

TEMA 5

DISTRIBUCIÓN DE FRECUENCIAS BIDIMENSIONALES.

En las distribuciones bidimensionales entran en juego dos variables o dos atributos. Si hablamos de variables cuantitativas las tablas de frecuencias donde se reúnen los datos se denominan tablas de correlación. Si por el contrario estamos ante variables cualitativas o atributos se denominan tablas de contingencia.

x_i	y_j	1	2	3	$n_{i.}$
0	3	0	1		4
3	0	4	2		6
5	1	1	6		8
$n_{.j}$		4	5	9	$N = 18$

Para trabajar con distribuciones bidimensionales, debemos calcular las distribuciones marginales.

x_i	0	3	5
$n_{i.}$	4	6	8

y_j	1	2	3
$n_{.j}$	4	5	9

$$\bar{x} = \frac{\sum x_i n_{i.}}{N} \quad \bar{y} = \frac{\sum y_j n_{.j}}{N}$$

$$\sigma_x^2 = \frac{\sum (x_i - \bar{x})^2 n_{i.}}{N} \quad \sigma_y^2 = \frac{\sum (y_j - \bar{y})^2 n_{.j}}{N}$$

MOMENTOS CON RESPECTO AL ORIGEN.

$$a_{rs} = \frac{n_{ij}}{N} \sum \sum x_i^r y_j^s$$

$$a_{10} = \frac{\sum x_i n_{i.}}{N} \quad a_{01} = \frac{\sum y_j n_{.j}}{N} \quad a_{20} = \frac{\sum x_i^2 n_{i.}}{N} \quad a_{02} = \frac{\sum y_j^2 n_{.j}}{N}$$

$$a_{11} = \frac{\sum \sum x_i y_j n_{ij}}{N}$$

MOMENTOS CON RESPECTO A LA MEDIA.

$$m_{rs} = \sum \sum (x_i - \bar{x})^r (y_j - \bar{y})^s \frac{n_{ij}}{N}$$

Los momentos con respecto a la media pueden ponerse en relación con los momentos con respecto al origen.

$$m_{20} = \sigma_x^2 = a_{20} - (a_{10})^2$$

$$m_{02} = \sigma_y^2 = a_{02} - (a_{01})^2$$

$$m_{11} = Cov(xy) = a_{11} - a_{10} \cdot a_{01}$$

1. Con la tabla anterior, calcular las medias ,desviaciones típicas marginales y la covarianza.

$$\bar{x} = a_{10} = \frac{\sum_1^3 x_i n_i}{N} = \frac{0 \cdot 4 + 3 \cdot 6 + 5 \cdot 8}{18} = \frac{58}{18} = \frac{29}{9}$$

$$\bar{y} = a_{01} = \frac{\sum y_j n_j}{N} = \frac{1 \cdot 4 + 2 \cdot 5 + 3 \cdot 9}{18} = \frac{41}{18}$$

$$s_x^2 = m_{20} = a_{20} - (a_{10})^2 = \frac{\sum x_i^2 n_i}{N} - \left(\frac{\sum x_i n_i}{N} \right)^2 = \frac{3^2 \cdot 6 + 5^2 \cdot 8}{18} - \left(\frac{29}{9} \right)^2 = 3,73$$

$$s_x = \sqrt{3,73} = 1,9132$$

$$s_y^2 = m_{02} = a_{02} - (a_{01})^2 = \frac{\sum y_j^2 n_j}{N} - \left(\frac{\sum y_j n_j}{N} \right)^2 = \frac{1^2 \cdot 4 + 2^2 \cdot 5 + 3^2 \cdot 9}{18} - \left(\frac{41}{18} \right)^2 =$$

$$= 0,422939$$

$$s_y = \sqrt{0,422939} = 0,65033$$

$$Cov(xy) = m_{11} = a_{11} - a_{10} \cdot a_{01} = \frac{\sum x_i y_j n_{ij}}{N} - (\bar{x} \cdot \bar{y}) = \frac{0 \cdot 1 \cdot 3 + 0 \cdot 2 \cdot 0 + 0 \cdot 3 \cdot 1 +$$

$$+ 3 \cdot 1 \cdot 0 + 3 \cdot 2 \cdot 4 + 3 \cdot 3 \cdot 2 + 5 \cdot 1 \cdot 1 + 5 \cdot 2 \cdot 1 + 5 \cdot 3 \cdot 6}{18} - \left(\frac{29}{9} \cdot \frac{41}{18} \right) = \frac{147}{18} - \frac{1189}{162} =$$

$$= 0,82716$$

DEPENDENCIA ESTADÍSTICA ENTRE DOS O MÁS VARIABLES

La teoría de la correlación estudia la dependencia estadística entre variables, esta relación puede adoptar tres resultados:

- No existe dependencia entre las variables por lo tanto se dice que hay independencia funcional o correlación nula.
- Cuando existe una función tal que a cada valor de la variable X le corresponde uno sólo valor de Y, y viceversa, se dice que hay dependencia funcional o correlación funcional.
- Cuando existe algún grado de relación entre las variables diremos que hay dependencia estadística parcial. Estas relaciones pueden ser positivas o negativas. A través de la teoría de la regresión decidiremos qué tipo de función explica mejor la dependencia.

Correlación o grado de dependencia lineal entre dos variables.

Se utiliza el coeficiente de correlación de Pearson $r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{m_{11}}{\sqrt{m_{20} \cdot m_{02}}}$

Su campo de variación es : $-1 \leq r_{xy} \leq 1$

$r_{xy} = 1$ relación lineal perfecta positiva o directa entre las variables.

$r_{xy} = -1$ relación lineal perfecta negativa o inversa entre las variables.

$-1 < r_{xy} < 1$ existe dependencia estadística entre las variables:

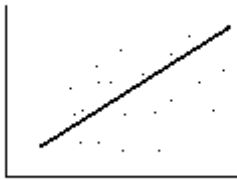
REGRESIÓN ENTRE DOS O MÁS VARIABLES.

Para poder realizar predicciones sobre los valores que adoptará un variable, es necesario ajustar una ecuación (recta, parábola, logarítmica, exponencial, etc) a la nube de puntos proveniente de la tabla de correlación.

Si la ecuación fuese una recta $Y = a + bX$ siendo Y la variable dependiente o endógena y X la variable independiente o exógena.

Para realizar el ajuste entre la función y la nube de puntos que representa los datos de la tabla de correlación, utilizaremos el ajuste por mínimos cuadrados.

Se introduce el concepto de error e_i en la predicción, siendo necesario que el error sea mínimo: $e_i = Y_i - Y_i^*$ $Y_i = Y_i^* + e_i$



El ajuste por mínimos cuadrados consiste en minimizar la suma al cuadrado de los errores

$$S = \sum e_i^2 \quad S = \sum (Y_i - Y_i^*)^2 = \sum (Y_i - a - bX_i)^2$$

$$\frac{\partial S}{\partial a} = 0 \quad \frac{\partial S}{\partial b} = 0$$

$$\frac{\partial S}{\partial a} = 2 \sum (Y_i - a - bX_i)(-1) = 0$$

$$-2(\sum Y_i - \sum a - b \sum X_i) = 0 \Rightarrow (\sum Y_i - \sum a - b \sum X_i) = 0$$

$$\sum Y_i = Na + b \sum X_i$$

$$\frac{\partial S}{\partial b} = 2 \sum (Y_i - a - bX_i)(-X_i) = 0$$

$$-2(\sum Y_i X_i - a \sum X_i - b \sum X_i^2) = 0 \Rightarrow (\sum Y_i X_i - a \sum X_i - b \sum X_i^2) = 0$$

$$\sum Y_i X_i = a \sum X_i + b \sum X_i^2$$

Obtenemos un sistema de ecuaciones normales:

$$\sum Y_i = Na + b \sum X_i$$

$$\sum Y_i X_i = a \sum X_i + b \sum X_i^2$$

Si dividimos las ecuaciones por la frecuencia total N, obtenemos:

$$\frac{\sum Y_i}{N} = \frac{Na}{N} + b \frac{\sum X_i}{N} \rightarrow \bar{Y} = a + b\bar{X} \rightarrow a = \bar{Y} - b\bar{X} \rightarrow a = a_{01} - ba_{10}$$

