Tema 7 : DATOS BIVARIADOS. CORRELACION Y REGRESION.

Distribuciones uni- y pluridimensionales.

Hasta ahora se han estudiado los índices y representaciones de una sola variable por individuo. Son las distribuciones unidimensionales o univariadas .

En un individuo se pueden estudiar conjuntamente dos o más variables con objeto de ver si hay relación o dependencia entre ellas. Tenemos entonces distribuciones pluridimensionales, también llamadas plurivariadas. Cuando son dos se llaman bivariadas o bidimensionales. Son las únicas que veremos nosotros.

La simple medida de más de una variable en un individuo no tiene categoría de pluridimensional, sólo se tiene una serie de variables unidimensionales. ¡Hace faltar estudiarlas conjuntamente!

Estudio de variables bidimensionales

A una de las variables se la llama variable independiente y se representa por X. A la otra se la denomina variable dependiente y su símbolo es Y. (también se usan las minúsculas: x e y).

Los datos deben de ir siempre apareados. Para cada individuo se dan su X y su Y. ("Cada oveja con su pareja"). El nº de individuos se representa por N.

N es el nº de individuos, no el nº de datos, que siempre será el doble de N, pues cada individuo nos proporciona dos. ¡Es un error observado con frecuencia en los exámenes!

Ambas variables pueden ser cuantitativas (CT) o cualitativas (CL). En este tema veremos el caso de que ambas variables sean CT (que se completará en el tema 18) . En el tema 16 veremos la relación entre dos variables CL, expresada mediante la Odds ratio (OR). El caso de una variable CL y otra CT se trata en el tema 17.

--Ejemplos de variables bidimensionales

talla y peso, edad y tensión arterial, frecuencia cardiaca y frecuencia respiratoria, sexo y hábito de fumar, sexo y peso al nacer, velocidad de un vehículo y distancia de frenada...

Cuando ambas variables son CT, se pueden presentar:

a) cada variable por separado (con sus tablas, gráficos e índices)
b) conjuntamente (objeto de este tema) mediante:
    a. la tabulación y representación gráfica de los datos
    b. el cálculo de dos índices:
        i. coeficiente de correlación
        ii. ecuación de regresión

Tabulación

---de los datos originales

se hace una tabla, vertical u horizontal, con una columna (o fila) para X y otra para Y. Es opcional añadir otra para el número de orden del individuo. Los datos se ordenan en función del orden de los individuos o de los valores de X o de los valores de Y o no se ordenan en absoluto.

Ejemplo: Para X = (1 , 1 , 3 , 6 , 2 , 3 , 5 , 6) e Y = (1 , 1 , 4 , 4 , 2 , 5 , 1 , 5) :

Indiv.	**X**	**Y**
1	**1**	**1**
2	**1**	**1**
3	**3**	**4**
4	**6**	**4**
5	**2**	**2**
6	**3**	**5**
7	**5**	**1**
8	**6**	**5**

o

Ind.	1	2	3	4	5	6	7	8
X	**1**	**1**	**3**	**6**	**2**	**3**	**5**	**6**
Y	**1**	**1**	**4**	**4**	**2**	**5**	**1**	**5**

---de los datos agrupados en clases

Los valores de X e Y se agrupan en clases, siguiendo el método visto en el tema 4. La tabla es bidimensional: en la primera columna se representan las clases de X y en la primera fila las clases de Y. Al hacer el recuento los valores de cada individuo quedarán dentro de la casilla de la tabla que englobe a ambos.

