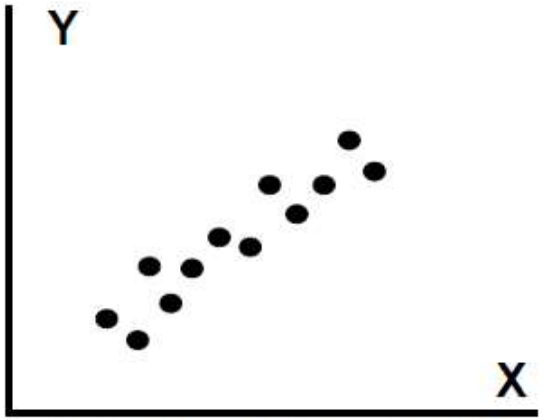
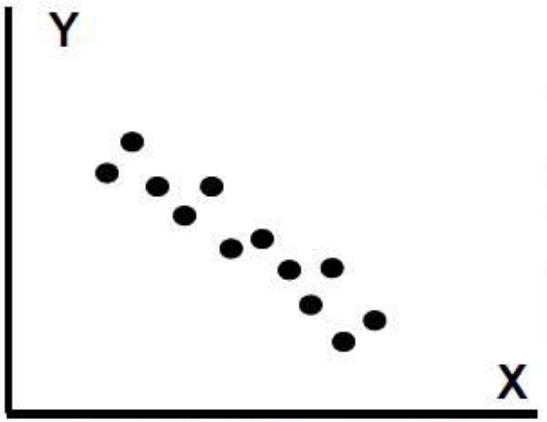
Unidad V

Análisis de Correlación



Correlación positiva



Correlación negativa

Relación entre variables aleatorias

En muchos problemas existe una relación inherente entre dos o más variables, y resulta de utilidad explorar la naturaleza de esta relación.

La matemática estudia variables relacionadas de manera determinística. Dos variables X e Y están relacionadas en forma ***determinística*** si conocido el valor de X, el valor de Y queda especificado por completo.

Ejemplo: Si deseamos determinar el área de un círculo, siendo **X** =" radio del circulo"(cm) e **Y** = "área del circulo", entonces $\mathbf{Y = \pi X^2}$.
Si **X = 2**, entonces $\mathbf{Y = \pi\ 2^2 = 12.57\ cm^2}$.
Cada valor de **X** determina de manera exacta el valor correspondiente de **Y.**

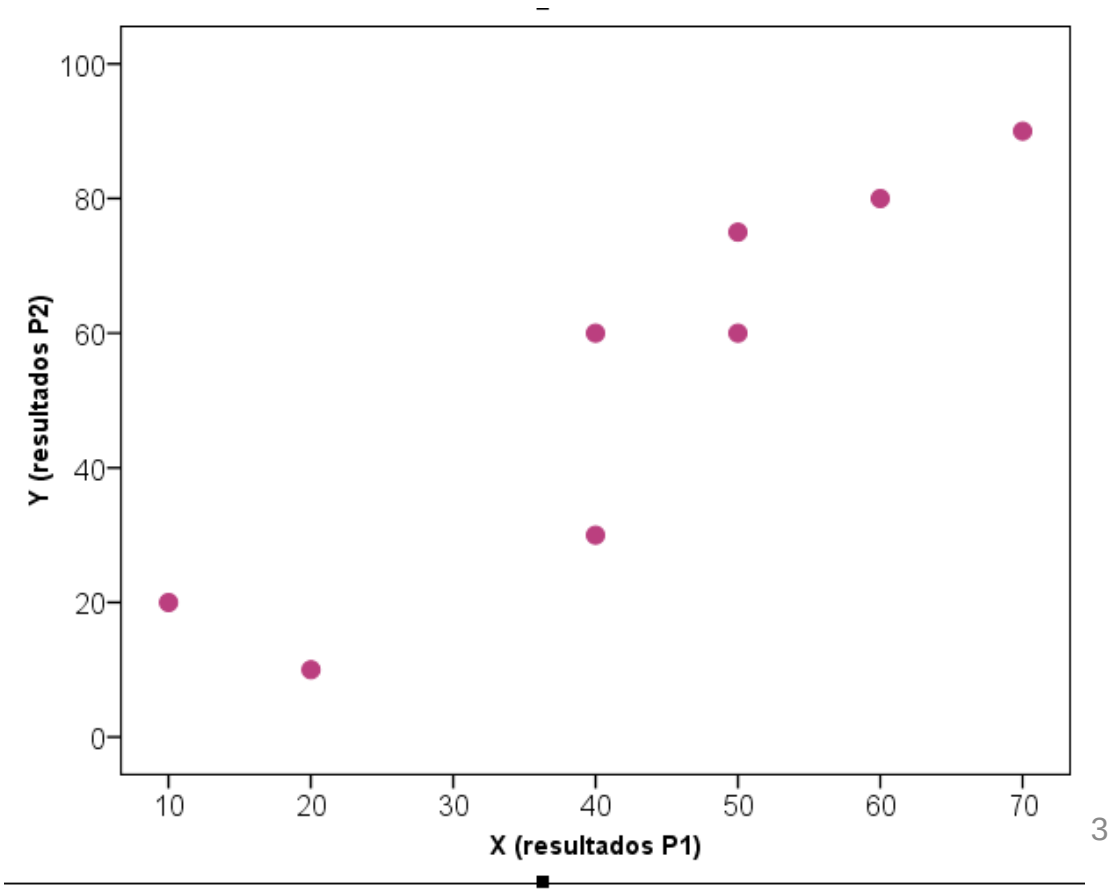
Existen muchas variables X e Y que están relacionadas entre sí, pero **no** de manera **determinística.**

Ejemplo, si **X** ="la altura, en m, de un estudiante de primer año de la UNS", e **Y** ="el peso, en Kg.", el valor de **Y no** queda especificado por completo si se conoce **X**; dos estudiantes podrían tener el mismo valor de **X**, pero tener valores de peso (Y) muy diferentes.

Ejemplo: Las calificaciones obtenidas por **9** alumnos de Estadística en el primer parcial y en el segundo parcial son:

Calificaciones								
P1 (X)	10	20	40	40	50	50	60	70
P2 (Y)	20	10	30	60	60	75	80	90

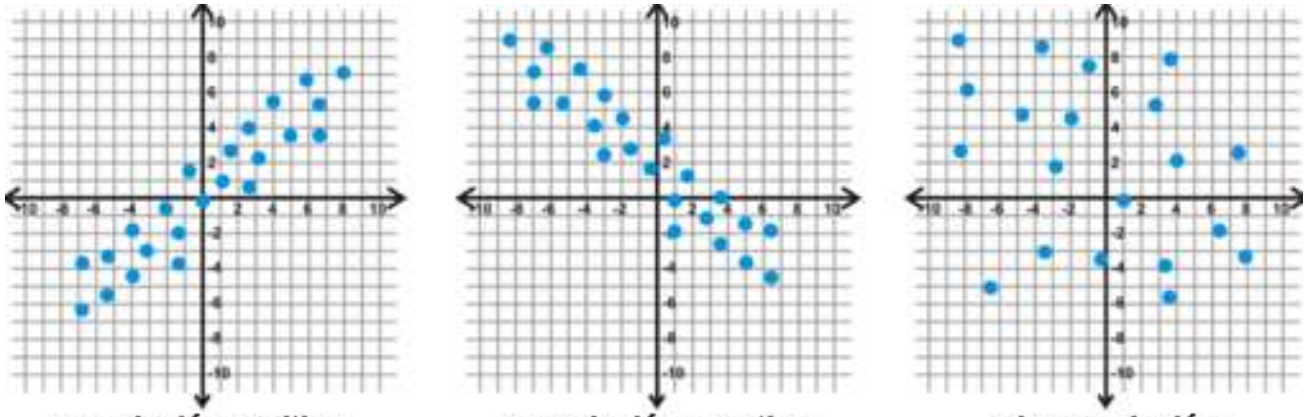
¿Existe relación entre los resultados del primer y segundo parcial?