$$\frac{\sum Y_i X_i}{N} = a \frac{\sum X_i}{N} + b \frac{\sum X_i^2}{N} \rightarrow a_{11} = (a_{01} - ba_{10})a_{10} + ba_{20} \rightarrow$$

$$\rightarrow a_{11} = a_{10} \cdot a_{01} - ba_{10}^2 + ba_{20} \rightarrow a_{11} - a_{10}a_{01} = b(a_{20} - a_{10}^2) \rightarrow$$

$$\rightarrow b = \frac{a_{11} - a_{10}a_{01}}{a_{20} - a_{10}^2} = \frac{m_{11}}{m_{20}} = \frac{s_{xy}}{s_x^2}$$

La recta de regresión de Y sobre X $(Y - \bar{Y}) = b(X - \bar{X}) \Rightarrow (Y - a_{01}) = \frac{m_{11}}{m_{20}}(X - a_{10})$

Para la recta de regresión de X sobre Y.

En este caso la variable dependiente o endógena será la X y la variable independiente o exógena será la Y. Las operaciones son idénticas pero definiendo la regresión sobre $X_i = a' + b'Y_i$

$$b' = \frac{m_{11}}{m_{02}} \quad a' = a_{10} - ba_{01} \quad (X - \bar{X}) = b(Y - \bar{Y}) \rightarrow (X - a_{10}) = \frac{m_{11}}{m_{02}}(Y - a_{01})$$

BONDAD DEL AJUSTE Y PREDICCIONES.

Al afectar una regresión es necesario saber hasta que grado es posible sustituir la función estimada a los datos de la que se obtuvo, y el grado de dependencia entre las variables, para ello calculamos la varianza residual y el coeficiente de determinación.

- a) Varianza residual: $S_e^2 = \frac{\sum e_i^2}{N} = \frac{\sum (Y_i - Y_i^*)^2}{N} = m_{02} - bm_{11}$ si la medida es alta significa que la función estimada se aleja bastante de los valores originales (los residuos son grandes), si la medida es baja, entonces se dice que la regresión es bastante representativa.
- b) Coeficiente de determinación. Es el grado de participación de la varianza explicada por la regresión en la varianza total de la variable observada Y.

$$\text{Partimos de la igualdad } S_y^2 = S_{y^*}^2 + S_e^2$$

$S_{y^*}^2$ es la varianza explicada por la regresión

S_e^2 contiene la variabilidad que no es explicada por el modelo lineal.

$$R^2 = \frac{S_{y^*}^2}{S_y^2} = \frac{S_y^2 - S_e^2}{S_y^2} = 1 - \frac{S_e^2}{S_y^2} = \frac{m_{11}^2}{m_{20}m_{02}} \quad 0 \leq R^2 \leq 1$$

Si R^2 es cero existe una representatividad nula.

Si R^2 es uno significa que el ajuste es perfecto entre la nube de puntos y la ecuación estimada.

Los valores intermedios nos indican mayor o menor representatividad. La regresión se entiende representativa cuando $R^2 \geq 0.75$

EJERCICIOS.

- ♦ Una compañía quiere realizar un estudio sobre la influencia del gasto en I+D sobre las ventas. Para ello dispone de los siguientes datos sobre los últimos años:

Años	Gastos millones	Ventas millones
1998	3,0	130
1999	3,3	155
2000	3,8	175

2001	4,2	210
------	-----	-----

- a) Obtenga un modelo lineal que permita predecir las ventas a partir de los gastos en I+D. Comente los resultados.
 b) Prediga las ventas del 2002 sabiendo que el gasto en I+D será de 4,5 millones.
 c) Juzgue la bondad del modelo estimado.

