Grado en Ingeniería Informática Relación 1: Análisis descriptivo de datos. Regresión.

21 de marzo de 2019

1. Hemos recibido el encargo de desarrollar una aplicación para móvil que convierta archivos en formato *.jpg* a *.pdf*. Uno de los requisitos solicitados por el cliente es que, una vez elegido el archivo *.jpg* a convertir, se muestre en pantalla el tiempo estimado que se tardará en convertir el archivo. Para implementar tal funcionalidad, decidimos usar un modulo de regresión para predecir el tiempo en función del tamaño del archivo *.jpg*, tomando como base los datos recolectados en la fase de pruebas de la aplicación, reflejados en la siguiente tabla:

Tamaño archivo (MB)	20	10	50	17	5
Tiempo de conversión (segs)	1,8	1	5,3	1.9	0,6

a) Construye las dos rectas de regresión correspondientes a los datos en la tabla anterior.

b) ¿Podemos concluir que la regresión es adecuada para resolver el problema planteado, en base a los datos de la tabla?.

c) Usando la recta de regresión adecuada, calcula el tiempo de conversión estimado que mostraremos en pantalla, para un archivo de 30MB.

---

a) Construye las dos rectas de regresión correspondientes a los datos en la tabla anterior.

Para construir las dos rectas de regresión tendremos que sacar una serie de datos:

Media :

$$\overline{X} = \sum_{i=1}^{N}(x_i * y_j) = \frac{20+10+50+17+5}{5} = 20{,}4MB$$

$$\overline{Y} = \sum_{i=1}^{N}(y_i * n_i) = \frac{1,8+1+5,3+1,9+0,6}{5} = 2{,}12Seg$$

Covarianza:

$$\text{Sxy} = \frac{1}{N}\sum_{i=1}^{N} x_i{*}y_j{*}n_{ij} - (\overline{X}{*}\overline{Y}) = \frac{(20*1,8)+(10*1)+(50*5,3)+(17*1,9)+(5*0,6)}{5} - (20,4*2,12) = 26,012MB/Segs$$

Varianzas:

$$Sx^2 = (\frac{1}{N}\sum_{i=1}^{N} x_i^2 * n_i) - \overline{X}^2$$

$$Sx^2 = \frac{3314}{5} - 20{,}4^2 = 246{,}64MB_2$$

$$Sy^2 = (\frac{1}{N}\sum_{i=1}^{N} y_j^2 * n_j) - \overline{Y}^2$$

$$Sy^2 = \frac{1,8^2+1^2+5,3^2+1,9^2+0,6^2}{5} - 2{,}12^2 = 2,7656Seg_2$$

Regresión:

Recta de regresión X sobre Y: $(x - \overline{X}) = \frac{Sxy}{Sy^2}(y - 2{,}12)$

Recta de regresión Y sobre X: $(y - \overline{Y}) = \frac{Sxy}{Sx^2}(x - 20,4)$

b) ¿Podemos concluir que la regresión es adecuada para resolver el problema planteado, en base a los datos de la tabla?

Para saber si es adecuada, voy a estudiar el grado de correlación entre X e Y, ya que podremos saber si existe una buena correlación.

$$r = \frac{S_{xy}}{S_x * S_y} = \frac{26{,}012}{\sqrt[2]{246{,}64} * \sqrt[2]{2{,}765}} = 0,9960$$

Con lo que podemos ver que existe una buena relacion entre X e Y con lo que podemos concluir que la regresión es adecuada para resolver este problema

c) Usando la recta de regresión adecuada, calcula el tiempo de conversión estimado que mostraremos en pantalla, para un archivo de 30MB

$y = \overline{Y} + \frac{S_{xy}}{S_x^2}(x - \overline{X})$

$y = 2,12 + \frac{26,012}{246,64}(30 - 20,4) = 3,132468375 segs.$

---

2. Con objeto de determinar la relación entre el tiempo de respuesta (en segundos) de una determinada base de datos de consulta y el número de usuarios se han tomado 10 datos correspondientes a 2 semanas consecutivas, obteniéndose los siguientes resultados (tiempo — número de usuarios):

	Lunes	Martes	Miercoles	Jueves	Viernes
Semana1	4.32 — 15	7.14 — 18	9.21 — 20	9.71 — 20	15.39 — 26
Semana2	5.2 — 16	8.37 — 19	9.34 — 20	10.46 — 21	18.9 — 29

a) ¿Cuál de las dos variables es más homogénea?

b) Usando la recta de regresión adecuada, determinar el número de usuarios activos, si el tiempo de respuesta es de 12 segundos.

c) ¿Podemos concluir que la regresión es adecuada para resolver el problema planteado, en base a los datos de la tabla?

