Cátedra ESTADISTICA

TRABAJOS PRÁCTICOS

2020

Facultad de Ingeniería

Universidad Nacional de La Patagonia S. J. B.

Comodoro Rivadavia

TRABAJO PRÁCTICO Nº 7

REGRESIÓN Y CORRELACIÓN

PRE - REQUISITOS:

Se requiere lectura previa y manejo conceptual de los siguientes conceptos:

- **Estadística Descriptiva e Inferencial.**
- **Población y Muestra. Parámetros y Estimadores.**
- **Variables aleatorias. T. C. L.**
- **Distribuciones de Probabilidad en Inferencia.**
- **Estimación puntual y por Intervalo de Confianza.**
- **Dócimas.**

CONSIGNA PARTICULAR:

Se recomienda atender especialmente a cuáles conceptos aprehendidos con anterioridad son empleados y cómo se relacionan.

Tenga en cuenta que en este momento debería ser capaz de razonar estadísticamente, hacer inferencias y concluir con la terminología específica adecuada.

RECUERDE QUE YA NO SERÁ NECESARIO ACLARAR EN TODO MOMENTO QUE SE DEBEN INTERPRETAR LOS RESULTADOS NUMÉRICOS Y QUE SE ESPERA LA JUSTIFICACIÓN DE ANÁLISIS Y PROCEDIMIENTOS.

EJERCICIOS:

1.- . Qué tipo de análisis realizaría con los siguientes pares de variables? Justifique su respuesta, aclarando su objetivo.

a) Se hace un estudio sobre la antigüedad en años de los automotores de una marca determinada y se piensa que la cantidad de nafta que consumen cada 100 km depende de la antigüedad de los vehículos de esa marca.

b) Se toma una muestra de 15 personas y en cada una de ellas se mide el tamaño del perímetro encefálico y se observa el éxito en la vida.

c) Se tiene una distribución estadística bidimensional que representa el precio del kg de pan en $ y el consumo mensual en kg.

d) Se tienen datos sobre la velocidad de un río y la profundidad en distintos puntos del mismo. Se desea analizar:

*)si la velocidad está relacionada en forma directa con la profundidad.

**)si existe relación entre las variables y cuál es la fuerza de esta relación.

2. - En una investigación, los analistas de costos tratan de predecir el consumo mensual de agua de una planta química como una función de la producción mensual, para lo cual se cuenta con los siguientes datos:

Producción mensual	Consumo de agua en m^3
10	7.5
19	9
20	14
29	16

¿Qué función propondría usted? ¿Por qué?

3. -. Una compañía de productos químicos desea estudiar los efectos que el tiempo de extracción tiene en la eficiencia de una operación de extracción, obteniéndose los datos que aparecen en la siguiente tabla:

Tiempo de Extracción (minutos)	Eficiencia de Extracción %
57	27
64	45
80	41
46	19
62	35
72	39
52	19
77	49
57	15
68	31

a) Dibuje un diagrama de dispersión para verificar que una línea recta se ajustará relativamente bien a los datos, bosqueje una línea recta “a ojo”, y con ella prediga en forma aproximada la eficiencia en la extracción que puede esperarse cuando el tiempo de extracción es de 55 minutos.

b) Ajuste una línea recta a los datos dados con el método de los mínimos cuadrados y utilícela para predecir ahora concretamente la eficiencia de extracción que puede esperarse cuando el tiempo de extracción es de 55 minutos.

4. - Este ejercicio tiene algunos ítem resueltos, a fin de que en este momento aplique los conceptos aprehendidos, completando lo que sea necesario y analizando y discutiendo las cuestiones que se le presentan.

Para determinar la relación que existe entre el esfuerzo normal y la resistencia al corte del suelo, se llevó a cabo un experimento con una caja de esfuerzo cortante, obteniéndose los siguientes resultados:

Esfuerzo Normal	11	13	15	17	19	21
Resistencia	15.2	17.7	19.3	21.5	13.9	25.4
al corte	14.8	18.3	18.7	19.9	22.9	24.3
(kN/m²)	17.3			21.8	24.1	26.9

Datos: **Σx = 260** **Σx² = 4424** **Σy = 322** **Σy² = 6710.92** **Σxy = 5398.4**

a) Construya el dispersograma. ¿Cuántos pares ha observado? ¿Significa algo la suma: 11 + 13 + 15 + 17 + 19 + 21 = 96? ¿La usará para algo?¿Por qué?
COMPLETE

b) Halle la recta de regresión estimada e interprete.
(Debería encontrar “0,834” y “6,578”, aproximadamente)
COMPLETE

c) Pruebe la hipótesis que crea más importante para decidir si continúa con el análisis del problema de regresión. Concluya e interprete.

