EJERCICIO DE REPASO

En el análisis técnico de un prototipo de robótica es necesario estudiar dos índices, X e Y. La siguiente tabla informa sobre el resultado de 20 mediciones conjuntas de tales índices

X \ Y	[0-8]	(8-20]	(20-50]
-1	0	0	4
0	0	2	3
1	4	2	0
2	5	0	0

a) Supuesto que $|X| \leq 1$, determinar el número de mediciones del índice Y cuyos valores oscilan entre el valor modal y el valor mínimo del 25% de los índices más altos.
b) ¿Qué índice medio es más representativo de su distribución, el de X o el de Y?
c) Estudiar la interdependencia lineal de las variables estadísticas X e Y.
d) Estimar mediante una recta de regresión mínimo cuadrática el valor de Y cuando X=0. ¿Coincide este valor con la estimación de menor error cuadrático medio de Y cuando X=0?

SOLUCIÓN

a) En primer lugar tenemos que calcular la moda de la distribución condicionada $Y/|X| \leq 1$.

$I_j = (e_{j-1}, e_j]$	n_j	$a_j = e_j - e_{j-1}$	$d_j = \frac{n_j}{a_j}$
[0, 8]	4	8	0.500
(8, 20]	4	12	0.333
(20, 50]	7	30	0.233

La distribución es unimodal. La densidad de frecuencia máxima es 0.500 y el intervalo modal es $I_1 = [0, 8]$. $Mo_{Y/X<2} \in I_1 = [0, 8]$ y se calcula como se muestra en la Figura 1.

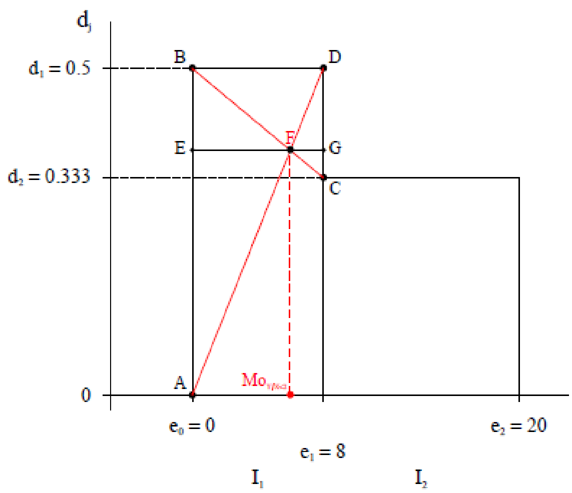

Figura 1

Los triángulos $\widehat{AFB}$ and $\widehat{CFD}$ son semejantes, y entonces

$$\frac{EF}{EF + FG} = \frac{AB}{AB + CD}$$

o, equivalentemente,

$$\frac{Mo_{Y/X<2} - e_0}{a_1} = \frac{d_1 - 0}{(d_1 - 0) + (d_1 - d_2)}$$

Así,

$$Mo_{Y/X<2} = e_0 + \frac{d_1}{2\,d_1 - d_2}\,a_1 = 0 + \frac{0.5}{2 \times 0.5 - 0.333}\,8 = 5.997.$$

Una vez conocido el valor de la moda, calculamos el número de individuos cuyo valor de la variable es inferior a la moda, utilizando la función de distribución:

$I_j = (e_{j-1}, e_j]$	n_j	$N_j = n_j + N_{j-1}$
[0, 8]	4	4
(8, 20]	4	8
(20, 50]	7	15

$F_{Y/X<2}(5.997)$ y $n\,F_{Y/X<2}(5.997)$ son, respectivamente, la proporción y el número de datos menores o iguales a 5.997. Puesto que $5.997 \in I_1 = (e_0, e_1] = (0, 8]$, resulta que

$$N_0 = 0 < n\,F_{Y/X<2}(5.997) < 4 = N_1$$

y $n\,F_{Y/X<2}(5.997)$ se calcula por interpolación lineal entre los puntos

$$(e_0, N_0) = (0, 0)$$

y

$$(e_1, N_1) = (8, 4)$$

como muestra la Figura 2.

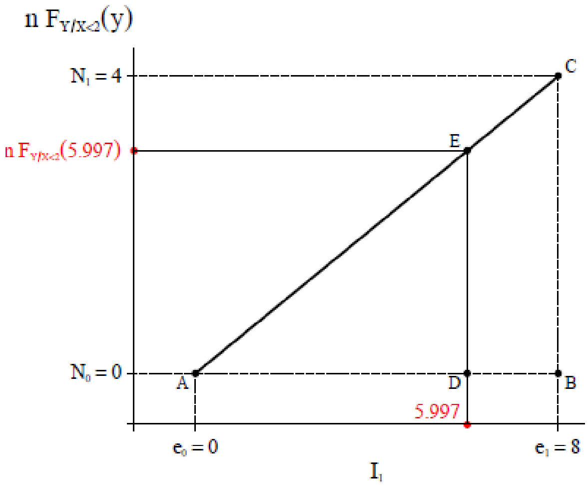

Figura 2

El triángulo $\widehat{ABC}$ es semejante al triángulo $\widehat{ADE}$, y por tanto $\dfrac{DE}{BC} = \dfrac{AD}{AB}$, esto es,

$$\frac{n\,F_{Y/X<2}(5.997) - N_0}{N_1 - N_0} = \frac{5.997 - e_0}{e_1 - e_0}$$

Así,

$$n\,F_{Y/X<2}(5.997) = N_0 + \frac{5.997 - e_0}{e_1 - e_0}(N_1 - N_0) = 0 + \frac{5.997 - 0}{8 - 0}(4 - 0) = 2.998$$

y

$$F_{Y/X<2}(5.997) = \frac{2.998}{15} = 0.2.$$

El porcentaje de datos menores o iguales que 5.997 es del 20 %, y el porcentaje de datos mayores que 5.997 es del 80 %.

