los datos anteriores. Sin hacer ningún cálculo razonar si los datos muestran correlación positiva o negativa.
b) Calcular la covarianza e el coeficiente de correlación entre X e Y.
c) Calcular la recta de regresión de Y sobre X. Explicar su utilidad.

8 Se ha realizado un estudio sobre las preferencias de las ratas con respecto a la temperatura del agua. Un grupo de 10 ratas fue sometido a dos temperaturas del agua diferentes y se midió el tiempo de permanencia en este medio. Estos fueron los resultados:

26ºC	6	7	8	5	9	3	6	4	7	2
30ºC	1	8	3	10	4	7	9	2	4	6

Calcular la covarianza y el coeficiente de correlación. Analizar que tipo de dependencia existe entre las variables.

9 La media de los pesos de una población es 65 Kg. y la de las estaturas 170 cm, mientras que las desviaciones típicas son respectivamente 5 Kg y 10 cm, siendo la covarianza entre ambas variables 40. Calcular la recta de regresión de los pesos respecto de las estaturas. ¿Cuánto estima que pesará un individuo de 180 cm de estatura?

10 La siguiente tabla ofrece los resultados de 6 pares de observaciones, realizadas para analizar el grado de relación existente entre 2 variables X e Y.

X	2	2	3	3	3	4
Y	0	1	1	2	4	3

Obtener:
a) Recta de regresión de Y sobre X.
b) Representación gráfica de la misma, así como de los pares de observaciones anteriores.
c) ¿Qué grado de relación lineal existe entre ambas variables?

11 Calcula la recta de regresión correspondiente a la distribución siguiente:

Altura sobre el nivel del mar	0	184	231	481	730	911	1550
Presión atmosférica	760	745	740	720	700	685	650

¿Qué presión atmosférica habría sobre una aldea que está a 2600 metros de altitud?

12 A dos grupos de 8 profesores de letras (grupo A) y de ciencias (grupo B), se les ha planteado un test de cultura general de 100 preguntas, arrojando el siguiente número de contestaciones acertadas:

Grupo A	46	48	49	50	50	51	52	54
Grupo B	10	18	30	50	50	70	82	90

Halla para cada uno de los grupos la media, moda y mediana, así como la desviación típica. Interpreta los resultados.

13 Para estudiar algunos efectos de la altitud, un grupo de 10 jóvenes aficionados a la investigación científica ha llevado a cabo un experimento. Cada uno de ellos a acudido a un lugar distinto de la misma comarca y ha obtenido medidas sobre:

- Altura en metros sobre el nivel del mar.
- Números de plantas de una cierta especie en 1 dam^2.
- Presión atmosférica en mm de Hg.
- Número de pulsaciones por minuto del experimentador.

Estos son los resultados:

a	Altura	0	184	231	481	730	911	1343	1550	1820	2184
n	Nºde Plantas	0	0	4	14	23	18	12	3	0	0
Pa	Presión atmosférica	760	745	740	720	700	685	650	630	610	580
Pu	Pulsaciones	73	78	75	78	83	80	89	80	85	92

Considera las siguientes distribuciones bidimensionales: I) a,n; II) a, Pa; III) a,Pu. Representa cada una de ellas en un diagrama cartesiano mediante la nube de puntos, traza a ojo su recta de regresión y estima su coeficiente de correlación. Efectúa después, los cálculos de forma rigurosa y compara los resultados. Estima, sobre la correspondiente recta de regresión la presión atmosférica correspondiente a una altura de 2000 m.

14 De la siguiente distribución bidimensional:

Y / X	[0,1)	[1,2)	[2,3)	[3,4)	[4,5)	[5,6)	[6,7)
[1,2)	2	2	1	-	-	-	-
[2,3)	-	1	2	4	1	1	-
[3,4)	-	-	-	-	1	2	3

Obtener:

a) Recta de regresión de Y sobre X.

b) Coeficiente de correlación lineal e interpretar el resultado.

15 Un sociólogo afirma que las mujeres se casan más jóvenes que los hombres. Para apoyar dicha afirmación presenta la siguiente tabla obtenida en una encuesta a 50 parejas, donde H representa la edad de los hombres y M la de las mujeres

M H	[15,20)	[20,25)	[25,30)	[30,35)
[15,20)	**1**	**4**	**3**	**0**
[20,25)	**1**	**5**	**7**	**3**
[25,30)	**0**	**2**	**8**	**8**
[30,35)	**0**	**0**	**2**	**6**

a) ¿Es correcta la afirmación del psicólogo? Razona la respuesta
b) ¿Qué edad se puede esperar para un hombre casado con una mujer de 25 años?
c) Estudia la fiabilidad de la predicción del apartado anterior.