Relación lineal entre variables aleatorias

La técnica estadística que permite la investigación de la relación *no determinística lineal* entre dos o más variables es el ***Análisis de Correlación*****.**

Este análisis trata de medir *el grado de asociación lineal* entre dos variables aleatorias cuantitativas **X** e **Y**. En qué medida un cambio en una de ellas afecta a la otra.



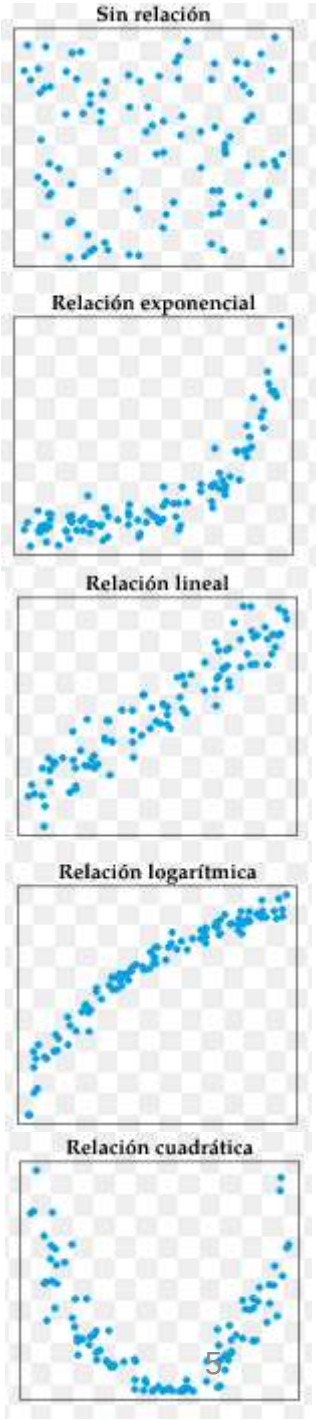
correlación positiva | correlación negativa | sin correlación

Herramienta gráfica: Diagrama de Dispersión

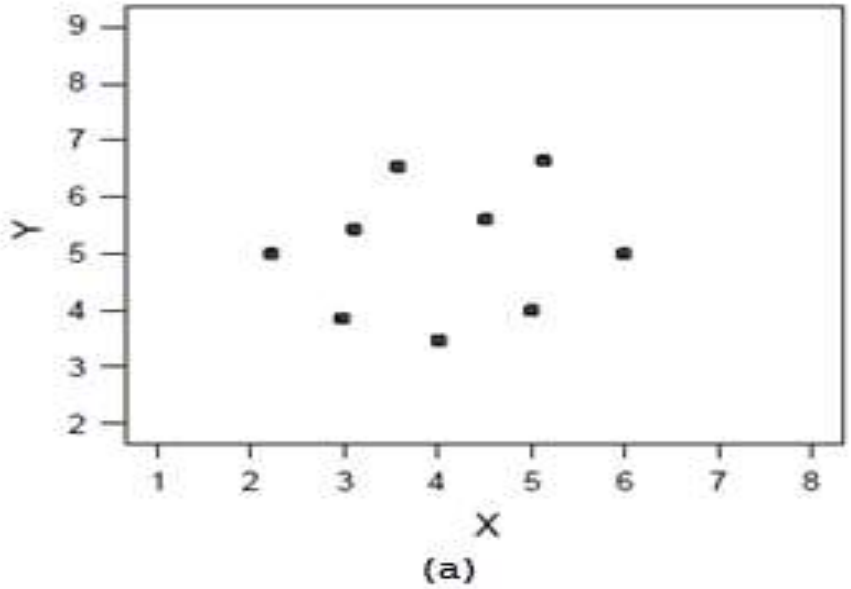
Supongamos que deseamos observar si existe **relación lineal** entre la v.a. **X** y la v.a. **Y.** Si medimos estas v.a. en ***n*** unidades experimentales (individuos, objetos, etc.), habremos recolectado los siguientes datos:

$$(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_n, y_n).$$

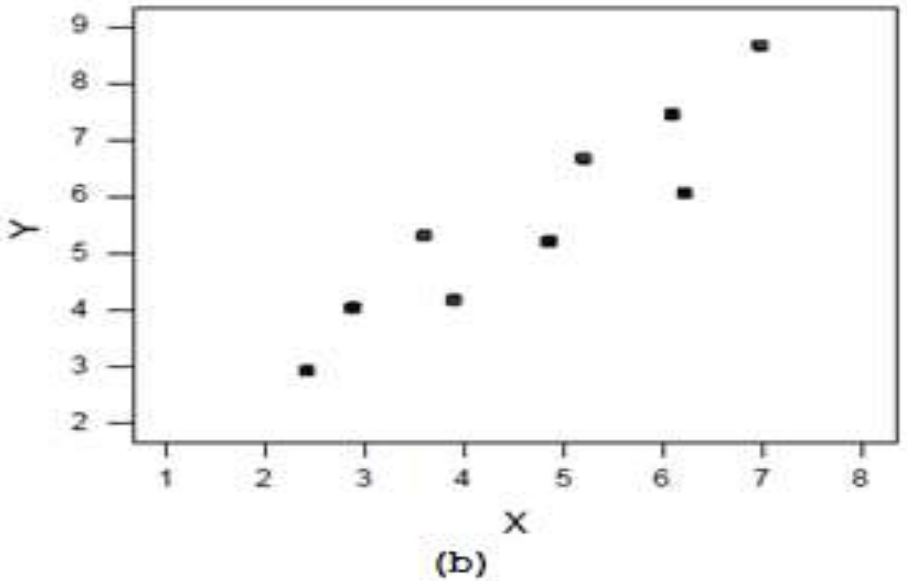
El ***diagrama de dispersión*** ("**scatterplot**") permite tener una primera impresión sobre el tipo de relación existente entre dos v.a.. Este método gráfico coloca una de las v.a. (X) en el eje de las abscisas y la otra (Y), en el de las ordenadas, grafica por medio de puntos, (x_i, y_i), los valores correspondientes de las variables bajo estudio para cada una de las unidades experimentales analizadas. Queda así determinada una ***nube de puntos***. La forma de esta nube de puntos informa sobre el tipo de relación existente entre las variables. Permite establecer algún patrón de comportamiento gráfico.



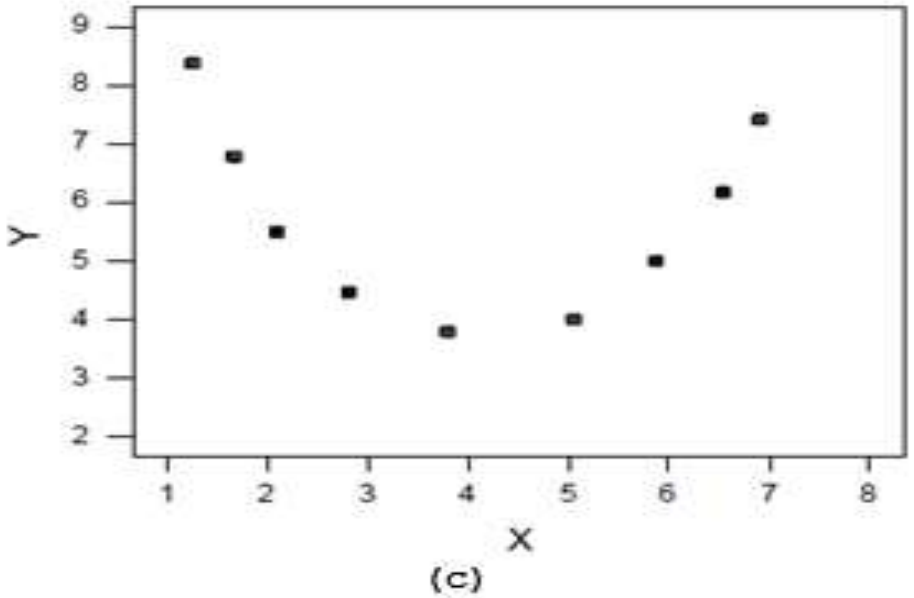
Diagramas de Dispersión que reflejan situaciones de relaciones diferentes entre las variables.