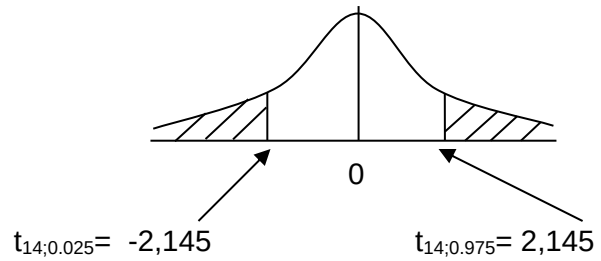
1) Ho : β =0

H_1 : $\beta \neq 0$

2) $\alpha = 0.05$

3) $v.p.: t_{n-2;} = \dfrac{b-\beta}{Sb}$

4)



R.D. : Rechazo Ho si tcal $\geq$ 2,145 ò tcal $\leq$ -2,145

No rechazo Ho si –2,145 < tcal < 2,145

5)

$$S^2e = \frac{1}{n-2}\left\{\left\{\sum y^2 - \frac{\left(\sum y\right)^2}{n} - b\left[\sum xy - \left(\sum x \sum y\right)/n\right]\right\}\right.$$

$$S^2_b = \frac{S_e^2}{\sum (x-\bar{x}^2)^2} = \frac{S_e^2}{\sum x^2 - (\sum x)^2/n}$$

S^2e = 6,59 Se = 2,567 S_b = 0.1820

tcal = (0.834 –0)/0.1820

tcal = 4,58

6) Como tcal es > 2,145 rechazo Ho

Conclusión: Con un nivel de significación del 5 % tengo evidencias suficientes para suponer que la verdadera pendiente de la recta de regresión que explica la variación de la resistencia al corte en función del esfuerzo normal es distinta de cero, o sea, existe regresión lineal entre las variables mencionadas.

d) Pruebe si la pendiente difiere significativamente de una pendiente predicha en forma teórica igual a 1. Interprete.

COMPLETE

e) Obtenga un intervalo de confianza del 95% para α e interprete.

$$P\left(\alpha - t_{n-2;\alpha/2} Sa < \alpha < \alpha + t_{n-2;\alpha/2} Sa\right) = 0.95$$

$$Sa = Se\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x-\bar{x})^2}}$$

Sa = 3.0248

$$\left(6{,}578 - 2{,}145(3{,}0248) < \alpha < 6{,}578 + 2{,}145(3{,}0248)\right)$$

(0,0915< α < 13.066)

Con una confianza de 95 %, podría decir que el intervalo (0,09; 13,07) encerraría al verdadero valor de la ordenada al origen de la recta de regresión entre las variables esfuerzo normal y resistencia al corte. Esto es, con una confianza de 95 %, podría decir que el intervalo (0,09 ; 13) encerraría al verdadero promedio de la resistencia al corte, para un valor “cero” del esfuerzo normal, si esto tiene sentido.

f) Obtenga un intervalo de confianza del 95% para β e interprete.

COMPLETE

g) Estime el valor medio de la resistencia al corte para x = 16 kN/m^2, luego obtenga un intervalo de confianza con 1 - α = 0,95. Interprete.

h) Suponiendo que la recta de regresión puede extrapolarse a x = 25 kN/m^2, determine la estimación del valor medio de la resistencia al esfuerzo cortante para ese valor, así como una estimación por intervalo de confianza, con 1 - α = 0,95. ¿Tiene sentido la estimación?

COMPLETE

i) ¿En qué condiciones es válido calcular e interpretar “r”? Asuma las condiciones necesarias y hágalo.