---

a) ¿Cuál de las dos variables es más homogénea?

Para saber cual es de las dos variables es más homogénea usaremos el coeficiente de variación de Pearson, para ello necesitaremos la media de ambas variables, su varianza y su desviación tipica.

Medias:

$\overline{X} = \frac{1}{N} \sum_{i=1}^{N} x_i * n_i$

$$\overline{Y} = \frac{1}{N} \sum_{i=1}^{N} x_j * n_j$$

Donde la media de x es:

$$\overline{X} = \frac{4{,}32+7{,}14+9{,}21+9{,}71+15{,}39+5{,}2+8{,}37+9{,}34+10{,}46+18{,}9}{10} = 9,804 Segundos$$

y la media de y es :

$$\overline{Y} = \frac{15+18+20+20+26+16+19+20+21+29)}{10} = 20,4 Usuarios$$

Varianza:

$$S_x^2 = (\frac{1}{n} \sum_{i=1}^{N} x_i^2 * n_i) - \overline{X}^2$$

$$S_y^2 = (\frac{1}{n} \sum_{i=1}^{N} x_j^2 * n_j) - \overline{Y}^2$$

Donde la Varianza de x es:

$S_x^2 = \frac{4{,}32^2+7{,}14^2+9{,}21^2+9{,}71^2+15{,}39^2+5{,}2^2+8{,}37^2+9{,}34^2+10{,}46^2+18{,}9^2}{10} - 9,804^2 = 17,537224 Segundos^2$

y la Varianza de y:

$$S_y^2 = \frac{15^2+18^2+20^2+20^2+26^2+16^2+19^2+20^2+21^2+29^2}{10} - 20,4^2 = 16,24 Usuarios^2$$

Desviación tipica:

$$S_x = \sqrt[2]{15,537224} = 4,187746888$$

$$S_y = \sqrt[2]{16,24} = 4,029888336$$

Coeficiente de variación de pearson:

Y por ultimo ya obtenemos el coeficiente de variación de Pearson,

$$V(x) = \frac{S_x}{\overline{X}} = \frac{4{,}187746888}{9{,}804} = 0,4271467654$$

$V(y) = \frac{S_y}{\overline{Y}} = \frac{4{,}029888336}{20{,}4} = 0,1975435459$

Con lo que podemos ver que la variable X es más homogénea que Y

b) Usando la recta de regresión adecuada, determina el número de usuarios activos, si el tiempo de respuesta es de 12 segundos.

Con lo sacado en el apartado A y nos dice 12 segundos, realizaremos la recta de regresión de Y sobre X.

Covarianza:

$$S_{xy} = (\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} x_i * y_j * n_{ij}) - \overline{X} * \overline{Y}$$

$S_{xy} = (\frac{(4{,}32*15)+(7{,}14*18)+(9{,}21*20)+(9{,}71*20)+(15{,}39*26)+(5{,}2*16)+(8{,}37*19)+(9{,}34*20)+(10{,}46*21)+(18{,}9)+(29)}{10}) - 9,804 * 20,4 = 16,8634 Segundos/Usuarios$

Usaremos la recta de regresión de Y sobre X:

$y - \overline{Y} = \frac{S_{xy}}{S_x^2}(x - \overline{X})$

$y = 20,4 + \frac{16{,}8634}{17{,}537224}(12 - 9,804) = 22,51162419 Usuarios$

c) ¿Podemos concluir que la regresión es adecuada para resolver el problema planteado, en base a los datos de la tabla?