Llamaremos x a los Gastos e y a las Ventas, por lo tanto ajustaremos una recta de Y sobre X .

x (Gastos)	y (Ventas)	x^2	y^2	xy
3	130	9	16900	390
3.3	155	10.89	24025	511.5
3.8	175	14.44	30625	665
4.2	210	17.64	44100	882
14.3	670	51.97	115650	2448.5

$$a_{10} = \frac{14.3}{5} = 2,86 \quad a_{01} = 134 \quad a_{11} = 489,7 \quad a_{20} = 10,394$$

$$a_{02} = 23.130 \quad m_{20} = 2,2144 \quad m_{02} = 5174 \quad m_{11} = 106,46$$

$$b = \frac{106,46}{2,2144} = 48,076$$

el coeficiente de regresión nos da la medida en que aumentarán las ventas al aumentar en un millón los gastos en I+D

$$y - 134 = 48,076(x - 2,86) \Rightarrow y = 48,076x - 3,49736$$

Las ventas para 2002 serán: $y_{2002} = 48,076(4,5) - 3,49736 = 212,84 \text{ millones}$

$$R^2 = 1 - \frac{S_{ry}^2}{S_y^2} = \frac{m_{11}^2}{m_{20}m_{02}} = \frac{(106,46)^2}{(2,2144 \cdot 5174)} = 0,9892 \quad \text{El } 98,92\% \text{ de la varianza de } y \text{ está}$$

explicada por x , a través de la función ajustada y_{ti} . Podemos decir que el modelo lineal se acepta.

- ♦ En Mercamadrid se ha observado durante un periodo de tiempo, las cantidades de Kg vendidos de un producto y el precio correspondiente en euros, obteniéndose los

$$\text{siguientes resultados } \sum_1^6 y = 21 \quad \sum_1^6 x = 84 \quad \sum_1^6 xy = 273 \quad \sum_1^6 x^2 = 1202 \quad \sum_1^6 y^2 = 91$$

Calcular: 1- La recta de regresión de Y sobre X . 2- La varianza de Y , la varianza explicada por la regresión y la varianza residual. 3- El coeficiente de determinación. 4- Qué cantidad de producto se vendería a un precio de 10€/Kg. Comente cada apartado.

$$a_{01} = 3,5 \quad a_{10} = 14 \quad a_{11} = 45,5 \quad a_{20} = 200,33 \quad b = \frac{-3,5}{4,33} = -0,80798$$

$$a_{02} = 15,16 \quad m_{20} = 4,33 \quad m_{02} = 2,9166 \quad m_{11} = -3,5$$

Esto significa que por cada euro por kilo que se aumente, las ventas se reducirán en un 80,798%

$$y - 3,5 = -0,808(x - 14) \Rightarrow y = -0,808x + 14,814$$

$$S_y^2 = 15,16 - (3,5)^2 = 2,9166 \quad S_{ry}^2 = m_{02} - \frac{m_{11}^2}{m_{20}} = 0,08965 \quad S_{yti}^2 = S_y^2 - S_{ry}^2 = 2,82695$$

La varianza de la variable endógena (y) y la varianza explicada por la regresión (y_{ti}) están muy aproximadas. Esto significa que al calcular el coeficiente de determinación el valor esté muy próximo a 1.

$$R^2 = 1 - \frac{S_{ry}^2}{S_y^2} = 1 - \frac{0,08965}{2,9166} = 0,9662 \quad \text{El } 96,62\% \text{ de la varianza de la variable endógena}$$

está explicada por la variable dependiente a través de la función ajustada. El modelo lineal es aceptable.

$$\text{Si } x = 10\text{€/Kg} \quad y = -0'808 \cdot (10) - 14'814 \Rightarrow \quad y = 6,734\text{Kg}$$

♦ Se dispone de la siguiente información relativa a dos variables:

$$\sum y_i = 186,95 \quad \sum x_i = 55 \quad \sum y_i^2 = 4144,07 \quad \sum x_i^2 = 385 \quad \sum y_i x_i = 1258,4$$

Se pide:

- Ajustar los coeficientes de la recta de regresión utilizando las fórmulas mínimo cuadráticas.
- Qué tanto por ciento de la variabilidad de y es explicada por la regresión.
- Obtener la predicción del valor de y para $x = 12$