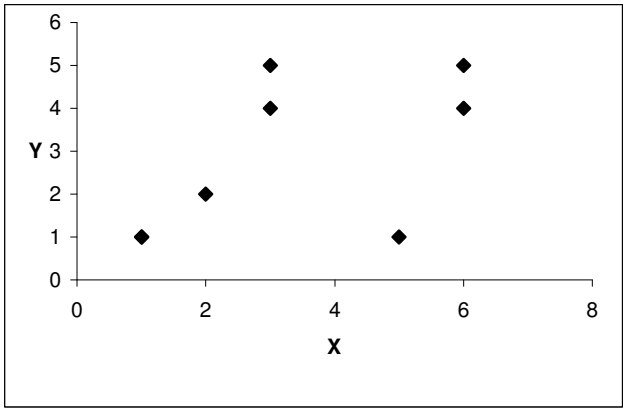
Ejemplo: Para los datos ya vistos la tabla podría ser así (presentada de forma simplificada y no del todo ortodoxa para mayor claridad):

Y / X	1-2	3-4	5-6	TOTAL
1-2	3	0	0	3
3-4	0	1	1	2
5-6	1	1	1	3
TOTAL	4	2	2	8

Gráficos

--datos originales, aislados

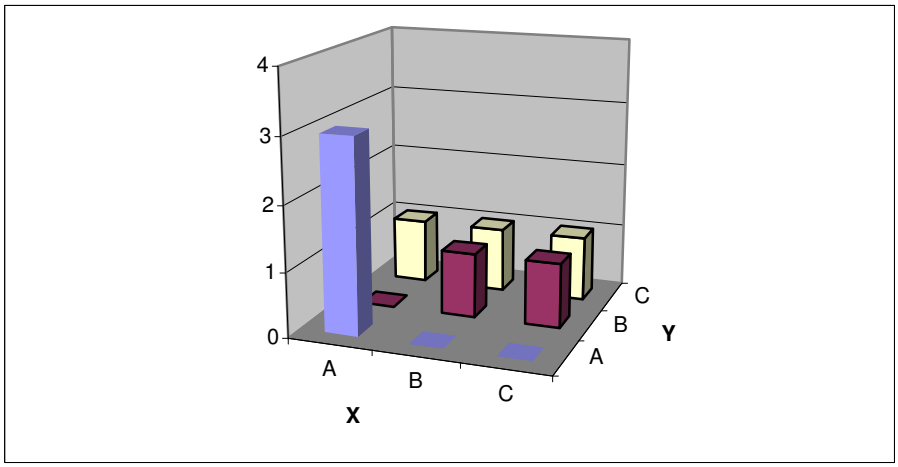
Es el diagrama de puntos, también llamado de dispersión o de nube de puntos. Los valores de cada individuo llevados aun eje de coordenadas originan un punto.



---datos agrupados en clases

El gráfico es el Estéreograma. Cada casilla de la tabla (que es la conjunción de dos clases, una de X y otra de Y) está representada por un prisma o cilindro (o incluso por una línea) cuya altura es proporcional a la frecuencia.

Para mayor claridad las clases en vez de como 1-2, 3-4 y 5-6 se representan como A, B y C



Índices estadísticos

Los típicos de estas distribuciones, aparte de los de cada variable por separado, son el coeficiente de correlación y la ecuación de regresión. Son los llamados índices o parámetros de asociación. Son distintos en función del tipo de variables (CL-CL, CL-CT, CT-CT). en este tema sólo nos ocuparemos del caso en que ambas variables son CT.

Correlación significa relación mutua y expresa el grado de asociación existente entre las variables, el CUANTO de la relación. Su parámetro es el coeficiente de correlación. Su símbolo es **r**, que puede acompañarse, si la claridad lo exige, de un subíndice con la notación de las variables (p.e. r_{xy}). Se puede calcular la correlación entre dos variables o más (correlación múltiple).

La **regresión** es la forma, el COMO de esa asociación. Expresa la relación entre las dos variables, X e Y, mediante la ecuación de regresión y su representación gráfica la línea de regresión. Mediante ella conocida una variable es posible predecir la otra. Por consenso X es la variable independiente e Y la dependiente. De esta forma $Y = f(X)$.