$r = \frac{S_{xy}}{S_x * S_y}$

$r = \frac{16{,}8634}{4{,}187746888 * 4{,}029888336} = 0,9992443575$

Si ya que podemos ver existe un muy buena relación entre ambas variables

---

3. La siguiente tabla muestra los resultados de medir el tiempo que se tarda en transferir 5 archivos de distinto tamaño a través de una red:

Tamaño del archivo (KBytes)	300	500	120	600	400
Tiempo de transmisión (segundos)	1,1	1,9	0,3	2,1	1,5

a) Usando regresión lineal, determina el tiempo de transmisión de un archivo de 200 KBytes

b) Usando regresión lineal, determina el tamaño de un archivo que tarda 0,5 segundos en ser transferido

c) Determina si el ajuste es fiable o no

---

a) Usando regresión, determina el tiempo de transmisión de un archivo de 200 KBytes

$\overline{Y} sobre \overline{X}$

$(y - \overline{Y}) = \frac{S_{xy}}{S_x^2}(x - \overline{X})$ Recta de regresión

Medias:

$\overline{X} = \frac{1}{N}\sum_{i=1} N x_i * n_i = \frac{(300*1)+(500*1)+(120*1)+(600*1)+(400*1)}{5} = 384KB$

$\overline{Y} = \frac{1}{N}\sum_{j=1}^{N} y_j * n_j = \frac{(1,1*1)+(1,9*1)+(0,3*1)+(2,1*1)+(1,5*1)}{5} = 1,38Segundos$

Varianza:

$S_x^2 = (\frac{1}{N}\sum_{i=1}^{N}) x_i^2 * n_i) - \overline{X}^2 = (\frac{300^2+500^2+120^2+600^2+400^2}{5}) - 384^2 = 27424KB^2$

Covarianza:

$S_{xy} = (\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} x_i y_j nij) - \overline{XY} = \frac{(300*1,1)+(500*1,9)+(200*0,3)+(600*2,1)+(400*1,5)}{5} - (384*1,38) = 105,28KBSegundo$

Por lo tanto la recta de regresión es:

$y = 1,38 + \frac{105,28}{27424}(200 - 384) = 0,67 Segundos$

b) Usando regresión lineal, determina el tamaño de un archivo que tarda 0.5 segundos en ser transferido

$\overline{X} sobre \overline{Y}$

$x - \overline{X} = \frac{S_{xy}}{S_y^2}(y - \overline{Y})$

$S_y^2 = \frac{(1,1)^2+(1,9)^2+(0,3)^2+(2,1)^2+(1,5)^2}{5} - 1{,}38^2 = 0,4096 Segundo^2$

$x = 384 + \frac{105,28}{0,4096}(0,5 - 1,38) = 157,8125 KB$

c) Determina si el ajuste es fiable o no

$S_x = \sqrt[2]{27424} = 165,60$

$S_y = \sqrt[2]{0,4096} = 0,64$

$r = \frac{S_{xy}}{S_x S_y} = \frac{105,28}{165,60*0,64} = 0,9933574$

Es casi perfecta la predicción, por lo que si es fiable

---

4. Se han obtenido los siguientes datos sobre número de usuarios en función del número de fallos en los PCs del aula:

Número de usuarios	47	41	54	50	42
Número de fallos	5	4	6	5	3

a) Obtén la recta de regresión necesaria para predecir el número de usuarios en función del número de fallos

b) Determina cómo de fiable es el ajuste

---

a) Obtén la recta de regresión necesaria para predecir el número de usuarios en función del número de fallos

Media:

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} x_i n_i$$

$$\overline{X} = \frac{47+41+54+50+42}{5} = 46,8 Usuarios$$

$$\overline{Y} = \frac{1}{N}\sum_{j=1}^{N} y_j n_j$$

$$\overline{Y} = \frac{5+4+6+5+3}{5} = 4,6 Fallos$$

Covarianza:

$$S_{xy} = (\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} x_i y_j n_{ij}) - \overline{XY}$$

$$S_{xy} = \frac{(47*5)+(41*4)+(54*6)+(50*5)+(42*3)}{5} - 46,8 * 4,6 = 4,52$$

Varianza:

$$S_x^2 = (\frac{1}{N}\sum_{i=1}^{N} x_i^2 * n_i) - \overline{X}^2$$

$$S_x^2 = (\frac{(47^2)+(41^2)+(54^2)+(50^2)+(42^2)}{5}) - 46{,}8^2 = 23,76$$

$$S_y^2 = (\frac{1}{N}\sum_{j=1}^{N} y_j^2 * n_j) - \overline{Y}^2$$

$$S_y^2 = (\frac{(5^2)+(4^2)+(6^2)+(5^2)+(3^2)}{5}) - 4,6^2 = 1,04$$

La recta de regresión para predecir el número de usarios seria:

$$(x - \overline{X}) = \frac{S_{xy}}{S_y^2}(y - \overline{Y})$$

b) Determina cómo de fiable es el ajuste

Tendremos que sacer la desviación tipica:

$S_x = \sqrt[2]{23,76} = 4,874423043$

$S_y = \sqrt[2]{1,04} = 1,019803903$

$r = \frac{S_{xy}}{S_x S_y}$

$r = \frac{4,52}{4,874423043*1,019803903} = 0,9092819014$

Por lo que tenemos un ajuste muy fiable

---

5. La siguiente tabla muestra los resultados de medir, en 5 instantes durante un día, el número de usuarios conectados a un servidor Linux y la cantidad de RAM (en GB) disponible en el sistema:

Número de usuarios	5	3	2	6	1
RAM disponible	1,1	2,6	3	1	3,5

a) Usando regresión lineal, determina la cantidad de RAM disponible si hay 4 usuarios conectados.

b) Usando regresión lineal, determina el número de usuarios conectados si la RAM disponible es de 1.5 GB.

c) Determina cómo de fiable es el ajuste

---

a) Usando regresión lineal, determina la cantidad de RAM disponible si hay 4 usuarios conectados.

Media:

$\overline{X} = \frac{1}{N} \sum_{i=1}^{N} x_i n_i$

$\overline{X} = \frac{5+3+2+6+1}{5} = 3,4 Usuarios$

$\overline{Y} = \frac{1}{N} \sum_{j=1}^{N} y_j n_j$

$\overline{Y} = \frac{1,1+2,6+3+1+3,5}{5} = 2,24 GB RAM$

Covarianza:

$$S_{xy} = (\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} x_i y_j n_{ij}) - \overline{XY}$$

$$S_{xy} = \frac{(5*1,1)+(3*2,6)+(2*3)+(6*1)+(1*3,5)}{5} - (3,4*2,24) = -1,856$$

Varianzas y desviaciones tipicas

$S_x^2 = \frac{5^2+3^2+2^2+6^2+1^2}{5} - 3,4^2 = 1,854723699 Usuarios^2$

$S_y^2 = \frac{1,1^2+2,6^2+3^2+1^2+3,5^2}{5} - 2,24^2 = 1,013114011 Ram^2$

$S_x = \sqrt[2]{3,44} = 1,854723699$

$S_y = \sqrt[2]{1,0264} = 1,013114011$

Recta de regresión:

$y = 2,24 + \frac{-1,856}{3,44}(4 - 3,4) = 1,912470588 RAM$

b) Usando regresión lineal, determina el número de usuarios conectados si la RAM disponible es de 1.5 GB.

$x = 3,4 + \frac{-1,856}{1,0264}(1,5 - 2,24) = 4,738113796 Usuarios$

c) Determina cómo de fiable es el ajuste

$r = \frac{S_{xy}}{S_x * S_y} = \frac{-1,856}{1,854723699*1,013114011} = -0,9877349681$

Con lo cual podemos decir que tenemos una buen predición, teniendo casi una correlación perfecta negativa, ya que viendo su covarianza que es negativa, significa que las rectas son decrecientes.

---

6. La siguiente tabla muestra los resultados de medir la cantidad de datos (X, en KB) tecleada por cinco operadores en un día de trabajo y el tamaño de su monitor (Y, en pulgadas).

X	150	175	210	230	276
Y	15	17	19	21	26

a) Usando regresión lineal, determina la cantidad de datos que se puede predecir para un operador cuyo monitor sea de 24 pulgadas.

b) Determina cómo de fiable es el ajuste.

---

a) Usando regresión lineal, determina la cantidad de datos que se puede predecir para un operador cuyo monitor sea de 24 pulgadas.