COMPLETE

5. - .La siguiente tabla indica cuántas semanas trabajó una muestra de seis personas en una estación de inspección de automóviles y el número de unidades que cada uno inspeccionó entre el medio día y las 2 P.M. en un día cualquiera:

Número de semanas empleadas "x" por cada persona	Número de automóviles inspeccionados "y"
2	13
7	21
9	23
1	14
5	15
12	21

a) Resuelva las ecuaciones normales para calcular la recta de mínimos cuadrados que le permitan predecir el valor de y en función de x. Utilice los valores que se le dan más adelante.

b) Con el resultado de la parte a) calcule cuántos automóviles puede esperarse que inspeccione alguien que ha estado trabajando en la estación de inspección durante 8 semanas en un período determinado de 2 horas.

c) Pruebe la hipótesis nula: β=1,2 contra la hipótesis alternativa: β<1,2 con un nivel de significación de 0,05.

d) Encuentre un intervalo con un nivel de confianza de 95% para el número promedio de automóviles que en el período determinado inspecciona una persona que ha estado trabajando en la estación de inspección por un periodo de 8 semanas.

e) Encuentre los límites de predicción de 95% para el número de automóviles a inspeccionar por una persona que trabaje en la estación de inspección durante 8 semanas.

f) Calcule el coeficiente de determinación e interprete.

g)Calcule el coeficiente de correlación.

g.1) Tiene sentido calcularlo? ¿Porqué?

g.2) Interprete.

Cálculos:

$\Sigma x = 36$ $\Sigma x^2 = 304$ $\Sigma y = 107$ $\Sigma y^2 = 2001$ $\Sigma xy = 721$

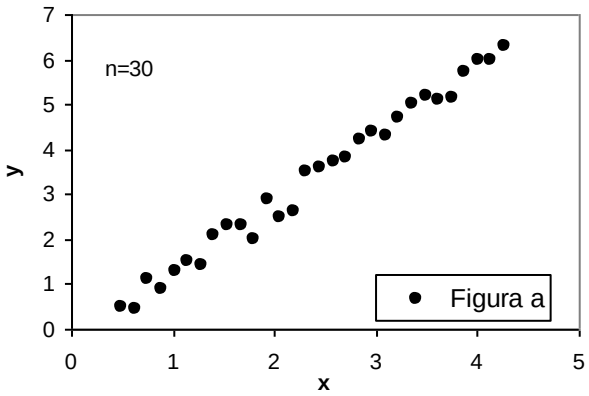
$\sum(x-\bar{x})^2 = 88$ $\sum(x-\bar{x})(y-\bar{y}) = 79$

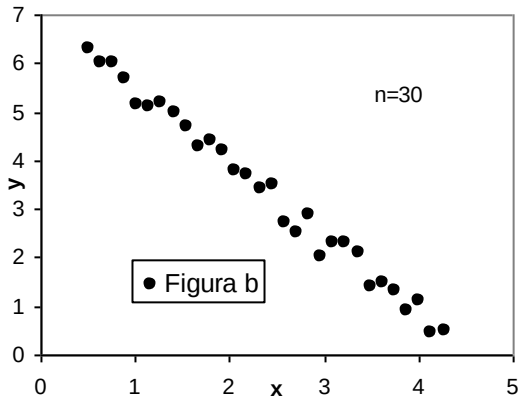
$\sum(y-\bar{y})^2 = 92{,}83$ $b^2 \sum(x-\bar{x}) = 70{,}92$

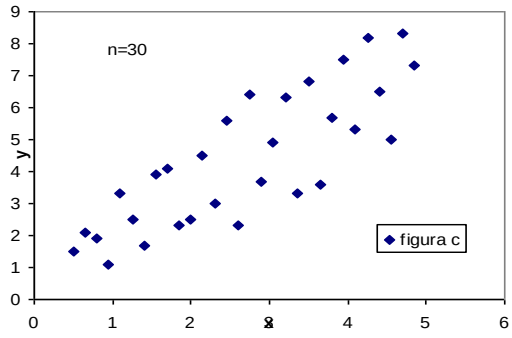
$b \sum(x-\bar{x})(y-\bar{y}) = 70{,}92.$

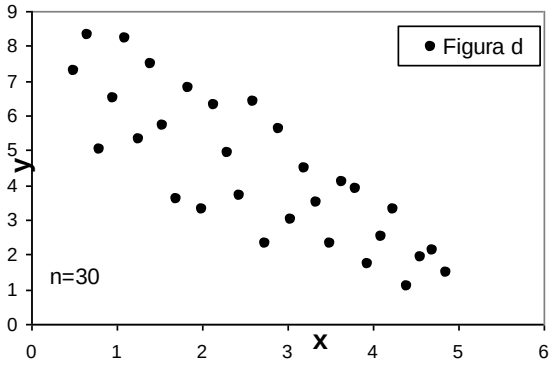
6. - Dados los siguientes gráficos de dispersión entre las variables “x” e “y”, asignar los valores de "r" dados a cada uno de ellos, según su criterio:

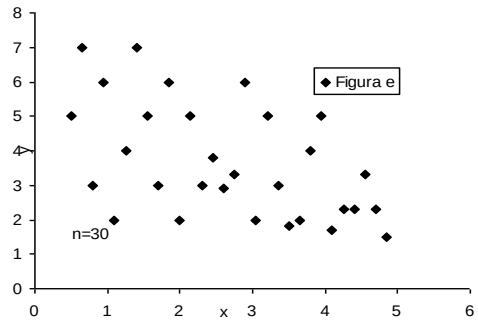
r	0.6	0.9	0	-0.6	-0.9
Figura Nº	.....	.....	.....	.....	.....

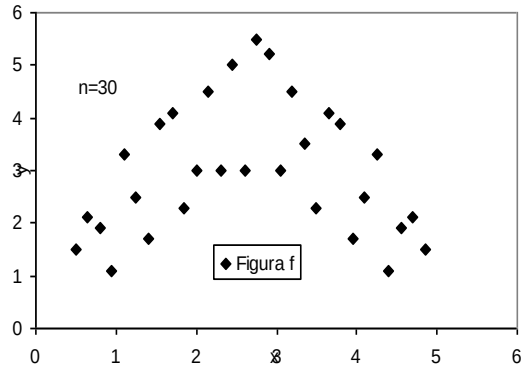












7. - Observe las figuras mostradas a continuación que corresponden a los mismos pares de datos: ¿qué diferencias nota en cuanto al grado de asociación de las variables? ¿Por qué sucede esto? Describa posible causas (pero no invente, sólo observe y razone.
Se tienen datos correspondientes a empresas A y B mezclados indiscriminadamente o bien separados en dos estratos, según algún factor de estratificación.

