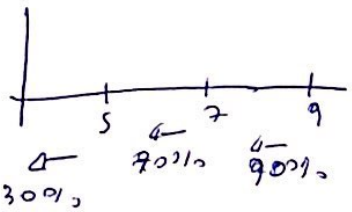
Javier Gómez López

1. Indica si son verdaderas o falsas las siguientes afirmaciones, justificando la respuesta:

1.1a) Podemos afirmar que $P_{30} = 5$, $P_{70} = 7$ $P_{90} = 9$



observamos que el porcentaje entre el percentil 30 y 70 es del 40% y coincide con el intervalo $(5, 7]$, mientras que entre el percentil 70 y 90 hay un 20% y corresponde al intervalo $(7, 9]$. Por tanto es verdadera.

b)

1.2 Si X e Y son estadísticamente independientes, $m_{rs} = m_{r0} \cdot m_{0s}$

Así, $m_{32} = m_{30} \cdot m_{02} = 4 \cdot 5 = 20 \neq 18$. Falsa.

1.3. $y = ax + b$ si pasa por el origen, $b = 0$

$b = \bar{y} - a\bar{x}$; $\bar{y} = a\bar{x}$; $a = \frac{\bar{y}}{\bar{x}} = \frac{m_{01}}{m_{10}}$. Falsa

1.4 $r^2 = \frac{\sigma_{xy}^2}{\sigma_x^2 \sigma_y^2}$

2. Sea $\{X, Y\}$ una variable estadística bidimensional con distribución de frecuencias $\{(x_i, y_j)_i n_{ij}\}$. Obtener la media y la varianza de la distribución marginal de X.

$$\bar{x} = \sum_{i=1}^{k} f_{i\cdot} x_i = \frac{1}{n} \cdot \sum_{i=1}^{k} n_{i\cdot} x_i = \frac{1}{n} \sum_{j=1}^{p} n_{ij} \cdot \sum_{i=1}^{k} x_i = \sum_{j=1}^{p} \sum_{i=1}^{k} \frac{n_{ij}}{n} x_i =$$

$$= \sum_{j=1}^{p} \sum_{i=1}^{k} \frac{n_{\cdot j}}{n} \cdot \frac{n_{ij}}{n_{\cdot j}} x_i = \sum_{j=1}^{p} \sum_{k=1}^{k} f_{\cdot j} f_i^j x_i = \sum_{j=1}^{p} f_{\cdot j} \sum_{i=1}^{k} f_i^j x_i = \sum_{j=1}^{p} f_{\cdot j} \bar{x}_j$$

$$\sigma_x^2 = \sum_{i=1}^{k} f_{i\cdot} (x_i - \bar{x})^2 = \sum_{i=1}^{k} \frac{n_{i\cdot}}{n} (x_i - \bar{x})^2 = \sum_{j=1}^{p} \sum_{i=1}^{k} \frac{n_{ij}}{n} (x_i - \bar{x})^2 =$$

$$= \sum_{j=1}^{p} \sum_{i=1}^{k} \frac{n_{\cdot j}}{n} \cdot \frac{n_{ij}}{n_{\cdot j}} (x_i - \bar{x})^2 = \sum_{j=1}^{p} \sum_{i=1}^{k} f_{\cdot j} f_i^j (x_i - \bar{x})^2 =$$

$$= \sum_{j=1}^{p} f_{\cdot j} \left[\sum_{i=1}^{k} f_i^j (x_i - \bar{x})^2 \right] = \sum_{j=1}^{p} f_{\cdot j} \left[\sum_{i=1}^{k} f_i^j (x_i - \bar{x}_j + \bar{x}_j - \bar{x})^2 \right] =$$

$$= \sum_{j=1}^{p} f_{\cdot j} \left[\sum_{i=1}^{k} f_i^j (x_i - \bar{x}_j)^2 + \sum_{i=1}^{k} f_i^j (\bar{x}_j - \bar{x})^2 + 2 \cdot \sum_{i=1}^{k} f_i^j (x_i - \bar{x}_j)(\bar{x}_j - \bar{x}) \right] =$$

$$= \sum_{j=1}^{p} f_{\cdot j} \left[\sum_{i=1}^{k} f_i^j (x_i - \bar{x}_j)^2 + \sum_{i=1}^{k} f_i^j (\bar{x}_j - \bar{x})^2 \right] =$$

$$= \sum_{j=1}^{p} f_{\cdot j} \left[\sigma_{x_j}^2 + (\bar{x}_j - \bar{x})^2 \sum_{i=1}^{k} f_i^j \right] = \sum_{j=1}^{p} f_{\cdot j} \left[\sigma_{x_j}^2 + (\bar{x}_j - \bar{x})^2 \right] =$$

$$= \sum_{j=1}^{p} f_{\cdot j} \sigma_{x_j}^2 + \sum_{j=1}^{p} f_{\cdot j} (\bar{x}_j - \bar{x})^2$$