Media:

$\overline{X} = \frac{1}{N} \sum_{i=1}^{N} x_i n_i$

$\overline{X} = \frac{150+175+210+230+276}{5} = 208,2KB$

$\overline{Y} = \frac{1}{N} \sum_{j=1}^{N} y_j n_j$

$\overline{Y} = \frac{15+17+19+21+26}{5} = 19,6 pulgadas$

Covarianza:

$S_{xy} = (\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} x_i y_j n_{ij}) - \overline{XY}$

$S_{xy} = \frac{(150*15)+(175*17)+(2210*19)+(230*21)+(276*26)}{5} - 208,2*19,6 = 163,48 KB/Pulgada$

Varianza:

$S_x^2 = (\frac{1}{N} \sum_{i=1}^{N} x_i^2 * n_i) - \overline{X}^2$

$S_x^2 = (\frac{(150^2)+(175^2)+(210^2)+(230^2)+(276^2)}{5}) - 208,2^2 = 1912,96$

$$S_y^2 = (\frac{1}{N}\sum_{j=1}^{N} y_j^2 * n_j) - \overline{Y}^2$$

$$S_y^2 = (\frac{(15^2)+(17^2)+(19^2)+(21^2)+(26^2)}{5}) - 19,6^2 = 14,24$$

La recta de regresión para predecir el número de datos es:

$$(x - \overline{X}) = \frac{S_{xy}}{S_y^2}(y - \overline{Y})$$

$$x = 208,2\frac{163,48}{14,24}(24 - 19,6) = 10537,11258KB$$

b) Determina cómo de fiable es el ajuste.

$$S_x = \sqrt[2]{1912,96} = 43,73739819$$

$$S_y = \sqrt[2]{14,24} = 3,773592453$$

$$r = \frac{S_{xy}}{S_x S_y}$$

$$r = \frac{163,48}{43,73739819*3,773592453} = 0,9905050402$$

Por lo que podemos decir que es fiable el ajuste

---

7. Los siguientes datos representan el número de accesos (X) a un servidor remoto y el número de cortes en la comunicación (Y) sufridos durante diez días consecutivos

X	170	160	210	140	180	240	160	140	210	230
Y	8	7	11	5	9	12	8	6	10	15

a) Comprueba si las variables X e Y son independientes

b) Calcula, usando una recta de regresión, el número de cortes en la comunicación durante un día en el que se producen 150 accesos. ¿Es bueno el ajuste?

---

a) Comprueba si las variables X e Y son independientes

media:

$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} x_i n_i$

$\overline{X} = \frac{170+160+210+140+180+240+160+140+210+230}{10} = 184 Mensajes$

$\overline{Y} = \frac{1}{N}\sum_{j=1}^{N} y_j n_j$

$\overline{Y} = \frac{8+7+11+5+9+12+8+6+10+15}{10} = 8,9 MensajesdeSPAM$

Varianza:

$S_x^2 = (\frac{1}{N}\sum_{i=1}^{N} x_i^2 * n_i) - \overline{X}^2$

$S_x^2 = (\frac{(170^2)+(160^2)+(210^2)+(140^2)+(180^2)+(240^2)+(160^2)+(140^2)+(210^2)+(230^2)}{10}) - 184^2 = 1184$

$S_y^2 = (\frac{1}{N}\sum_{j=1}^{N} y_j^2 * n_j) - \overline{Y}^2$

$S_y^2 = (\frac{(8^2)+(7^2)+(11^2)+(5^2)+(9^2)+(12^2)+(8^2)+(6^2)+(10^2)+(13^2)}{10}) - 8,9^2 = 6,09$

$r = \frac{82,4}{34,40930107 * 2,467792536} = 0,97038196$

Por lo que tienen una buena relación entra ambas variables

b) Calcula, usando una recta de regresión, el número de cortes en la comunicación durante un día en el que se producen 150 accesos. ¿Es bueno el ajuste?

$y = 8,9 + \frac{82,4}{6,09}(150 - 184) = 6,53\overline{378}$

Si es buen ajuste ya que tienen una buena correlación ambas variables

---

8. Los siguientes datos representan el número total de mensajes de correo

electrónico (X) manejados por un servidor y el número de mensajes tipo SPAM (Y) correspondiente a diez días consecutivos:

X	170	160	210	140	180	240	160	140	210	230
Y	8	7	11	5	9	12	8	6	10	15

a) ¿Cuál de las dos variables es más dispersa?.

b) Calcula, usando una recta de regresión, el número total de mensajes manejados durante un día en el que se recibieron 13 mensajes tipo SPAM. ¿Es bueno el ajuste realizado mediante la recta de regresión?