Coeficiente de correlación

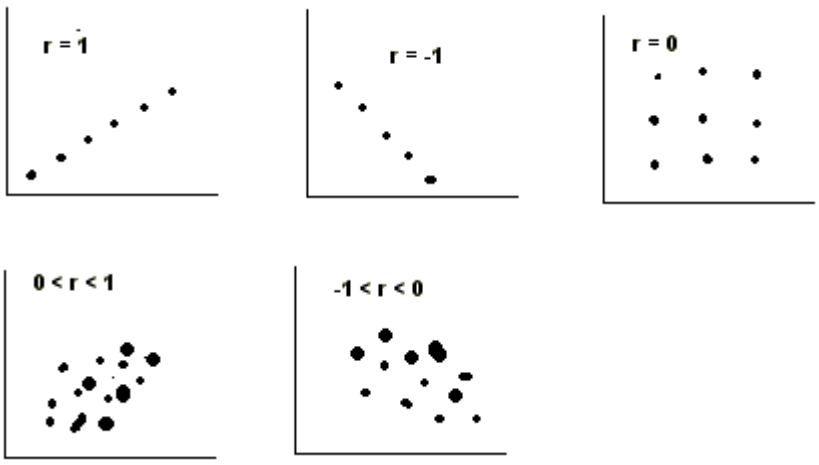
Mide la intensidad de la asociación entre las variables. Es un número abstracto, independiente de la unidad de medida de las variables. Puede adoptar cualquier valor entre –1 y 1. Dicho de otra forma: $r = \in (-1 \div 1)$. Suele expresarse con 3 decimales, a no ser que valga –1, 0 ó 1. Aparte de su valor descriptivo sirve para ver la significación estadística de la relación (tema 18)

Aquí veremos sólo la correlación entre dos variables. Su coeficiente de correlación se llama de Pearson, aunque cuando se dice simplemente coeficiente de correlación, se sobreentiende que es éste. En el tema 18 se verá otro coeficiente, el de Spearman, que se usa cuando no puede utilizarse el de Pearson.

Si se observa una correlación aparentemente alta entre X e Y puede tratarse de dos situaciones:
--una variación de X provoca otra en Y. Por ejemplo, el aumento de la temperatura corporal produce un aumento de la frecuencia cardiaca.
--X e Y varían a la par por efecto de un a tercera o más variables. La correlación existente es pura coincidencia. Son las llamadas correlaciones espurias, ya citadas en el tema 1. Son las más frecuentes. De forma automática correlación ≠ causalidad. Se requiere un estudio experimental con resultado significativo.

Si r = 1 hay una correlación total (perfecta) positiva.
Si r = -1 hay una correlación total (perfecta) negativa.
Si r = 0 no hay correlación.
Si está entre –1 y 0 , la correlación es parcial y negativa.
Si está entre 0 y 1, la correlación es parcial y positiva.
Una r de 0, -1 ó 1 apenas se encuentra en la práctica

Gráficamente esto se puede representar así:



Cálculo de coeficiente de correlación

Veremos únicamente el cálculo a partir de los datos originales, aislados.

$$r = \frac{N\sum XY - \sum X \sum Y}{\sqrt{\left[N\sum X^2 - \left(\sum X\right)^2\right]\left[N\sum Y^2 - \left(\sum Y\right)^2\right]}}$$

Para este cálculo y el de la ecuación de regresión es de gran ayuda construirse una tabla auxiliar como la que se utiliza en el siguiente ejemplo:

X = (2 , 1 , 3 , 2 , 5) ; Y = (3 , 5 , 4 , 2 , 6)

X	**Y**	**X^2**	**Y^2**	**XY**
2	3	4	9	6
1	5	1	25	5
3	4	9	16	12
2	2	4	4	4
5	6	25	36	30
13	*20*	*43*	*90*	*57*

$$\mathbf{r = \frac{(5 * 57) - (13 * 20)}{\sqrt{[(5 * 43) - 13^2][(5 * 90) - 20^2]}} =}$$

$$\mathbf{= \frac{25}{\sqrt{46 * 50}} = 0,521}$$

Este valor de r es el valor puntual. Cada día se utiliza más el valor por intervalo, cuyo cálculo veremos en el tema 13, en el que se estudian los intervalos de confianza (IC).