3. Se realiza un estudio para observar el tiempo que tardan en resolver unos escolares que han seguido un curso de formación por módulos. Se observa el número de módulos que han superado (X) junto con el tiempo en minutos que tardan en resolver el problema (Y)

$X \backslash Y$	$[1-9]$	$(9-21]$	$(21-39]$	$n_{i\cdot}$	$n_{i\cdot} x_i$	$n_{i\cdot} x_i^2$
2	0	1	5	6	12	24
4	0	5	5	10	40	160
5	5	3	0	8	40	200
8	15	1	0	16	128	1024
$n_{\cdot j}$	20	10	10	40	220	1408
$n_{\cdot j} y_j$	100	150	300	550		
$n_{\cdot j} y_j^2$						

a) Sabiendo que $\sigma_Y^2 = 104'6875$, ¿qué valor medio es más representativo, ¿el de X o el de Y?

Primero calculamos la media de X e Y:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{4} n_{i\cdot} x_i = 5'5 \text{ módulos superados}$$

$$\bar{y} = \frac{1}{n} \sum_{j=1}^{3} n_{\cdot j} y_j = \frac{550}{40} = 13'75 \text{ minutos}$$

Ahora calculamos sus varianzas y sus desviaciones típicas:

$\sigma_Y^2 = 104'6875$ minutos^2

$$\sigma_X^2 = m_2 - m_1^2 = \frac{1}{n} \sum_{i=1}^{4} n_{i\cdot} x_i^2 - (5'5)^2 = \frac{1408}{40} - 5'5^2 = 4'25 \text{ módulos}^2$$

$\sigma_X = 2'225$ módulos

$\sigma_Y = 10'232$ minutos

Ahora calculamos sus coeficientes de variación de Pearson

$$C.V(X) = \frac{\sigma_X}{\bar{x}} = \frac{2'225}{5'5} = 0'405$$

$$C.V(Y) = \frac{\sigma_Y}{\bar{y}} = \frac{10'232}{13'25} = 0'744$$

[Por lo tanto el valor medio más representativo es el de X]

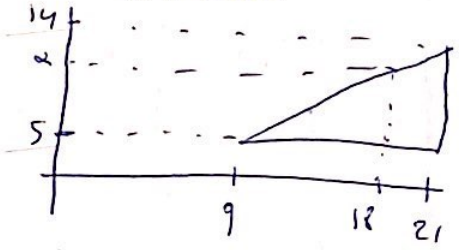
Javier Gómez López

5) Para los estudiantes que superan menos de 6 módulos, ¿qué porcentaje tarda menos de 18 minutos en resolver el problema? ¿Cuál es el tiempo de respuesta más frecuente?

X \ Y	[1-9]	(9-21]	(21-39]	→ (Marginal de Y / x < 6
2	0	1	5	
4	0	5	5	
5	5	3	0	
n.º i	5	9	10	
h_i	$5/18$	$9/12$	$10/18$	

Calculamos cuantos tardan menos de 18 min

(No escala)



observamos que $18 \in (9-21]$

Hacemos semejanza de triángulos

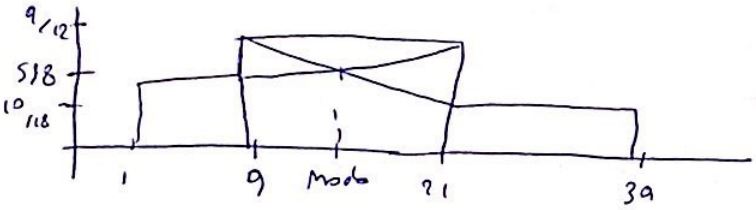
$$\frac{18-9}{21-9} = \frac{a-5}{14-5} \; ; \; a = 11'75$$

Es decir, 11'75 alumnos tardan menos de 18 min en superar 6 o menos módulos

$\frac{11'75}{24} = 0'4896$. Entonces, podemos afirmar que el 48'96% tardan menos de 18 min en resolver menos de 6 módulos.