---

a) ¿Cuál de las dos variables es más dispersa?.

Para ver el grado de dispersión de ambas variables tenemos que fijarnos en su coeficiente de variacion de Pearson:

media:

$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} x_i n_i$

$\overline{X} = \frac{170+160+210+140+180+240+160+140+210+230}{10} = 184 Mensajes$

$\overline{Y} = \frac{1}{N}\sum_{j=1}^{N} y_j n_j$

$\overline{Y} = \frac{8+7+11+5+9+12+8+6+10+15}{10} = 9,1 Mensajes de SPAM$

Varianza:

$S_x^2 = (\frac{1}{N}\sum_{i=1}^{N} x_i^2 * n_i) - \overline{X}^2$

$S_x^2 = (\frac{(170^2)+(160^2)+(210^2)+(140^2)+(180^2)+(240^2)+(160^2)+(140^2)+(210^2)+(230^2)}{10}) - 184^2 = 1184$

$$S_y^2 = (\frac{1}{N}\sum_{j=1}^{N} y_j^2 * n_j) - \overline{Y}^2$$

$$S_y^2 = (\frac{(8^2)+(7^2)+(11^2)+(5^2)+(9^2)+(12^2)+(8^2)+(6^2)+(10^2)+(15^2)}{10}) - 9,1^2 = 8,09$$

Desviación tipica:

$$S_x = \sqrt[2]{1184} = 34,40930107$$

$$S_y = \sqrt[2]{8,09} = 2,844292531$$

Y por ultimo ya obtenemos el coeficiente de variación de Pearson:

$$V(x) = \frac{S_x}{\overline{X}} = \frac{34,40930107}{184} = 0,3845054617$$

$$V(y) = \frac{S_y}{\overline{Y}} = \frac{2,844292531}{9,1} = 0,4301645784$$

Por lo que la variable Y es mas dispersa que la variable X

b) Calcula, usando una recta de regresión, el número total de mensajes manejados durante un día en el que se recibieron 13 mensajes tipo SPAM. ¿Es bueno el ajuste realizado mediante la recta de regresión?

$$S_{xy} = (\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} x_i y_j n_{ij}) - \overline{XY}$$

$$S_{xy} = \frac{(170*8)+(160*7)+(210*11)+(140*5)+(180*9)+(240*12)+(160*8)+(140*6)+(210*10)+(230*15)}{10} - 184 * 9,1 = 91,6$$

$$x = \overline{X} + \frac{S_{xy}}{S_y^2}(y - \overline{Y})$$

$$y = 184 + \frac{91,6}{0,3125596188}(13 - 9,1) = 1326 Mensajes manejaos$$

Para ver si fue un buen ajuste estudiaremos

$$r = \frac{S_{xy}}{S_x S_y}$$

$$r = \frac{91,6}{34,40930107 * 2,844292531} = 0,9359342958$$

Por lo que podemos ver que tiene un buena relación ambas variables por lo que es buen ajuste

---

9. En una empresa los empleados se clasifican en tres categorías: técnicos, especialistas y administrativos. El salario medio mensual y la varianza de los salarios de cada categoria en el mes de Diciembre de 2015 son los que aparecen en el siguiente cuadro:

Categoría	Salario medio mensual (euros)	Varianza de los salarios
Técnicos	2500	10
Especialistas	2000	25
Administrativos	1500	15

I.- ¿En qué grupo de empleados los salarios son más homogéneos?

II.- En la discursión para fijar los salarios de 2016 han sido propuestas dos alternativas:

A: La elevación de todos los salarios en un 5 %

B: La elevación de todos los salarios en 50 euros mensuales.

Calcula los salarios medios que resultan de aplicar las dos alternativas y la dispersión relativa en cada caso. ¿Qué alternativa provoca menos dispersión en los salarios de los grupos?