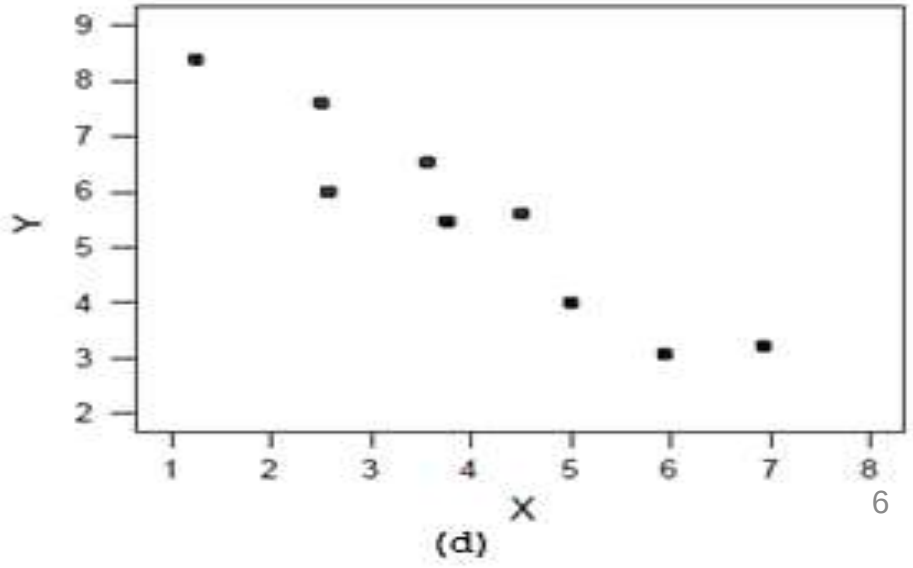
(a)



(b)



(c)



(d)

Importante!!!

El ***Diagrama de Dispersión*** en algunos casos puede que no resulte tan práctico para medir el **grado** de **asociación lineal**. Esto es debido a que la relación entre dos variables no siempre es **perfecta** o **nula**; habitualmente no es ni lo uno ni lo otro.

Un índice que da a conocer el **sentido de la asociación lineal** entre las v.a. X e Y, si es que ésta existe, es la **covarianza**.

$$\mathbf{Cov(X,Y)} = E[(X - E(X)).(Y - E(Y))] = E(X.Y) - E(X).E(Y)$$

Covarianza poblacional

Dados los puntos $(x_1, y_1), (x_2, y_2),..., (x_n, y_n)$, una muestra bidimensional obtenida al medir X e Y en ***n*** individuos u objetos, la **covarianza** de X e Y, denotada **Cov(X, Y)** puede ser estimada por la **Covarianza muestral**, $\mathbf{S_{XY}}$ cuya expresión es,

$$S_{XY} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

donde $\bar{X}$ e $\bar{Y}$ son las **medias muestrales**.

Interpretación del signo de la covarianza

Autor:
Chema Falcó

$$S_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

https://www.geogebra.org/m/f8zWM9Bq

Covarianza de dos variables X e Y

$$\mathbf{S_{XY}} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y})}{n-1}$$

La **covarianza muestral** entre dos variables nos indica si la posible relación lineal entre dos variables cuantitativas es **positiva** (**directa**) o **negativa** (**inversa**) ó **no existe**. El **sentido** de la relación.

– positiva: $\mathbf{S_{xy} > 0}$ (cuanto mayores son las puntuaciones en la variable **X**, mayores son las puntuaciones en la variable **Y**). Varían en igual sentido.

– negativa: $\mathbf{S_{xy} < 0}$ (cuanto mayores son las puntuaciones en la variable **X**, menores son las puntuaciones en la variable **Y**). Varían en sentidos opuestos.

– Incorreladas: $\mathbf{S_{xy} = 0}$ (no existe relación lineal entre **X** e **Y**).

El **signo de la covarianza** nos dice si el aspecto de la nube de puntos es **creciente** o **no**, pero **no nos dice nada** sobre el **grado de relación** entre las variable aleatorias.

Ejemplo: Eficiencia de programas informáticos

Un administrador de computadoras necesita saber cómo la eficiencia de su nuevo programa de computadora depende del tamaño de los datos entrantes. La eficiencia se medirá por la cantidad de solicitudes procesadas por hora. Aplicando el programa a conjuntos de datos de diferentes tamaños, obtiene los siguientes resultados,

X = "Tamaño de un conjunto de datos (gigabytes)"	6	7	7	8	10	10	15
Y = "N° de solicitudes procesadas por hora"	40	55	50	41	17	26	16

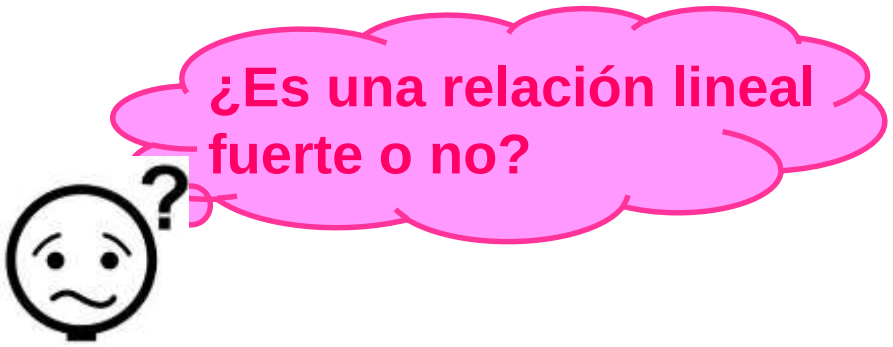
¿El N° de solicitudes procesadas por hora depende del tamaño del conjunto de datos?

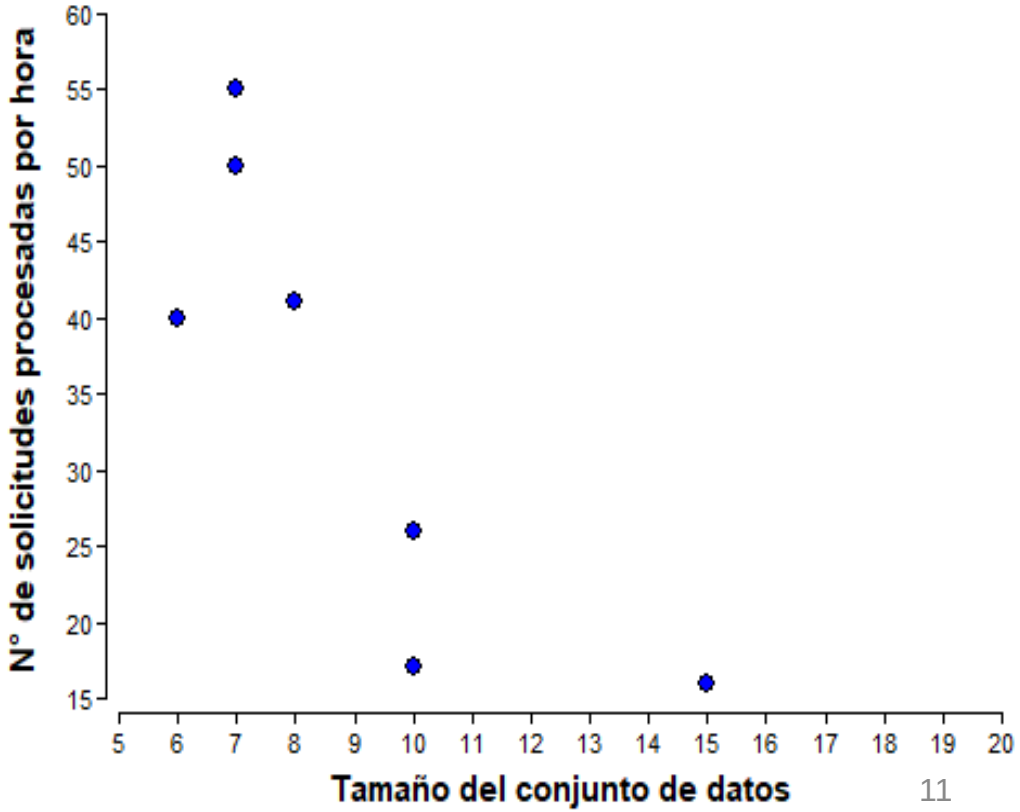
¿El N° de solicitudes procesadas por hora depende del tamaño de conjunto de datos?