Septiembre 2000 ADE

a) Ajustaremos la recta de regresión de Y sobre X $y - a_{01} = b(x - a_{10})$

$$a_{10} = \frac{\sum x}{N} \quad a_{01} = \frac{\sum y}{N} \quad a_{11} = \frac{\sum xy}{N} \quad a_{20} = \frac{\sum x^2}{N} \quad a_{02} = \frac{\sum y^2}{N}$$

$$m_{11} = a_{11} - a_{10} \cdot a_{01}$$

$$m_{20} = a_{20} - (a_{10})^2 \quad b = \frac{m_{11}}{m_{20}}$$

$$m_{02} = a_{02} - (a_{01})^2$$

Observamos que falta el dato de la frecuencia total, por lo tanto hagamos que $N = 10$

$$a_{10} = \frac{55}{10} = 5,5 \quad a_{01} = \frac{186,95}{10} = 18,695 \quad a_{11} = \frac{1258,4}{10} = 125,84 \quad a_{02} = \frac{4144,07}{10} = 414,407$$

$$m_{11} = 125,84 - (5,5 \cdot 18,695) = 23,0175$$

$$a_{20} = \frac{385}{10} = 38,5$$

$$m_{20} = 38,5 - (5,5)^2 = 8,25$$

$$b = \frac{23,0175}{8,25} = 2,79$$

$$m_{02} = 414,407 - (18,695)^2 = 64,904$$

$$y - 18,695 = 2,79(x - 5,5) \Rightarrow y = 2,79x + 3,35$$

b) El coeficiente de determinación es:

$$R^2 = 1 - \frac{S_{ry}^2}{S_y^2} = 1 - \frac{m_{02} - \frac{m_{11}^2}{m_{20}}}{m_{02}} = \frac{m_{02} - m_{02} + \frac{m_{11}^2}{m_{20}}}{m_{02}} = \frac{m_{11}^2}{m_{02} \cdot m_{20}}$$

$$R^2 = \frac{(23,0175)^2}{64,904 \cdot 8,25} = 0,9894 \quad \text{El } 98,94\% \text{ de varianza es explicada por la regresión.}$$

c) Si $x = 12$ $y = 2,79 \times 12 + 3,35 = 36,83$

AJUSTES NO LINEALES.

1. **Función polinómica:** $Y = a_0 + a_1X + a_2X^2 + \dots + a_nX^n$

- Con los siguientes datos, ajustar una parábola que exprese la relación entre ambas variables.

y_i	X_i
5	1
8	2
10	3
15	5
19	7

La curva a ajustar sería: $y_j^* = a + bx_i + cx_i^2$

$$\begin{aligned}\sum y_i &= Na + b\sum x_i + c\sum x_i^2 \\ \sum y_i x_i &= a\sum x_i + b\sum x_i^2 + c\sum x_i^3 \\ \sum y_i x_i^2 &= a\sum x_i^2 + b\sum x_i^3 + c\sum x_i^4\end{aligned}$$

y_j	x_i	x^2	x^3	x^4	$y_j x$	$y_j x^2$
5	1	1	1	1	5	5
8	2	4	8	16	16	32
10	3	9	27	81	30	90
15	5	25	125	625	75	375
19	7	49	343	2401	133	931
57	18	88	504	3124	259	1433

$$57 = 5a + 18b + 88c$$

$$259 = 18a + 88b + 504c$$

$$1433 = 88a + 504b + 3124c$$

$$a = \frac{\begin{bmatrix} 57 & 18 & 88 \\ 259 & 88 & 504 \\ 1433 & 504 & 3124 \end{bmatrix}}{\begin{bmatrix} 5 & 18 & 88 \\ 18 & 88 & 504 \\ 88 & 504 & 3124 \end{bmatrix}} = \frac{17176}{7504} = 2'289$$

$$b = \frac{\begin{bmatrix} 5 & 57 & 88 \\ 18 & 259 & 504 \\ 88 & 1433 & 3124 \end{bmatrix}}{7504} = \frac{21436}{7504} = 2'8566$$

$$c = \frac{\begin{bmatrix} 5 & 18 & 57 \\ 18 & 88 & 259 \\ 88 & 504 & 1433 \end{bmatrix}}{7504} = \frac{-500}{7504} = -0'067$$

$$y_j = 2'289 + 2'8566b - 0'067c$$

2. Función potencial: $Y = aX^b$

Para resolverla hay que tomar logaritmos neperianos $\text{Ln}Y = \text{Ln}[aX^b]$ por las propiedades de los logaritmos $\text{Ln}Y = \text{Lna} + b\text{Ln}X$

Ahora llamamos:

$$\text{Ln}Y = Y'$$

$$\text{Lna} = a' \quad \text{quedando una ecuación lineal del tipo } Y' = a' + bX'$$

$$\text{Ln}X = X'$$

- La relación entre renta disponible y consumo en un determinado grupo, viene determinado por los siguientes datos. Ajustar una función de consumo potencial.

