-ESTADÍSTICA DESCRIPTIVA E INTRODUCCIÓN A LA PROBABILIDAD-
Prueba Temas 1-2

Doble Grado en Ingeniería Informática y Matemáticas
22 de abril de 2021

1. ***[3 puntos] Indicar si son verdaderas o falsas las siguientes afirmaciones, justificando la respuesta:***

1.1. De la distribución de la variable X='calificación en un examen' observada sobre 50 alumnos se sabe lo siguiente: sus observaciones se agruparon en los siguientes intervalos de clase: [0,5], (5,7], (7,9] y (9,10]; y los percentiles 30, 70 y 90 fueron 5, 7 y 9 puntos, respectivamente.

a) El número de alumnos observados cuyas calificaciones oscilan entre (5,7] duplica al de alumnos observados cuyas calificaciones oscilan entre (7,9].

b) La desviación típica de X vale 5.7 puntos.

1.2. Si dos variables estadísticas X e Y son independientes, m_{02}=5 y m_{30}=4, entonces m_{32}=18.

1.3. Si a una nube de puntos se le ajusta por mínimos cuadrados una recta de Y/X que pase por el origen, entonces la pendiente de esa recta vale m_{11}/m_{20}.

1.4. El coeficiente de correlación lineal entre dos variables X e Y'=-3Y+2 es el mismo que entre X e Y.

1.5. En una distribución simétrica, siempre coinciden la media aritmética, la mediana y la moda.

2. ***[2 puntos] Sea (X,Y) una variable estadística bidimensional con distribución de frecuencias dada por $\{(x_i, y_j); n_{ij}\}_{i=1,..k;\, j=1,..p}$. Obtener la media y varianza de la distribución marginal de X en función de las medias y varianzas de las distribuciones condicionadas de X a cada valor de Y, justificando todos los pasos de la demostración.***

3. ***[5 puntos] Se realiza un estudio para observar el tiempo que tardan en resolver un problema 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 \ Y	[1-9]	(9-21]	(21-39]
2	0	1	5
4	0	5	5
5	5	3	0
8	15	1	0

a) Sabiendo que $\sigma_Y^2 = 104.6875$, ¿qué valor medio es más representativo, el de X o el de Y?

b) 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?

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

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.

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.

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

Parcial EDIP

Alumno: José Alberto Hoces Castro

1. 1.1) 50 alumnos

[0,5] $\quad X \equiv$ calif. examen

(5,7] $\quad P_{30} = 5$ puntos

(7,9] $\quad P_{70} = 7$ " "

(9,10] $\quad P_{90} = 9$ " "

a) Como $P_{30} = 5$ puntos y $P_{70} = 7$ puntos, en el intervalo (5,7] tenemos al $70\% - 30\% = 40\%$ alumnos.

Como $P_{70} = 7$ puntos y $P_{90} = 9$ puntos, en el intervalo de clase (7,9] tenemos al $90\% - 70\% = 20\%$. Por lo tanto, es verdad que los alumnos en (5,7] duplican a los de (7,9]. Es verdadero.

b)

1.2) Como se vio en clase, si X e Y son independientes, $m_{rs} = m_{r0} m_{0s}$. Como en este caso nos dan $m_{02} = 5$ y $m_{30} = 4$, sabemos que $m_{32} = m_{30} m_{02} = 4 \cdot 5 = 20 \neq 18$, por lo tanto es falsa.

1.4) X e $Y' = -3Y+2$

Hallemos $\sigma_{xy'}$:

$$\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}\, x_i\, y'_j - m_{10}\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}\, y'_j =$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}\, x_i(-3y_j+2) - m_{10}\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}(-3y_j+2) =$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{p} -3f_{ij}x_i y_j + 2\sum_{i=1}^{k} f_{i.}x_i - m_{10}\left(-3\sum_{j=1}^{p} f_{.j}\, y_j + 2\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}}_{1}\right) =$$

$$= -3m_{11} + 2m_{10} - 2m_{10} + 3m_{10}m_{01} = 3m_{01}m_{01} - 3m_{11}$$

Y $\sigma_{y'}^2$:

$$\sigma_{y'}^2 = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}\, y_j'^2 - \left(\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}\, y_j'\right)^2 =$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}(-3y_j+2)^2 - \left(\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}(-3y_j+2)\right)^2 =$$

$$\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}(9y_j^2-12y_j+4) - (3m_{01}+2)^2 =$$

$$= 9m_{02} - 12m_{01} + 4 - (9m_{01}^2 - 12m_{01} + 4) =$$

$$= 9m_{02} - 9m_{01}^2 = 9(m_{02} - m_{01}^2) = 9\sigma_y^2 \Rightarrow \sigma_{y'} = 3\sigma_y$$

$$r = \frac{\sigma_{xy'}}{\sigma_x \sigma_{y'}} = \frac{-3m_{11} + 3m_{10}m_{01}}{\sigma_x \sigma_{y'}} = \frac{-3(m_{11} - m_{10}m_{01})}{\sigma_x \sigma_y 3} =$$

$$= \frac{-3\sigma_{xy}}{\sigma_x\sigma_y 3} = -\frac{\sigma_{xy}}{\sigma_x\sigma_y} \Rightarrow r^2 = \frac{\sigma_{xy}^2}{\sigma_x^2\sigma_y^2}$$

Verdadera. Tienen el mismo coeficiente de correlación lineal.