$$\mathbf{S_{XY}} = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}}{n-1} = \frac{1973 - 7 \cdot 9 \cdot 35}{6} = -\frac{232}{6} = \mathbf{-38.7}$$

Relación negativa:

$\mathbf{S_{xy}} < 0$ (cuanto mayor sea el tamaño del conjunto de datos, **X**, menor es la cantidad de solicitudes procesadas por hora, **Y**), es decir, el sentido de la asociación lineal es inverso.





Covarianza muestral de dos variables X e Y

$$\mathbf{S_{XY}} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y})}{n-1}$$

El **valor** de $\mathbf{S_{XY}}$ depende de :

a) las magnitudes de los valores y las unidades en que están medidas las v.a. **X** e **Y**.

b) el número de puntos de la muestra, ***n***.

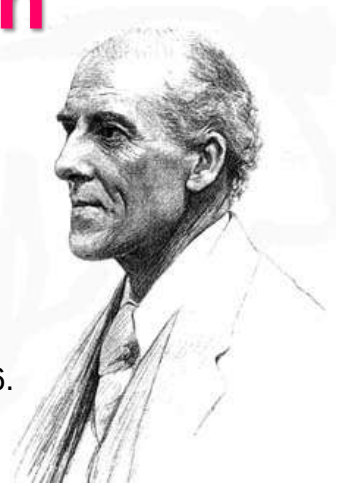
Por tal motivo, **no nos dice nada** sobre el **grado de relación lineal** entre las variable aleatorias.

Coeficiente de correlación de Pearson

Una medida del **grado o intensidad de relación lineal** entre dos variables cuantitativas **X** e **Y** está dada por el ***coeficiente de correlación de Pearson*** de las variables aleatorias **X** e **Y**, denotado: ρ_{XY}, se define:

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_x \, \sigma_y}$$

Propuesto por Karl Pearson en1896.



ρ_{XY}, puede estimarse por medio del **coeficiente de correlación muestral** **r,** definido por

$$\mathbf{r} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y - \bar{y})^2}} = \frac{S_{XY}}{\sqrt{S_X{}^2 \, S_Y{}^2}}$$

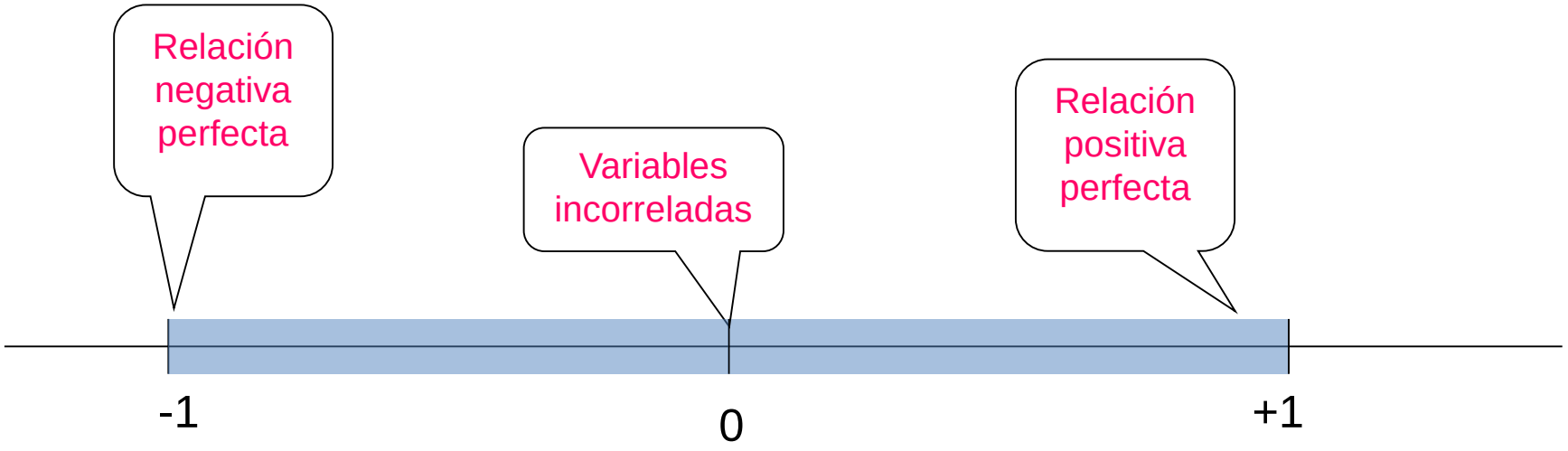
donde $\mathbf{S_{XY}}$ es la *covarianza muestral* y $\mathbf{S^2_X}$ y $\mathbf{S^2_Y}$ las *varianzas muestrales* de los valores de **X** e **Y,** respectivamente.

- **r** indica si los puntos tienen una tendencia a disponerse **alineadamente** (excluyendo rectas **horizontales** y **verticales**).
- Tiene el mismo signo que **Sxy,** por tanto, de su signo obtenemos si la posible relación es positiva o negativa.
- **r** es útil para determinar si hay **relación lineal** entre dos variables, pero no servirá para otro tipo de relaciones (cuadrática, logarítmica, etc.)