Si dicha moda deja por debajo al 20% de los datos y el valor mínimo del 25% de los valores más altos (percentil 75) deja por debajo al 75%, significa que entre dichos valores hay un 55% de los 15 datos; esto es, 15×55/100 = 8.25 ≅8 individuos.

b) Para estudiar la representatividad de las medias utilizamos los coeficientes de variación:

x_i	n_i	$x_i n_i$	$(x_i-\bar{x})^2 n_i$
−1	4	−4	10.24
0	5	0	1.80
1	6	6	0.96
2	5	10	9.80
	20	12	22.80

$x_i^2 n_i$
4
0
6
20
30

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{4} x_i n_i = \frac{12}{20} = 0.6. \qquad m_{2\,(X)} = \frac{1}{n}\sum_{i=1}^{4} x_i^2 n_i = \frac{30}{20} = 1.5$$

$$\sigma_X^2 = \frac{1}{n}\sum_{i=1}^{4}(x_i-\bar{x})^2 n_i = \frac{22.80}{20} = 1.14. \qquad \sigma_X^2 = m_{2\,(X)} - \bar{x}^2 = 1.14$$

$$\sigma_X = \sqrt{1.14} = 1.068.$$

$$CV_X = \frac{\sigma_X}{\bar{x}} = \frac{1.068}{0.6} = 1.78.$$

$I_j=(e_{j-1},e_j]$	n_j	y_j	$y_j n_j$	$(y_j-\bar{y})^2 n_j$
[0, 8]	9	4	36	1486.103
(8, 20]	4	14	56	32.490
(20, 50]	7	35	245	2305.957
	20		337	3824.550

$y_j^2 n_j$
144
784
8575
9503

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{3} y_j n_j = \frac{337}{20} = 16.85.$$

$$\sigma_Y^2 = \frac{1}{n}\sum_{j=1}^{3}(y_j-\bar{y})^2 n_j = \frac{3824.550}{20} = 191.227.$$

$$m_{2\,(Y)} = \frac{1}{n}\sum_{j=1}^{3} y_j^2 n_j = \frac{9503}{20} = 475.15$$

$$\sigma_Y^2 = m_{2\,(Y)} - \bar{y}^2 = 191.227$$

$$\sigma_Y = \sqrt{191.227} = 13.828.$$

$$CV_Y = \frac{\sigma_Y}{\bar{y}} = \frac{13.828}{16.85} = 0.821.$$

Por lo tanto, la media más representativa es la del índice Y.

c) Para estudiar la interdependencia lineal entre los índices tenemos que calcular el coeficiente de correlación lineal. Sólo tenemos que calcular la covarianza, pues las desviaciones típicas y las medias han sido ya calculadas en el anterior apartado.

X \ Y	[0, 8]	(8, 20]	(20, 50]	$n_{i\cdot}$	$x_i\sum_{j=1}^{3} y_j n_{ij}$
-1	0	0	4	4	-140
0	0	2	3	5	0
1	4	2	0	6	44
2	5	0	0	5	40
$n_{\cdot j}$	9	4	7	20	-56
y_j	4	14	35		

$$\sigma_{X,Y} = \frac{1}{n}\sum_{i=1}^{4}\sum_{j=1}^{3} x_i y_j n_{ij} - \bar{x}\bar{y} = \frac{-56}{20} - 0.6 \times 16.85 = -12.91.$$

Por lo tanto, el coeficiente de correlación lineal vale:

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{-12.91}{1.068 \times 13.828} = -0.8741$$

lo que significa que existe bastante relación lineal inversa.

d) Tenemos que calcular la recta de regresión de Y/X (y=ax+b)
$a = \frac{\sigma_{xy}}{{\sigma_x}^2} = \frac{-12.91}{1.14} = -11.34;\quad b = \bar{y} - a\bar{x} = 16.8 + 11.34 \times 0.6 = 23.654.$
La recta de regresión es, por tanto $y = -11.34x + 23.654$.
El valor estimado de Y cuando X=0 es 23.654.

Calculamos ahora la estimación de menor error cuadrático medio de Y cuando X=0; esto es, la de la media condicionada de Y al valor X=0.

$I_j = (e_{j-1}, e_j]$	n_j	y_j	$y_j n_j$
[0, 8]	0	4	0
(8, 20]	2	14	28
(20, 50]	3	35	105
	5		133

$$\bar{y}_{/X=0} = \frac{1}{n}\sum_{j=1}^{3} y_j n_j = \frac{133}{5} = 26.6.$$

Como vemos, dicho valor no coincide con el de la estimación mediante la recta de regresión.