Consumo C	Renta R
1	2
2	4
3	6
5	8

La función que debemos ajustar será: $C = aR^b$ llamaremos:

$$\text{Ln}C = c \quad \text{Ln}R = r \quad \text{Lna} = a' \quad \text{quedando } c = a' + br$$

Las ecuaciones quedarían:

$$\begin{aligned} \sum c &= Na' + b \sum r \\ \sum cr &= a' \sum r + b \sum r^2 \end{aligned}$$

C	R	c=LnC	r=LnR	r·c=LnR·LnC	$r^2 = (\text{Ln}R)^2$
1	2	0	0,693	0	0,480
2	4	0,693	1,386	0,960	1,921
3	6	1,099	1,792	1,969	3,211
5	8	1,609	2,079	3,345	4,322
		3,401	5,95	6,275	9,935

$$3'401 = 4a' + 5'95r$$

$$6'275 = 5'95a' + 9'935b$$

$$a' = \frac{\begin{bmatrix} 3'401 & 5'95 \\ 6'275 & 9'935 \end{bmatrix}}{\begin{bmatrix} 4 & 5'95 \\ 5'95 & 9'935 \end{bmatrix}} = \frac{-3'547}{4'3375} = -0'818$$

$$b = \frac{\begin{bmatrix} 4 & 3'401 \\ 5'95 & 6'275 \end{bmatrix}}{4'3375} = \frac{4'864}{4'3375} = 1'214$$

$$a' = \text{Lna} \Rightarrow -0'818 = \text{Lna} \Rightarrow a = e^{-0'818} \Rightarrow a = 0'4413$$

la función queda: $C = 0'4413R^{1'214}$

3. Función exponencial. $Y = ab^X$

La resolvemos tomando logaritmos $\text{Ln}Y = \text{Ln}(ab^X)$ quedando: $\text{Ln}Y = \text{Lna} + X \cdot \text{Lnb}$

$$\text{Ln}Y = y'$$

$$\text{Lna} = a' \quad y' = a' + b'X$$

$$\text{Lnb} = b'$$

- Ajustar una función exponencial dada la siguiente distribución bidimensional.

X	Y
1	2
2	1
3	5
4	6

Las ecuaciones quedarían:

$$\sum y' = a' + b' \sum x$$

$$\sum y'x = a' \sum x + b' \sum x^2$$

X	Y	$y' = \text{Ln}Y$	$y' \cdot X = \text{Ln}Y \cdot X$	X^2
1	2	0,693	0,693	1
2	1	0	0	4
3	5	1,609	4,827	9
4	6	1,792	7,168	16
10		4,094	12,688	30

$$4'094 = 4a' + 10b'$$

$$12'688 = 10a' + 30b'$$

$$a' = \frac{\begin{bmatrix} 4'094 & 10 \\ 12'688 & 30 \end{bmatrix}}{\begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix}} = \frac{-4'06}{20} = -0'203$$

$$b' = \frac{\begin{bmatrix} 4 & 4'094 \\ 10 & 12'688 \end{bmatrix}}{20} = \frac{9'812}{20} = 0'491$$

$$a' = Lna \Rightarrow -0'203 = Lna \Rightarrow a = e^{-0'203} \Rightarrow a = 0'816$$

$$b' = Lnb \Rightarrow 0'491 = Lnb \Rightarrow b = e^{0'491} \Rightarrow b = 1'634$$

$$Y = [(0'816) \cdot (1'634)]^X$$

4. **Función logarítmica:** $Y = a + bLnx$

Se resuelve llamando $Lnx = X'$ y resolviendo como una función lineal.