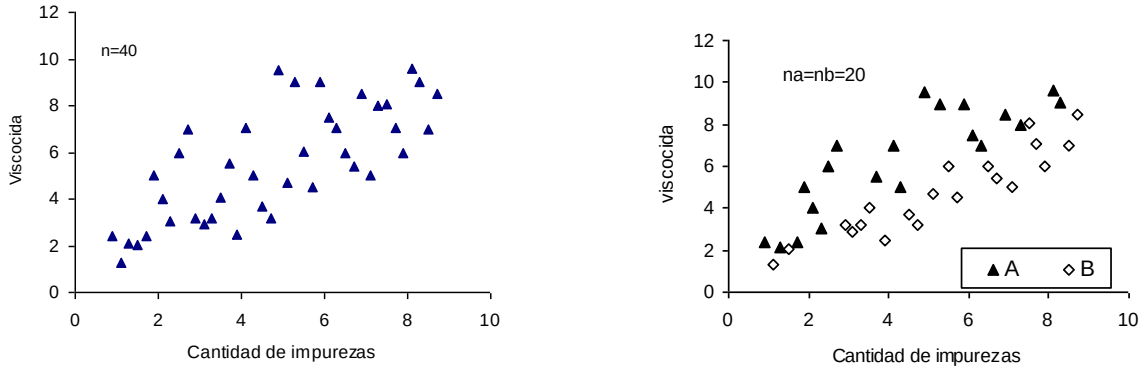


Figura 1: Estratificación en un diagrama de dispersión

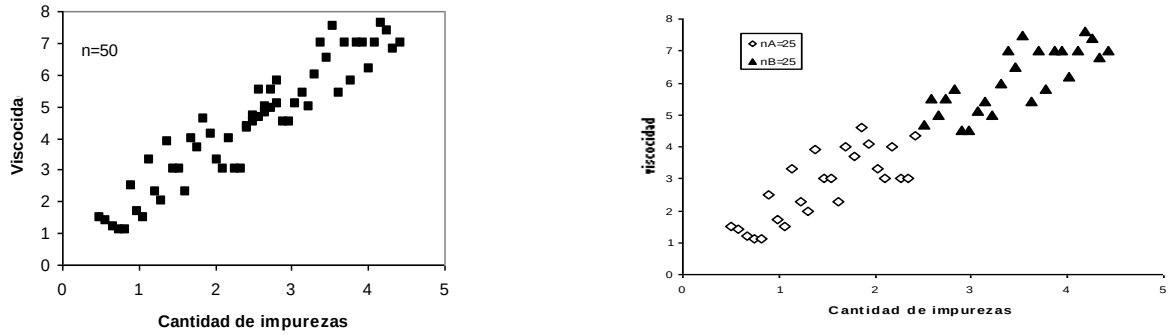


Figura 2: Estratificación en un diagrama de dispersión

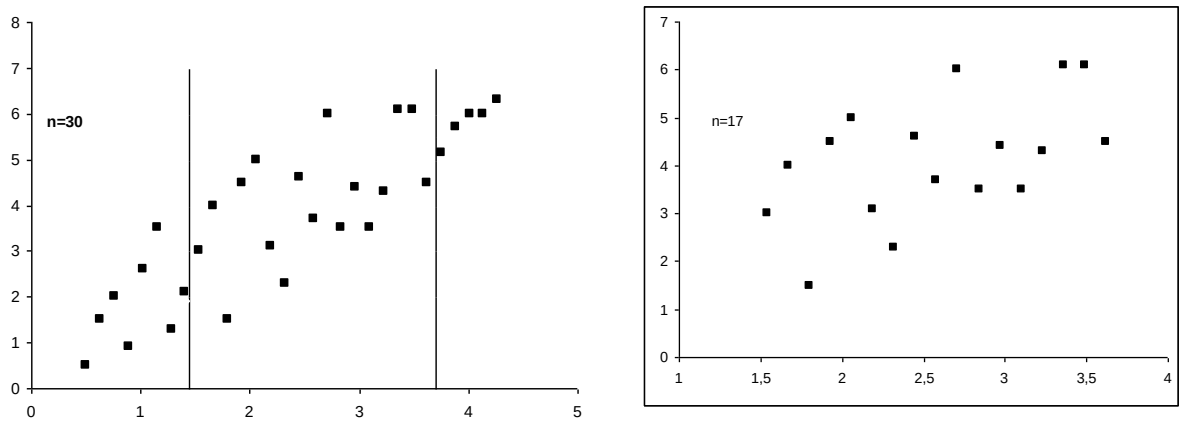


Figura 3: Efecto del rango de la variable

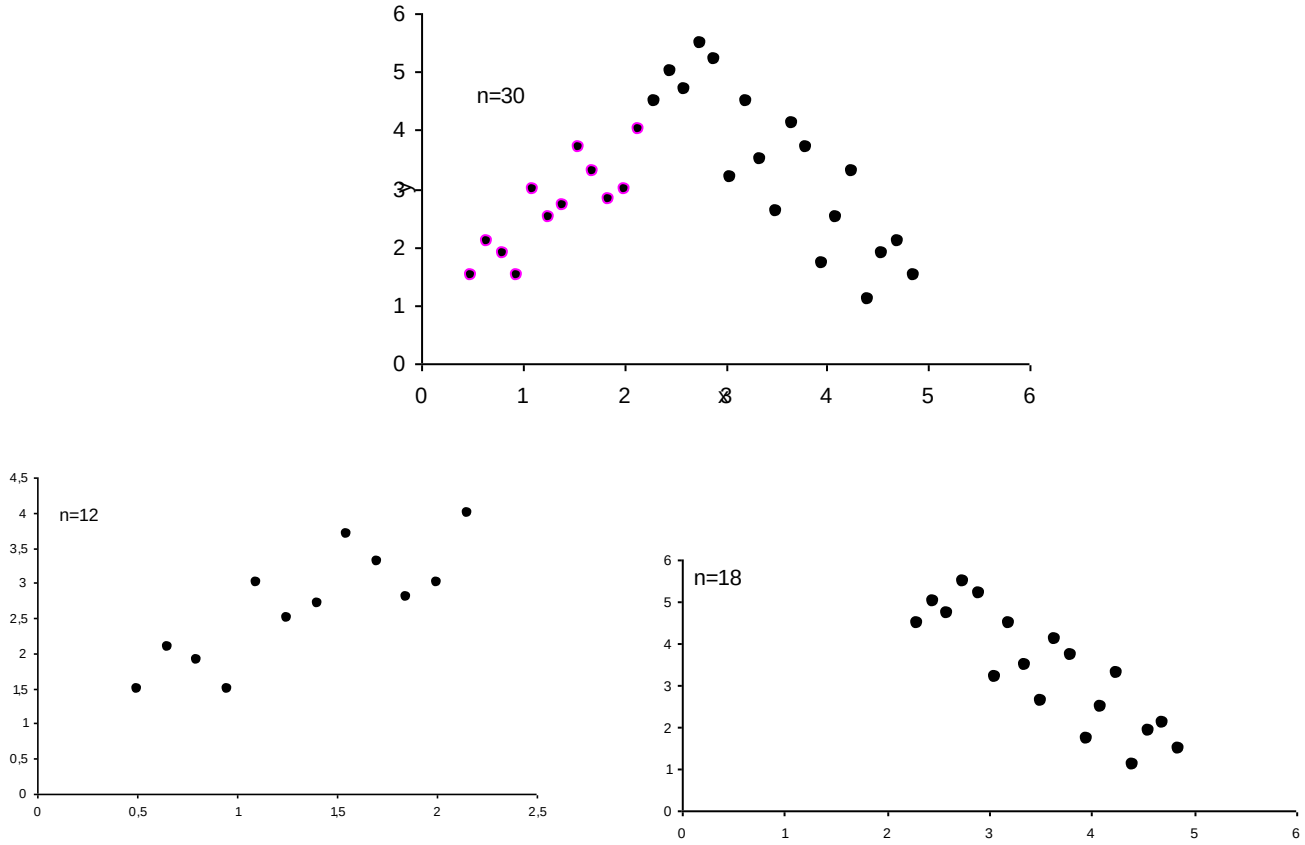


Figura 4: Efecto del rango de la variable

8.- La siguiente tabla proporciona el número de personas empleadas en laboratorios, y su salario mensual desde 1961 a 1966. Calcule la asociación y establezca en forma aproximada si es positiva o negativa, débil o fuerte. (interprete). Docime usando un nivel de significación de 1%.

Año	Empleados en millones	Salarios
1961	6.9	$151
1962	6.7	$155
1963	6.5	$159
1964	6.1	$162
1965	5.6	$171
1966	5.2	$185

9. - Retome el Ejercicio Nº 4 y realice las siguientes actividades:

a) Construya la tabla de análisis de varianza para el problema de regresión, y repita lo pedido en el ítem c). Luego compare los resultados.