Regresión

Ya hemos visto el concepto de regresión. La fórmula matemática que la expresa puede ser una ecuación de primer grado (regresión lineal: y = a+bx) u otras ecuaciones más complejas (cuadrática: $y=ax^2+bx+c$; exponencial: $y=ae^{bx}$; potencial: $y=ax^b$; hiperbólica: y=a(b/x) ; logarítmica: $y=a+bl_nx$; etc...), que no trataremos, pues son muy complejas. Nos limitaremos a la regresión lineal, también llamada recta de regresión, pues su representación gráfica es una línea recta, que representa lo mejor posible a todos los puntos del diagrama de dispersión. Realmente se podrían trazar muchas rectas de regresión, pero sólo nos interesa la llamada "mejor línea de ajuste", que es la que corresponde a la ecuación y=a+bx (ó y=bx+a; el orden de los sumandos no altera la suma).

En esta fórmula **b** es el coeficiente de regresión, también llamado pendiente, pues de él depende la inclinación de la recta y nos indica en cuanto se modifica **y** en media cuando **x** varía en una unidad.

a es el valor de y cuando x = 0 , por lo que también se la llama ordenada en el origen o intersección de y . Se ha comprobado que la mejor línea de ajuste es aquella en que la suma de los cuadrados de las diferencias entre cada punto original y la línea de regresión es la menor de todas las posibles. Por eso a este método se le llama "de los mínimos cuadrados". Afortunadamente no hay que calcularlos, pues se ha desarrollado una fórmula mucho más manejable para encontrar la ecuación.
En principio se considera a y variable dependiente y a x variable independiente, por lo que la regresión se dice que es de y sobre x. En este sentido b es realmente b_{yx} y así se entiende cuando no hay subíndice. Matemáticamente también se puede calcular la regresión de x sobre y. Si interesara este cálculo, lo que no es habitual, escribiríamos b_{xy} para evitar confusiones.

Cálculo
Seguiremos el procedimiento que calcula primero **b** y a partir de él calcula **a**

$$b = \frac{N\sum XY - \sum X \sum Y}{N\sum X^2 - \left(\sum X\right)^2} \qquad a = \overline{Y} - b\overline{X}$$

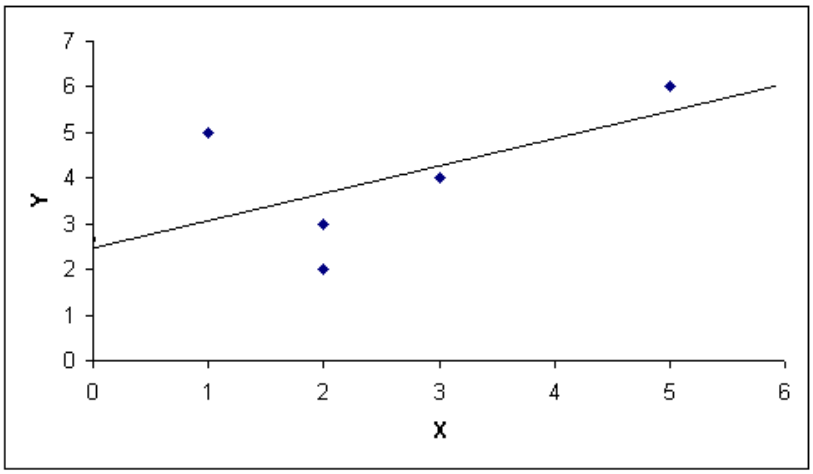
Ejemplo: Utilizando los datos empleados para calcular el coeficiente de correlación:

$$\bar{X} = \frac{13}{5} = 2,6 \qquad \bar{Y} = \frac{20}{5} = 4 \qquad b = \frac{(5*57) - (13*20)}{(5*43) - 13^2} = \frac{25}{46} = 0,54347$$

$$a = 4 - (0,54347 * 2,6) = 2,587$$

por tanto la ecuación es $y = 2,587 + 0,543x$

Representación gráfica
Para trazar una recta basta con dos puntos. En el diagrama de dispersión se busca el valor de y para x = 0. El otro punto se obtiene a partir de un valor cualquiera de **x** que nos de una **y** que no se salga del gráfico. En nuestro ejemplo: si x = 0 , y = 2,587 ; para x = 5 , y = 5,302



Se suele incluir en el gráfico la ecuación y el coeficiente de correlación y con menos frecuencia el IC (intervalo de confianza) de forma numérica y/o con dos rectas más que lo delimiten.