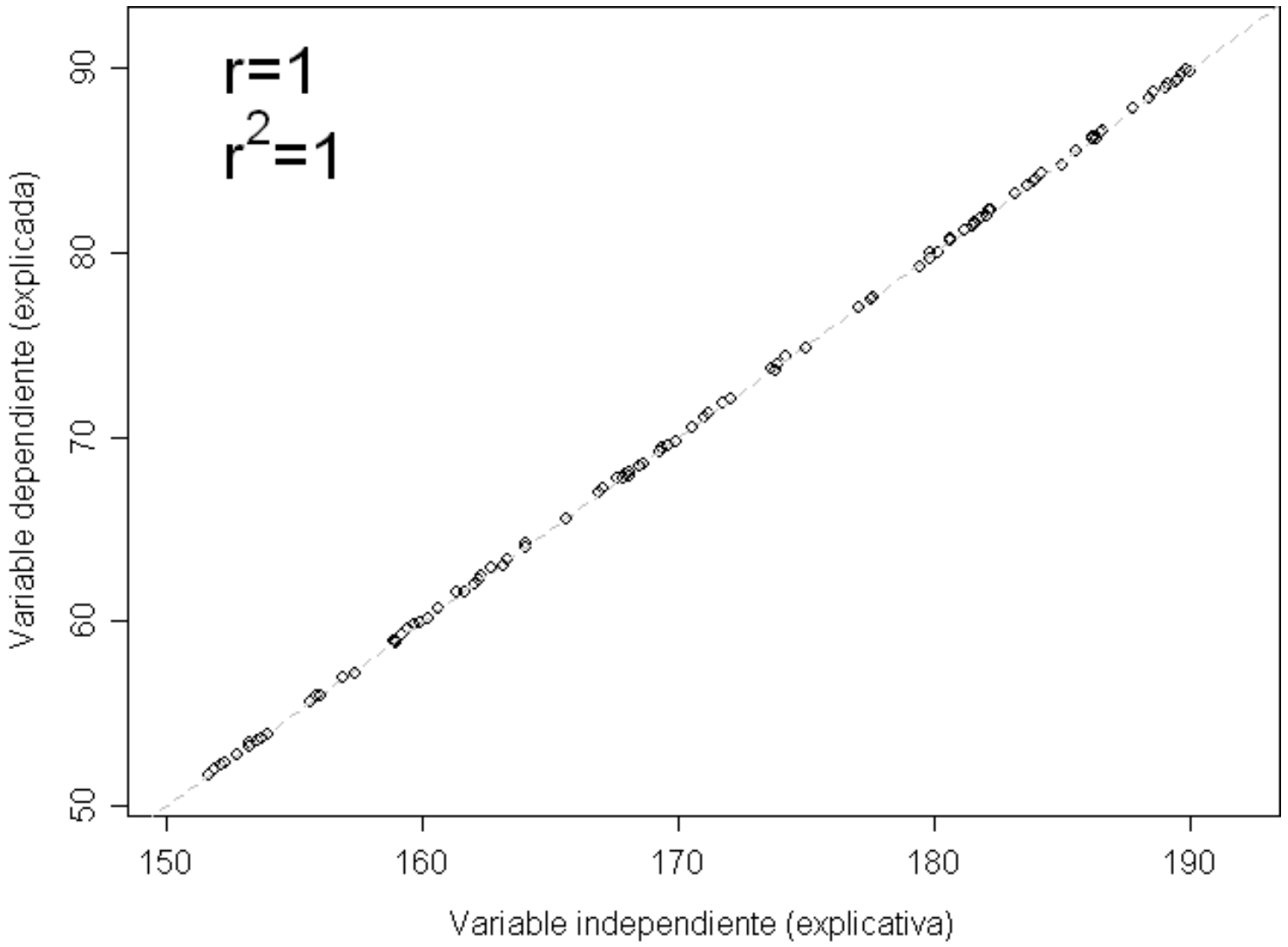
Propiedades de r

- Es ***adimensional*** (ya que su numerador y su denominador están medidos en las mismas escalas).
- Sólo toma valores en **[-1,1]**
- Las variables aleatorias cuantitativas son ***incorreladas*** ⇔ **r = 0**
- Relación ***lineal perfecta*** entre dos variables ⇔ **r = +1 o r = -1**

  Ojo!!! Se excluyen los casos de puntos alineados **horizontalmente** o **verticalmente.**
- Cuanto más cerca esté **r** de **+1** o **-1 mayor** será el **grado** o **intensidad** de relación lineal entre las variables.

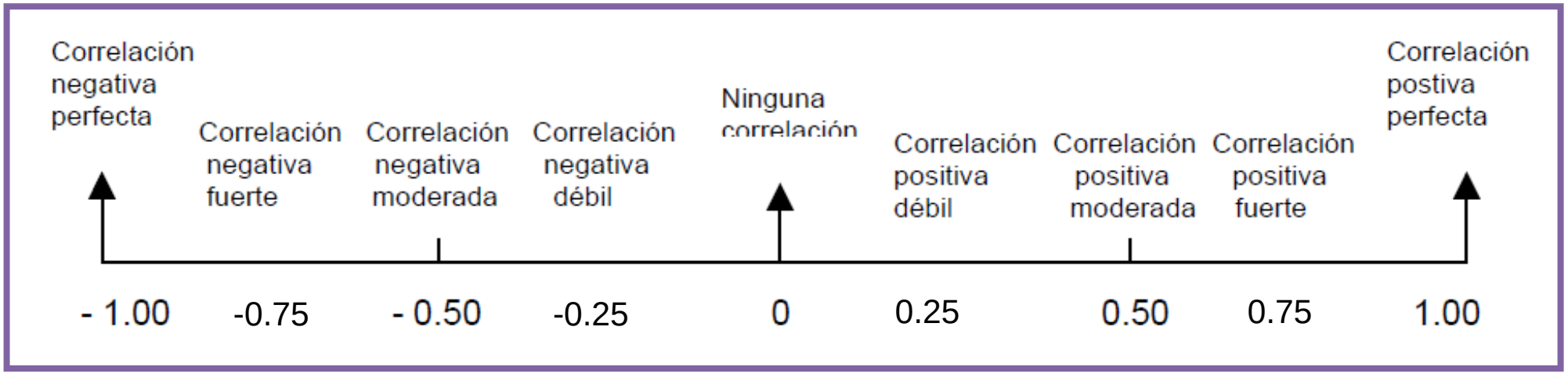


Animación: Evolución de r y diagrama de dispersión



Clasificación del coeficiente de correlación muestral r

Un criterio para clasificar la correlación entre dos v.a. cuantitativas, es el siguiente:



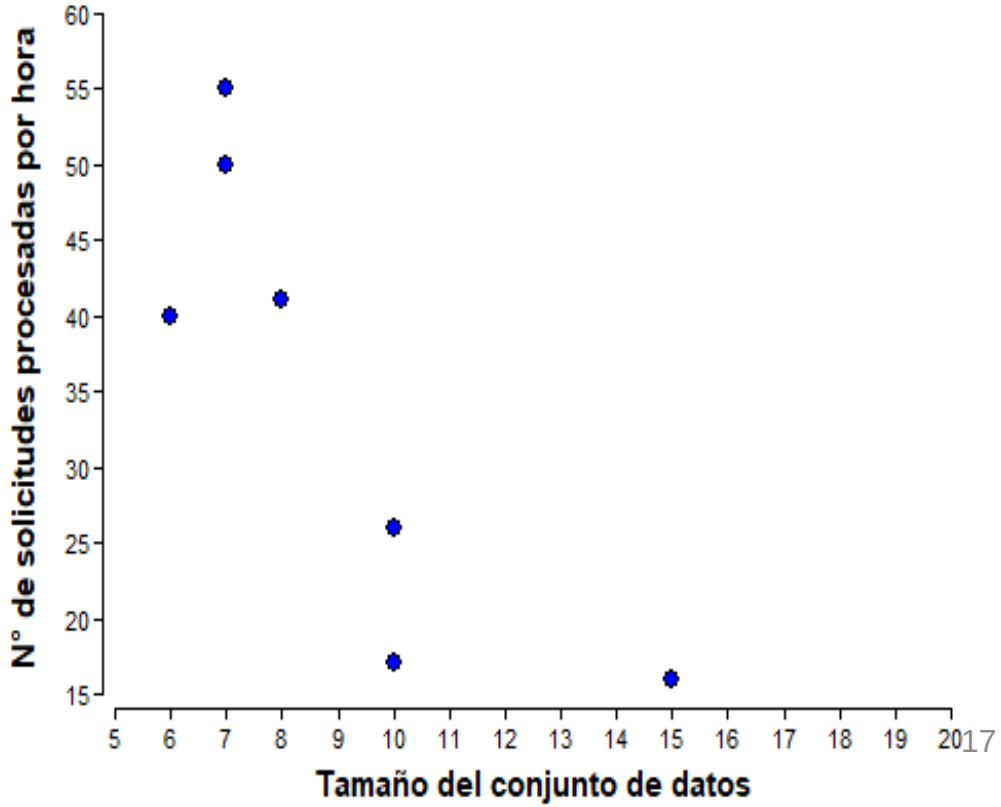
Considerando el **ejemplo** del tamaño del conjunto de datos (X), en gigabytes, y el N° de solicitudes procesadas por hora (Y),