b) Calcule R^2 e interprete.

c) ¿Puede calcular "r"? Si tiene sentido hágalo.

10. - Retome el Ejercicio Nº 5 y realice las siguientes actividades:

a) Construya la tabla de análisis de varianza para el problema de regresión y pruebe la hipótesis de interés con nivel de significación de 5 %.

b) A partir de la tabla de ANOVA calcule R^2 e interprete.

c) Si puede considerar un análisis de correlación, calcule "r".

11. - Retome el ítem c del Ejercicio Nº 4 y / o el ítem a del ejercicio Nº 9 y realice las siguientes actividades:

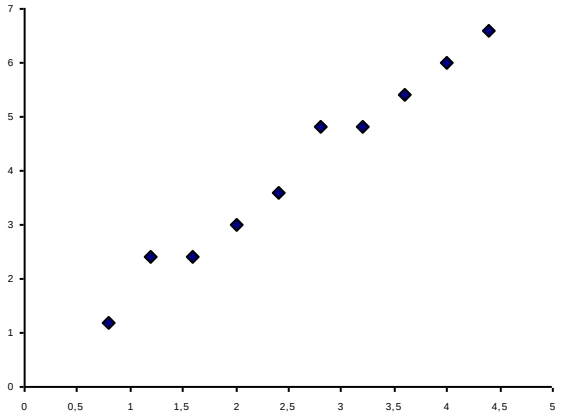
a) Diga si cree coherente encontrar un I.C. para la pendiente poblacional que tenga límite superior negativo. ¿Por qué?
b) Diga si cree coherente encontrar un I.C. para la pendiente poblacional que tenga límite inferior negativo y límite superior positivo. ¿Por qué?