Ahora calculamos la moda:

(No escala



Usando semejanza de triángulos:

$$\frac{9/12 - 5/18}{(9/12 - 5/18) + (9/12 - 10/18)} = \frac{Mo - 9}{(21 - Mo) + (Mo - 9)} \; ; \; [Mo = 13'696 \text{ minutos}]$$

Javier Gómez

5/18

c) Calcular el número máximo de módulos del 40 % de los estudiantes que más módulos superan

Nos piden el percentil 60, ya que deja por encima al 40 %.

$\frac{n \cdot k}{100} = \frac{40 \cdot 60}{100} = 24$ $\quad 24 = N_3 \rightarrow x_3 = 5$ módulos.

[Concluimos que, el número mínimo de módulos que superan el 40 % son más módulos es 5]

d) Sabiendo que $m_{11} = 57'5$, estimar el valor de Y cuando $x = 4$ mediante una recta de regresión mínimo cuadrática y dar una medida de la bondad de la predicción.

$y = ax + b \qquad a = \frac{\sigma_{xy}}{\sigma_x^2} \qquad b = \bar{y} - a\bar{x}$

$\sigma_{xy} = m_{11} - m_{10} \cdot m_{01} = m_{11} - \bar{x}\bar{y} = 57'5 - 5'5 \cdot 13'75 = -18'125$

$a = \frac{-18'125}{4'95} = -3'662 \qquad b = 13'75 - (-3'662) \cdot 5'5 = 33'89$

$y = -3'662x + 33'89$. Cuando $x = 4$, $[y = 19'242]$ minutos

Ahora calculamos la varianza de correlación. $r^2_{Y/X} = \frac{\sigma_{xy}^2}{\sigma_x^2 \sigma_y^2} = \frac{(-18'125)^2}{4'95 \cdot 104'6875}$

$[r^2_{Y/X} = 0'634]$

Javier Gómez

6/9

e) Ajustar a los datos un modelo de regresión hiperbólico para predecir el tiempo de respuesta conociendo el número de módulos superados y predecir el tiempo de respuesta de un estudiante que superó 4 módulos

$$a = \frac{\sigma_{zy}}{\sigma_z^2} \qquad b = \bar{y} - a\bar{z}$$

$y = a\frac{1}{x} + b$, $z = \frac{1}{x} \Rightarrow y = az + b$

Y	8	12	18
z			

$z \setminus Y$	3	15	30	$n_{i\cdot}$	$z_i \cdot n_{i\cdot}$	$z_i^2 n_{i\cdot}$
1/2	0	1	5	6	3	3/2
1/4	0	5	5	10	2'5	0'625
1/5	5	3	0	8	1'6	0'32
1/8	15	1	0	16	2	0'25
				40	9'1	2'695

Los medidos de y los tenemos ya.

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{4} z_i n_{i\cdot} = \frac{9'1}{40} = 0'2275 \text{ 1/módulos}$$

$$\sigma_z^2 = m_2 - m_1^2 = \frac{1}{n}\sum_{i=1}^{4} z_i^2 n_{i\cdot} - \bar{z}^2 = \frac{2'695}{40} - 0'2275^2 = 0'0156 \text{ 1/módulos}^2$$

$$\sigma_{zy} = m_{11} - m_{10}m_{01} = \frac{1}{n}\sum_{i=1}^{4}\sum_{j=1}^{3} z_i \cdot y_j \cdot n_{ij} - \bar{z}\bar{y} = \frac{164}{40} - 0'2275 \cdot 13'75 =$$

$$= 0'972 \text{ min/módulos}$$

$$a = \frac{0'972}{0'0156} = 62'3 \qquad b = -0'422$$

$$y = 62'3z - 0'422 \Rightarrow y = \frac{62'3}{x} - 0'422$$

[Si $x = 4$, $y = 15'153$ minutos]

Javier Gómez

f) Para predecir el tiempo de respuesta, ¿qué modelo de regresión es más ajustado, el lineal o el hiperbólico?

Calculemos la ratio de correlación $\eta^2_{y,x}$ de la hipérbola

$f(2) = 30'728$

$f(4) = 15'153$

$f(5) = 12'038$

$f(8) = 7'368$

$$\sigma_{e_y}^2 = \frac{1}{n}\sum_{i=1}^{4}\sum_{j=1}^{2} n_{ij}(f(x_i) - \bar{y})^2 =$$

$$= \frac{1}{40} \cdot (1729'515 + 19'684 + 23'448 + 652'608) =$$

$$= 60'610$$

$$\eta^2_{y/x} = \frac{\sigma_{e_y}^2}{\sigma_y^2} = \frac{60'61}{104'6075} = 0'5789$$

Como el valor de la ratio de correlación en la recta es mayor, afirmamos que este ajuste es mejor que el hiperbólico.