¿hay una relación lineal fuerte o no?

$$r = \frac{S_{XY}}{\sqrt{S_X{}^2 \, S_Y{}^2}} = \frac{-38.7}{\sqrt{9.3 * 242.1}} \cong -\mathbf{0.82}$$

Coeficiente de correlación = r = - 0.82

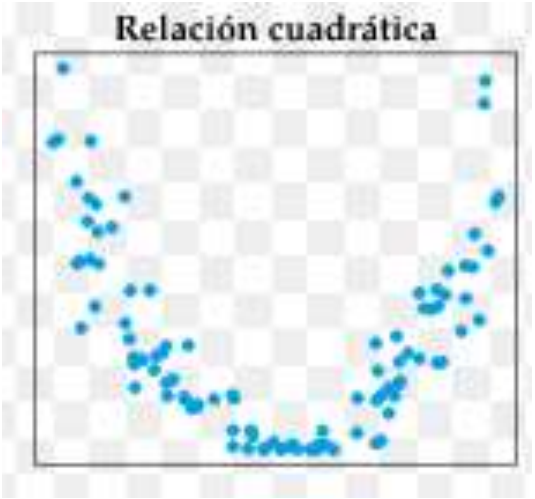
***Podemos concluir que, el tamaño del conjunto de datos* (X), en gigabytes, y el *N° de solicitudes procesadas por hora* (Y), poseen *una* relación lineal negativa fuerte**.



Pregunta frecuente:

¿Si r = 0 eso quiere decir que las variables son independientes?

Si **r** = 0, significa que **no hay relación lineal** entre las variables.
Pero puede existir otro tipo de relación entre ellas.



Lo contrario, si es cierto, es decir:

Independencia ⇒ **incorrelación**

Cuarteto de Ascombe

El **cuarteto de Anscombe** comprende cuatro conjuntos de datos que tienen las **mismas propiedades estadísticas**, pero se **comportan distinto** al inspeccionar sus **gráficos de dispersión** respectivos.

- **1**

Dato	1	2	3	4	5	6	7	8	9	10	11
X (x_i)	10	8	13	9	11	14	6	4	12	7	5
Y (y_i)	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.68

- **2**

Dato	1	2	3	4	5	6	7	8	9	10	11
X (x_i)	10	8	13	9	11	14	6	4	12	7	5
Y (y_i)	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74

- **3**

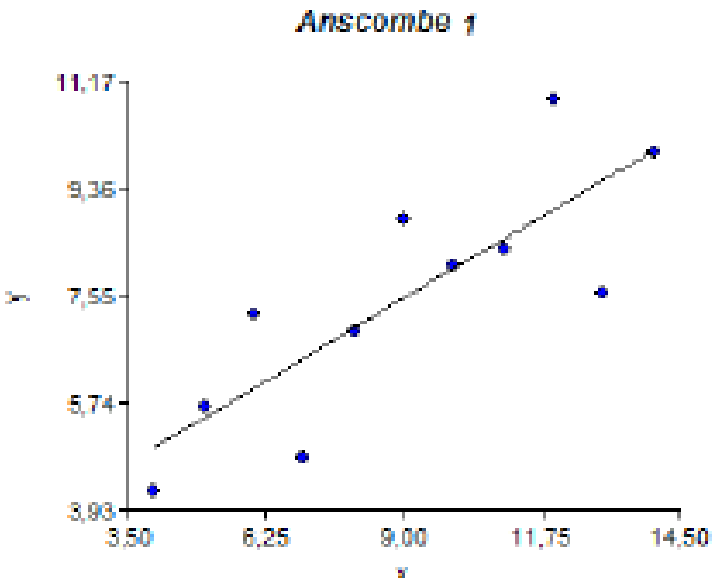
Dato	1	2	3	4	5	6	7	8	9	10	11
X (x_i)	10	8	13	9	11	14	6	4	12	7	5
Y (y_i)	7.46	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42	5.73

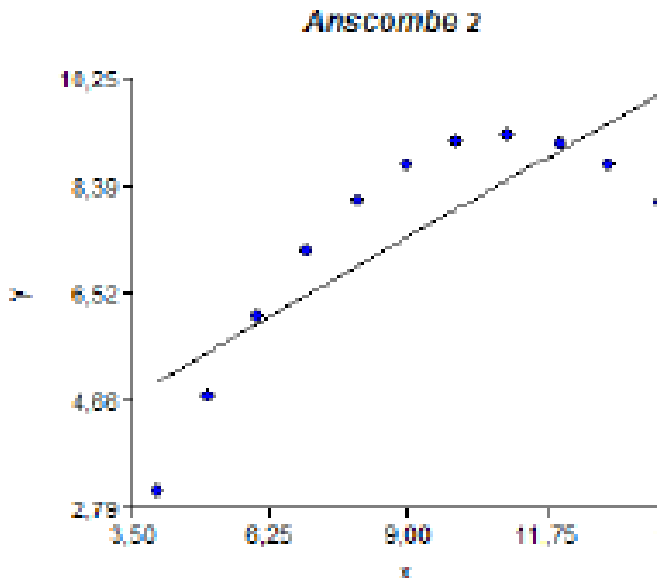
- **4**

Dato	1	2	3	4	5	6	7	8	9	10	11
X (x_i)	8	8	8	8	8	8	8	19	8	8	8
Y (y_i)	6.58	5.76	7.71	8.84	8.47	7.04	5.25	12.50	5.56	7.91	6.89

Cuarteto de Ascombe

Para los cuatro conjuntos de datos, se obtuvo:

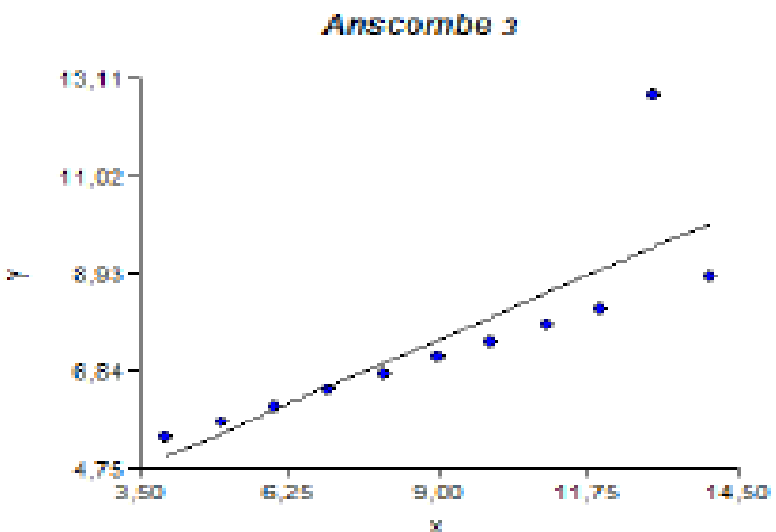


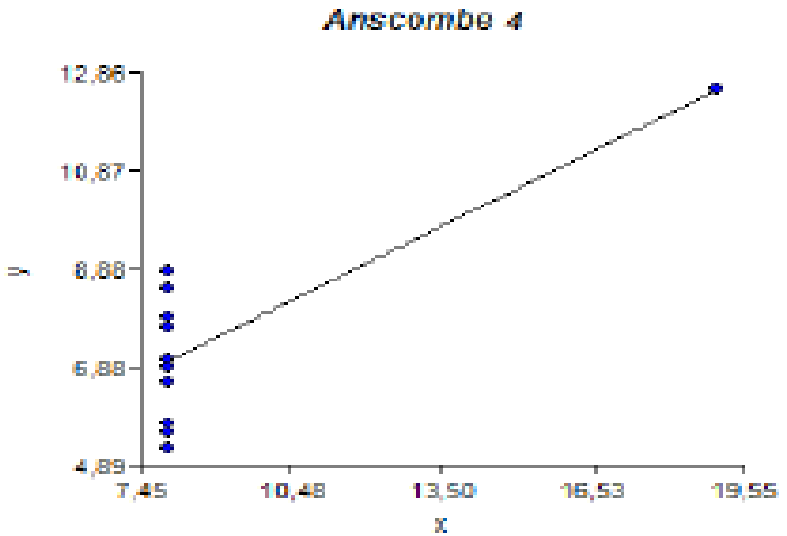


$\overline{X} = 9$ e $\overline{Y} = 7.5$

$S^2_X = 11$ y $S^2_Y = 4.12$

r = 0.816





- El primer gráfico muestra lo que parece una relación lineal simple, correspondiente a dos variables correlacionadas.