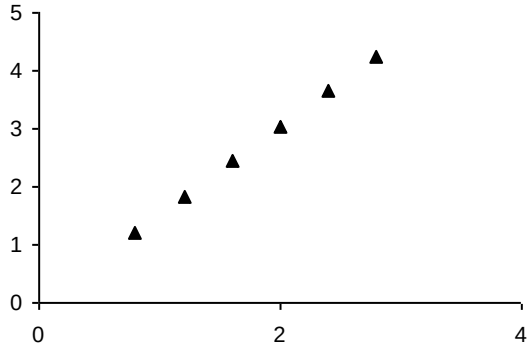
PRUEBA DE CONCEPTOS

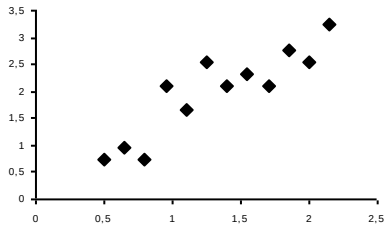
1) Dado y = 6 + 5x. Señale sin calcular cuáles intervalos podrían tener sentido:

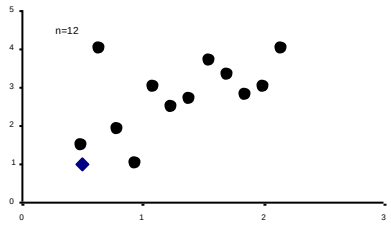
□ 5.5 < β < 6.5 □ -5 < β <-4 □ 4.8 < β <5.2 □ -5 < β < 5

2) Asigne los valores 0.43; 0.72; 0.97 y 1 a los coeficientes de correlación de las siguientes distribuciones bidimensionales:









....

3) Señale con una cruz la respuesta correcta. El coeficiente de correlación es un valor que:

□ es igual a 1. □ está entre 0 y 1. □ está entre 0 y -1. □ está entre –1 y 1.
□ es menor que -1. □ es mayor que -1. □ Ninguna de las anteriores.

AUTOEXÁMEN

1) ¿Cuál es el modelo de regresión simple y cuáles son sus parámetros?

2) Está de acuerdo o no con la siguiente afirmación: "si no existe una relación lineal, el coeficiente de correlación será cero, pero un coeficiente de correlación cero no significa que no existe ninguna relación".

3) ¿Qué diferencias y semejanzas encuentra entre regresión y correlación?

4) ¿Qué significa coeficiente de determinación y que significa coeficiente de correlación?

5) ¿Por qué se interesan los estadísticos frecuentemente en la pregunta ¿es $\beta = 0$? ¿Indica la magnitud de β qué tan bien pueden hacerse las predicciones? Discuta.

6) ¿Tiene sentido estimar α en la ecuación de regresión en todos los casos? (recuerde la interpretación del ítem "e" del ejercicio: 4 ¿Cuál sería el significado de $\alpha = 0$? ¿Podría ser su estimador "a" menor que cero si la variable y es, por ejemplo, una "longitud"?

7) ¿Cómo mide el coeficiente de correlación la fuerza de la relación lineal entre dos variables?

8) ¿Qué valor toma r si todos los puntos muestrales caen sobre la misma recta y si

a) la recta tiene pendiente positiva?

b) La recta tiene pendiente negativa?

OTRAS PREGUNTAS INTERESANTES:

1. ¿Cuál es el objetivo del análisis de regresión y cuál en un análisis de correlación?
2. Distinga entre: modelo teórico, modelo estadístico, modelo estimado.
3. Represente los modelos indicados en el ítem anterior.
4. Mencione los supuestos para poder realizar un análisis de regresión lineal.
5. Diga qué elementos tiene y cómo los relaciona la ecuación de regresión, cuando se hace referencia a la ecuación de regresión entre las variable aleatoria "y" y la variable "x".
6. Para poder trabajar con regresión lineal, ¿ambas variables deben de ser aleatorias? Explique.
7. ¿Qué significa en la regresión lineal que β sea cercano a cero?
8. Explique las diferencias entre un intervalo de predicción para un valor de y dado y el intervalo de confianza para la respuesta media $\mu_{y/x}$.
9. Explique por qué un intervalo de predicción para un evento específico es más amplio que un intervalo de confianza para la respuesta media correspondiente, utilizando el mismo coeficiente de confianza en ambos casos. ¿Significa esto que debería estar más interesado en estimar respuestas medias que en predecir eventos específicos?
10. ¿Qué mide el coeficiente de correlación lineal entre dos variables?