2.

Empezaremos con la media. Este es uno de los ejercicios propuestos de Semana Santa, por lo que sé cómo se expresa lo que se nos pide (la relación que buscamos. Por ello, empiezo con que $\overline{x} = \sum_{j=1}^{p} f_{\cdot j}\overline{x}_j$ y llegaré a $\sum_{i=1}^{k} f_{i\cdot}x_i$:

Hemos puesto la expresión general de $\overline{x}_j$

$f_{\cdot j} = \frac{n_{\cdot j}}{n}$

$f_{i/j} = \frac{n_{ij}}{n_{\cdot j}}$

$$\overline{x} = \sum_{j=1}^{p} f_{\cdot j}\overline{x}_j = \sum_{j=1}^{p} f_{\cdot j}\sum_{i=1}^{k} f_{i/j}x_i = \sum_{j=1}^{p}\sum_{i=1}^{k} f_{\cdot j}f_{i/j}x_i =$$

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

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

Y ahora vamos con la varianza de x:

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

$f_{i\cdot} = \frac{n_{i\cdot}}{n}$ $\qquad$ $n_{i\cdot} = \sum_{j=1}^{p} n_{ij}$

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

Sumamos y restamos $\overline{x}_j$

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

$f_{\cdot j}$ no depende de i

Aplicamos que $(a-b)^2 = a^2+b^2-2ab$

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

$$= \sum_{j=1}^{p} f_{\cdot j} \left[\sum_{i=1}^{k} f_{i/j} (x_i - \bar{x}_j)^2 + \sum_{i=1}^{k} f_{i/j} (\bar{x}_j - \bar{x})^2 + \underbrace{2 \sum_{i=1}^{k} f_{i/j} (x_i - \bar{x}_j)(\bar{x}_j - \bar{x})}_{0 \circledast} \right] =$$

$\circledast$

$$2 \sum_{i=1}^{k} f_{i/j} (x_i - \bar{x}_j)(\bar{x}_j - \bar{x}) \overset{j \text{ es indep. de } i}{=} 2(\bar{x}_j - \bar{x}) \underbrace{\sum_{i=1}^{k} f_{i/j} (x_i - \bar{x}_j)}$$

(Con el j fijo)

$$\sum_{i=1}^{k} f_{i/j} x_i - \sum_{i=1}^{k} f_{i/j} \bar{x}_j = \bar{x}_j - \bar{x}_j \sum_{i=1}^{k} f_{i/j} = \bar{x}_j - \bar{x}_j = 0 \checkmark$$

Retomamos por donde nos habíamos quedado:

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

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

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

3. $X \equiv$ nº módulos superados
$Y \equiv$ minutos en resolver el problema.

He sustituido directamente por las marcas de clase.

$X \setminus Y$	5	15	30	$n_{i.}$	$n_{i.} x_i$	$n_{i.} x_i^2$
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
$n_{.j}$	20	10	10	40	220	1408
$n_{.j} a_j$	100	150	300	550		

a) $\sigma_y^2 = 104.6875$. ¿Qué media es más representativa?
Hemos de centrarnos en las distribuciones marginales y hallar su respectivo coeficiente de variación de Pearson. A menor coeficiente, más homogeneidad y más representatividad:

$C.V.(X) = \frac{\sigma_x}{|\bar{x}|}$ $\quad \bar{x} = \frac{220}{40} = 5.5$ módulos superados

$\sigma_x = \sqrt{\frac{1408}{40} - 5.5^2} = 2.2249$ módulos superados

$C.V.(X) = \frac{2.2249}{5.5} = 0.40452$

$C.V(Y) = \frac{\sigma_y}{|\bar{y}|}$ $\quad \bar{y} = \frac{550}{40} = 13.75$ minutos

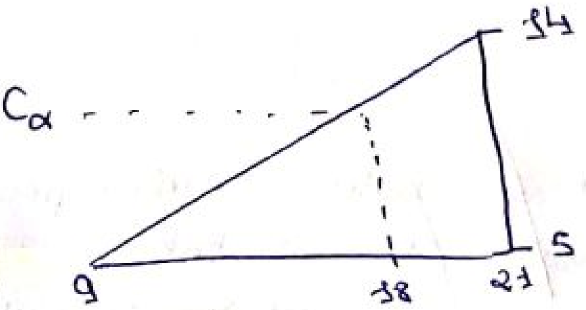
$C.V.(Y) = \frac{\sqrt{104.6875}}{13.75} = 0.74412$

Como $C.V.(X) < C.V.(Y)$, es más representativa la media de los módulos superados.