- El segundo gráfico, aunque se observa relación entre los datos, esta no es lineal y el coeficiente de correlación de Pearson, no es relevante.

- En la tercera gráfica, la distribución es lineal, pero con una línea de regresión diferente de la que se sale el dato extremo que influye lo suficiente como para alterar la recta y disminuir el coeficiente de correlación de 1 a 0.816.

- Por último, la cuarta gráfica, es un ejemplo de muestra en la que un valor atípico es suficiente para producir un coeficiente de correlación alto incluso cuando la relación entre las dos variables no es lineal.

Causalidad

X e Y están relacionadas → **una es causa de la otra???**

El hecho de que dos variables estén **relacionadas** **no necesariamente implica** que una sea **causa de la otra**, ya que puede ocurrir el hecho de que se esté dando una variación concomitante (simultánea), por el simple hecho de que **las dos son causa** de una **tercera**.

Por ejemplo, si se realiza un estudio en el que se analiza el número de canas (X) y la presión arterial (Y) podría encontrarse una relación lineal casi perfecta. Eso no significa que el tener canas aumente la presión arterial, lo que verdaderamente está ocurriendo es que es la edad (Z), la causante, de que se tengan más canas y una tendencia a tener más alta la presión arterial.

Correlación espuria

Se da cuando dos o más variables se **creen** estadísticamente relacionadas, porque el coeficiente de correlación presenta un valor cercano a **1** o **-1**, pero en realidad, **no** tienen una relación de causalidad entre ellas. Esto puede deberse a que existe un tercer factor no considerado, denominado "***factor de confusión***" o puede ocurrir sólo por ***azar***. Este tipo de correlaciones se denominan ***correlaciones espurias***

Por ejemplo, durante el verano en cierta ciudad aumenta la venta de helados y la tasa de asesinatos. Una correlación espuria sería identificar que el aumento en la venta de helados causa un aumento en la tasa de asesinatos. En este caso, se podría considerar una tercera variable, el calor, pueda estar asociada a ambas mostrando esta falsa correlación.

Correlación espuria

Importante!!!!

Los gráficos que se muestran a continuación no pretenden dar a entender la causalidad ni están destinados a crear desconfianza en la investigación o en los datos que están bajo estudio.

Correlaciones Espurias

Un **estudiante de criminología** de la Escuela de Leyes de la Universidad de Harvard, **Tyler Vigen**, diseñó un sistema informático que busca en grandes conjuntos de datos patrones de correlación y publica en su sitio miles de **correlaciones espurias**.