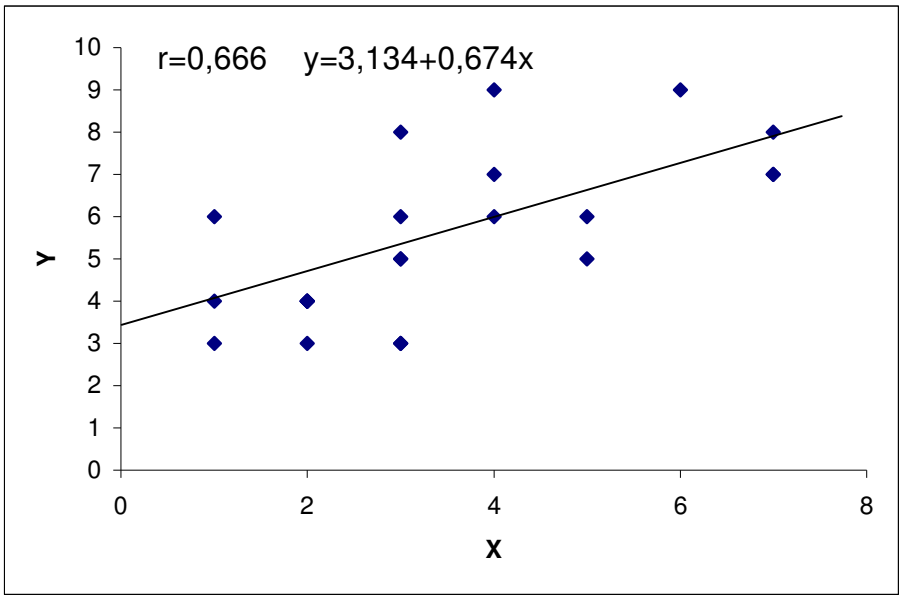
Coeficiente de determinación
Mide cuantitativamente la bondad o representatividad del ajuste de la recta a la nube de puntos. Es el cuadrado de r. Su símbolo es $\mathbf{r^2}$ o R. En nuestro ejemplo $r^2 = 0,302$. Cuando se calculan diversas ecuaciones de regresión (lineal, exponencial, logarítmica, etc.) la que tenga el r^2 más alto será la mejor, la más representativa. $\mathbf{r^2}$ unifica la fuerza de la asociación de positivos y negativos. (una r = -0,400 es más potente que una r = 0,350 ; sus $\mathbf{r^2}$ son 0,160 y 0,122)

Ejercicio resuelto con Excel.
Ejercítese en el cálculo de la media, desviación estándar, CV, coeficiente de correlación y ecuación de regresión.

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

				X	Y
Error est. Y	1,472	**media**		3,478	5,478
r	0,666	**s**		1,904	1,928
Ecuación: b	0,674	**CV**		54,7	35,2
Ecuación: a	3,134	**p50 ó M**		3,000	5,000

N = 23



Notas adicionales. 1) Con los datos del ejercicio anterior se han calculado otras ecuaciones de regresión con sus respectivos r y r^2. Se dan aquí a título puramente informativo para que se vea que la mejor ecuación que relaciona a X e Y es la cuadrática, ya que tiene la r^2 más alta.

ECUACION	a	b	c	r	r^2
Cuadrática	**-0,034**	**0,950**	**2,703**	**0,668**	**0,447**
Lineal	**3,134**	**0,674**		**0,666**	**0,443**
Exponencial	**3,334**	**0,125**		**0,659**	**0,434**
Logarítmica	**3,262**	**2,034**		**0,630**	**0,397**
Potencial	**3,412**	**0,378**		**0,625**	**0,390**

2) aunque no es lo correcto, en la práctica se calcula en ocasiones **r** cuando se contrastan 2 Vbles. CT procedentes de individuos distintos, siempre que estén emparejados. Aquí N es el nº de parejas de datos, no el de individuos.