I.- ¿En qué grupo de empleados los salarios son más homogéneos?

sabiendo que:

$\overline{x} = 2500$ y $S_x^2 = 10$

$\overline{y} = 2000$ y $S_y^2 = 25$

$\overline{z} = 1500$ y $S_z^2 = 15$

Sacamos la desviación tipica:

$S_x = \sqrt[2]{10} = 3,16227766$

$S_y = \sqrt[2]{25} = 5$

$S_z = \sqrt[2]{15} = 3,872983346$

Y podemos sacar el coeficiente de variación de Pearson:

$CV = \frac{S_x}{\overline{x}}$

$CV(x) = \frac{3,16227766}{2500} = 0,00126491106$

$CV(y) = \frac{5}{2000} = 0,0025$

$CV(z) = \frac{3,872983346}{1500} = 0,00258198889$

Con la conclusión que podemos sacar es que el grupo de los Administrativos es el más homogéneo.

II.- En la discursión para fijar los salarios de 2016 han sido propuestas dos alternativas:

A: La elevación de todos los salarios en un 5 %

B: La elevación de todos los salarios en 50 euros mensuales.

Calcula los salarios medios que resultan de aplicar las dos alternativas y la dispersión relativa en cada caso. ¿Qué alternativa provoca menos dispersión en los salarios de los grupos?

Aumentado en un 5 %

$\overline{x} = 2625$ y $S_x^2 = 10,5$

$\overline{y} = 2100$ y $S_y^2 = 26,25$

$\overline{z} = 1575$ y $S_z^2 = 15,75$

Sacamos la desviación tipica:

$S_x = \sqrt[2]{10} = 3,240370349$

$S_y = \sqrt[2]{25} = 5,123475383$

$S_z = \sqrt[2]{15} = 3,968626967$

Y podemos sacar el coeficiente de variación de Pearson:

$CV = \frac{S_x}{\overline{x}}$

$CV(x) = \frac{3,240370349}{2625} = 0,0012344268$

$CV(y) = \frac{5,123475883}{2100} = 0,00243975018$

$CV(z) = \frac{3,968626967}{1575} = 0,00251976315$

Sumando 50 €:

$\overline{x} = 2550$ y $S_x^2 = 60$

$\overline{y} = 2050$ y $S_y^2 = 75$

$\overline{z} = 1550$ y $S_z^2 = 65$

Sacamos la desviación tipica:

$S_x = \sqrt[2]{60} = 7,745966692$

$S_y = \sqrt[2]{75} = 8,660254038$

$S_z = \sqrt[2]{65} = 8,062257748$

Y podemos sacar el coeficiente de variación de Pearson:

$CV = \frac{S_x}{\overline{x}}$

$CV(x) = \frac{7,745966692}{2550} = 0,00126491106$

$CV(y) = \frac{8,660254038}{2050} = 0,0025$

$CV(z) = \frac{8,062257748}{1550} = 0,00520145661$

Atendiendo al Coeficiente de variación de Pearson, podemos decir que aumentar en un 5 % el salario tiene un menor dispersión de los datos respecto a la media aritmetica.

10. Se han estudiado las califiaciones de 100 alumnos en dos asignaturas: Matemáticas y Estadística, obteniéndose los siguientes datos:

$\overline{x} = 110, \overline{y} = 2{,}5, S_x = 10, S_y = 0{,}5, r = 0{,}85.$

a) ¿Qué nota se puede predecir para un alumno, que ha obtenido 125 puntos en Matemáticas, en la asignatura de Estadística?

b) ¿Se puede decir que aquellos alumnos que obtienen mayor calificación en Matemáticas sean los mismos que obtienen mayor calificación en Estadística?

c) ¿Cuál es la ecuación de la recta de regresión de X sobre Y?

---

a) ¿Qué nota se puede predecir para un alumno, que ha obtenido 125 puntos en Matemáticas, en la asignatura de Estadística?

Sabiendo que: $r = \frac{S_{xy}}{S_x S_y}$

podemos sacar la covarianza: $S_{xy=0{,}85*10*0{,}5=4{,}25}$

Ahora ya podemos predecir la nota del alumno que ha obtenido 125 puntos en Matemáticas, y con bastante buena prediccion

$y = 2,5 + \frac{4{,}25}{100}(125 - 110) = 2,659375$

b) ¿Se puede decir que aquellos alumnos que obtienen mayor calificación en Matemáticas sean los mismos que obtienen mayor calificación en Estadística?

Si, ya que existe una buena relación entre ambas variables

c) ¿Cuál es la ecuación de la recta de regresión de X sobre Y?

$x = \overline{X} + \frac{S_{xy}}{S_x^2}(y - \overline{Y})$