6) ~~Tenemos que fijarnos en los pares (x_i, y_j) con~~
Hemos de trabajar con la distribución condicionada $Y / X < 6$:

I_j	c_j	$n_{1j}+n_{2j}+n_{3j}$	a_j	N_j	h_j
[1-9]	5	5	8	5	0.625
(9-21]	15	9	12	14	0.75
(21-39]	30	10	18	24	$0.\hat{5}$
		24			

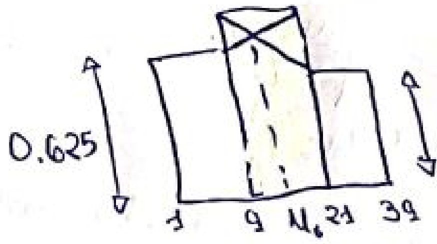
Si tardan menos de 18 minutos, hemos de hacer el cuantil inverso de la curva de distribución:



$$\frac{12}{11} = \frac{9}{n\alpha - 5}$$
$$\Downarrow$$
$$1.09n\alpha - 5.455 = 9$$
$$n\alpha = 13.2615$$
$$\alpha = 0.55256$$

55.256% de escolares que tardan menos de 18 minutos

Ahora nos falta la moda para la segunda pregunta. Busquemos en el intervalo modal, que es (9-21]:



$$\frac{0.75 - 0.625}{M_o - 9} = \frac{0.75 - 0.\hat{5}}{21 - M_o}$$
$$\frac{0.125}{M_o - 9} = \frac{0.19445}{21 - M_o}$$
$$-0.125M_o + 2.625 = 0.19445M_o - 1.75005$$
$$\Downarrow$$
$$4.37505 = 0.31945M_o \Rightarrow M_o = 13.696 \text{ minutos}$$

es el tiempo de respuesta más frecuente

c) Se nos está pidiendo el cuantil 0.60 de la distribución marginal X:

x_i	$n_{i.}$	$N_{i.}$
2	6	6
4	10	16
5	8	24
8	16	40

$n\alpha = 40 \cdot 0.60 = 24$

$\Downarrow$

Como $N_{3.}$ es justo 24, hemos de realizar la media entre x_3 y x_4:

$\frac{x_3+x_4}{2} = \frac{13}{2} = 6.5$ módulos como mínimo superan el 40% de los estudiantes que más módulos superan.

d) Como se pide estimar Y en función de X, se nos está pidiendo la recta de regresión Y|X que es de la forma

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

Solo nos falta la covarianza:

$\sigma_{xy} = \mu_{11} = m_{11} - m_{10}m_{01} = 57.5 - \bar{x}\bar{y} = 57.5 - 5.5 \cdot 13.75 =$
$= -18.125$

Y la recta será: $a = \frac{\sigma_{xy}}{\sigma_x^2} = -3.6615$

$b = \bar{y} - a\bar{x} = 33.89$

$$\boxed{y = -3.6615x + 33.89}$$

Si $x=4 \Rightarrow y = 19.244$ minutos

Para saber la bondad del ajuste nos hace falta saber el valor de r^2:

$r^2 = \frac{\sigma_{xy}^2}{\sigma_x^2\sigma_y^2} = 0.6339 \Rightarrow$ Explica menos del 64% de los casos, por lo que no es un ajuste demasiado fiable.

e) La hipérbola equilátera siempre presenta la misma forma:

$$y = b + \frac{a}{x}$$

Hacemos un cambio de variable $z = \frac{1}{x}$ y volvemos a hacer la tabla de doble entrada: $y = b + az$

$Z \backslash Y$	5	15	30	$n_{i.}$	$n_{i.} z_i$	$n_{i.} z_i^2$	$z_i \sum n_{ij} y_j$
0.5	0	1	5	6	3	1.5	82.5
0.25	0	5	5	10	2.5	0.625	56.25
0.2	5	3	0	8	1.6	0.32	14
0.125	15	1	0	16	2	0.25	11.25
				40	9.1	2.695	164

$$\bar{z} = \frac{9.1}{40} = 0.2275$$

$$\sigma_z^2 = \frac{2.695}{40} - 0.2275^2 = 0.015619$$

$$\sigma_{zy} = \frac{164}{40} - 0.2275 \cdot 13.75 = 0.9719$$

$$y = \frac{\sigma_{zy}}{\sigma_z^2} z + \bar{y} - \frac{\sigma_{zy}}{\sigma_z^2} \bar{z} \Rightarrow y = 62.2255z - 0.4063$$

$\Downarrow$ Deshacemos el cambio

$$\boxed{y = \frac{62.2255}{x} - 0.4063}$$

Si $x = 4 \Rightarrow y = 15.15$ minutos (Predicción)

f) ¿Qué modelo es más fiable?

Recta: $\eta^2_{Y/X} = r^2 = 0.6339$

Hipérbola: $\eta^2_{Y/X} = \frac{\sigma^2_{ey}}{\sigma^2_y} = \frac{60.4759}{104.6875} = 0.57768$

Es más adecuado el modelo lineal ya que su $\eta^2_{Y/X}$ (coef. determ. lineal) es mayor que el del hiperbólico