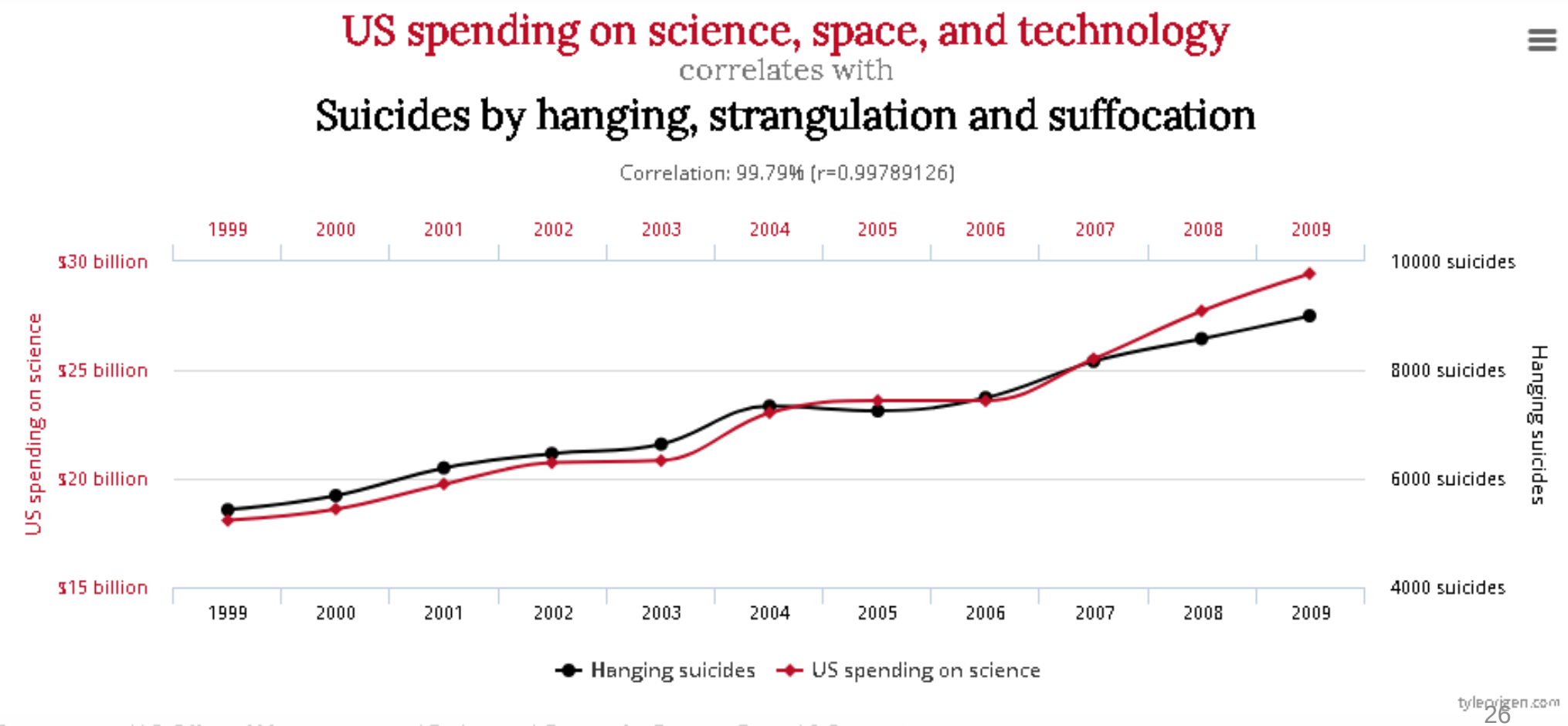
Spurious correlations

tylervigen.com

https://www.tylervigen.com/spurious-correlations

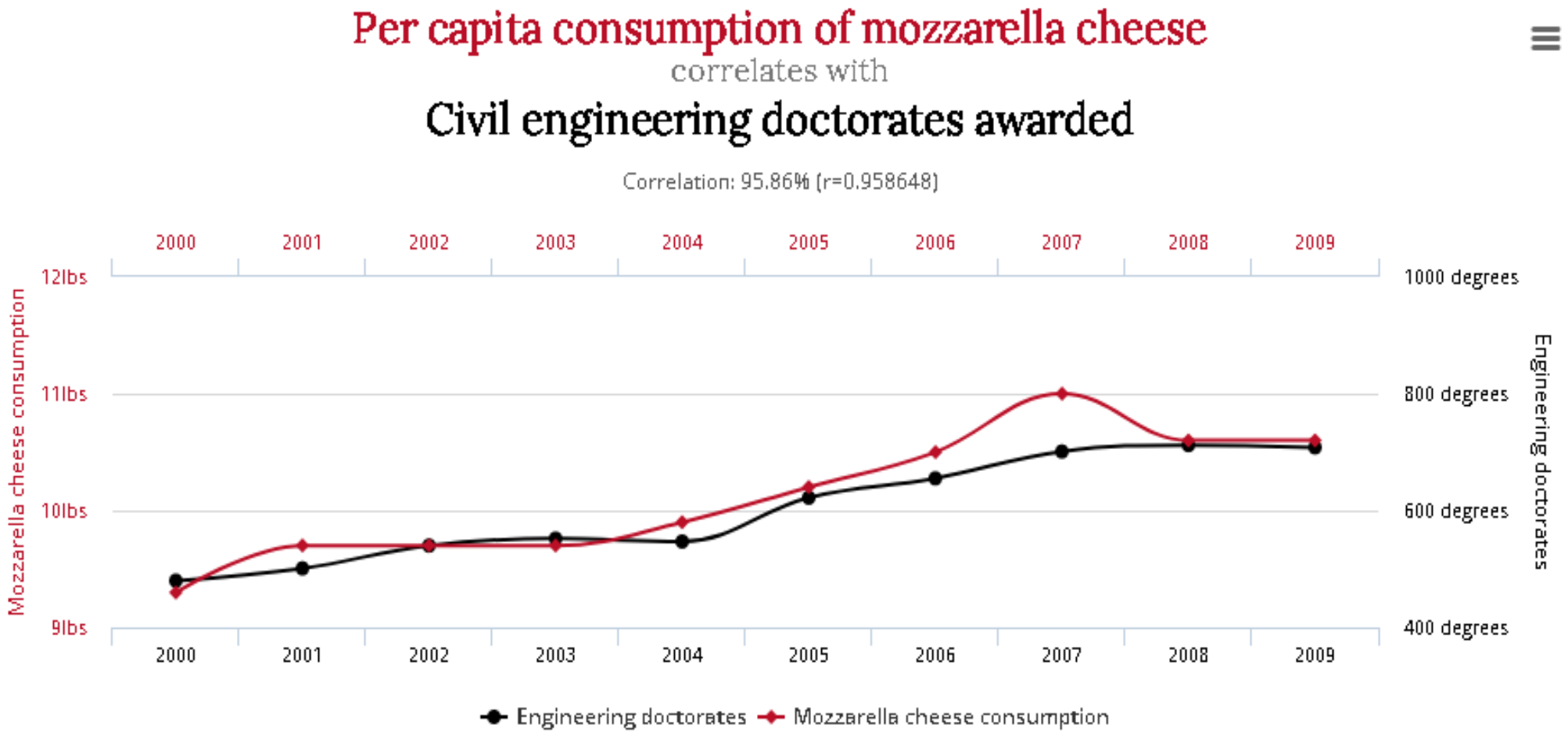
Correlación espuria

Gasto en ciencia, espacio, tecnología EE.UU. (rojo) vs. suicidios por asfixia, sofocación y estrangulamiento (negro). **r = 0.997**



Correlación espuria

Consumo de queso muzzarella per cápita (rojo) vs. cantidad de doctores en Ingeniería civil graduados (negro), r = **0.958.**



Data sources: U.S. Department of Agriculture and National Science Foundation

IMPORTANTE!!!

- Dos variables que estén correlacionadas no significa necesariamente que exista una relación de causalidad.

- Que dos variables tengan un comportamiento similar en el tiempo, no implica que exista una relación causal entre ellas.

Uso de la Calculadora

Las calculadoras permiten calcular el coeficiente de correlación muestral r, como así también la pendiente y la ordenada al origen de la recta de ajuste a una nube de puntos, entre otras cantidades.

Por ejemplo: en una calculadora **CASIO fx – 82MS**, se debe ir:

Mode >> 3 (Reg)
>> 1 (Lin)
Luego, ingresar los datos.

CÁLCULOS ESTADÍSTICOS PARA DOS VARIABLES

- Pulsar la tecla **MODE** tantas veces hasta que en el visor aparezcan las opciones:

COMP SD REG

- Pulsar la tecla del número asociado a la opción: **REG**, es decir, la tecla: **3**. En el visor se muestran varias opciones de regresión. Pulsar la tecla del número correspondiente a la opción: **Lin** (es la opción **3**, de regresión lineal, única que tratamos en el curso).

- **Limpiar la memoria**.

Pulsar las teclas:

SHIFT+ MODE

En el visor aparece:

Scl Mode All

Pulsar la tecla correspondiente a **Scl** y en el visor se muestra el mensaje:

Stat clear

Pulsar la tecla: **=**

Después pulsar la tecla: **AC**

Ejemplo: ingresos de datos

X	Y
6	10
4	11
5	10
8	14
7	12
8	16

Pulsar

Teclas: 6 , 1 0 M+
Teclas : 4 , 1 1 M+
Teclas: 5 , 1 0 M+
Teclas: 8 , 1 4 M+
Teclas: 7 , 1 2 M+
Teclas: 8 , 1 6 M+

Una vez ingresados los datos, las **sumatorias** y los **índices estadísticos** se encuentran asociadas a las teclas:

- **SHIFT + 1,** para los **sumatorias**, en el visor se muestra lo siguiente:

$$\underset{1}{\sum x^2} \quad \underset{2}{\sum x} \quad \underset{3}{n} \quad \rightarrow$$

La flecha de la derecha significa que aún hay otras sumatorias. Pulsando el lado derecho de la tecla direccional se accede a:

$$\leftarrow \quad \underset{1}{\sum y^2} \quad \underset{2}{\sum y} \quad \underset{3}{\sum xy}$$

- **SHIFT + 2** para los **índices estadísticos univariados** y los relativos a la **correlación** y **Regresión lineal Simple**.

Cuando se pulsa **SHIFT + 2,** en el visor se muestra:

$\bar{x}$ $x\sigma_n$ $x\sigma_{n-1}$ $\rightarrow$

1 2 3

Además de ésta, hay tres pantallas más:

$\leftarrow$ $\bar{y}$ $y\sigma_n$ $y\sigma_{n-1}$ $\rightarrow$

1 2 3

$\leftarrow A$ B r $\rightarrow$

1 2 3

Los coeficientes de la recta de ajuste de la regresión lineal y el coeficiente de correlación.

$\leftarrow \hat{x}$ $\hat{y}$

1 2

Las puntuaciones pronosticadas.

Para los datos de la tabla, algunos de los **resultados estadísticos** son:

$$\sum x^2 = 254$$

$$\sum y^2 = 917$$

$$\bar{x} = 6.333333$$

$$\bar{y} = 12.166666$$

$$r = 0.799023971$$

Uso de la Calculadora

Para el uso de este tipo de calculadoras en el **Análisis de Correlación** y en **Regresión Lineal Simple**, seguir paso a paso el siguiente tutorial.

https://www.youtube.com/watch?v=SrGHclGORX4

Prueba de Hipótesis

Significación de coeficiente de correlación de Pearson

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_x \, \sigma_y}$$

Prueba de Hipótesis: Coeficiente de Correlación

Objetivo: Permite analizar la existencia de correlación entre dos v. a. y ver si ésta es **significativa** o **no.**

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_x \ \sigma_y}$$ → **coeficiente de correlación poblacional**

$$r = \frac{S_{XY}}{\sqrt{S_X{}^2 \ S_Y{}^2}}$$ → **coeficiente de correlación muestral (estimador puntual de ρ_{XY})**

1. Planteo:

H_0: $\rho_{XY} = 0$ (No hay correlación entre las v.a.)
H_1: $\rho_{XY} \neq 0$ (Hay correlación entre las v.a.)

2. α **:** nivel de significación elegido.

3. Estadístico de Prueba bajo H_0 cierta:

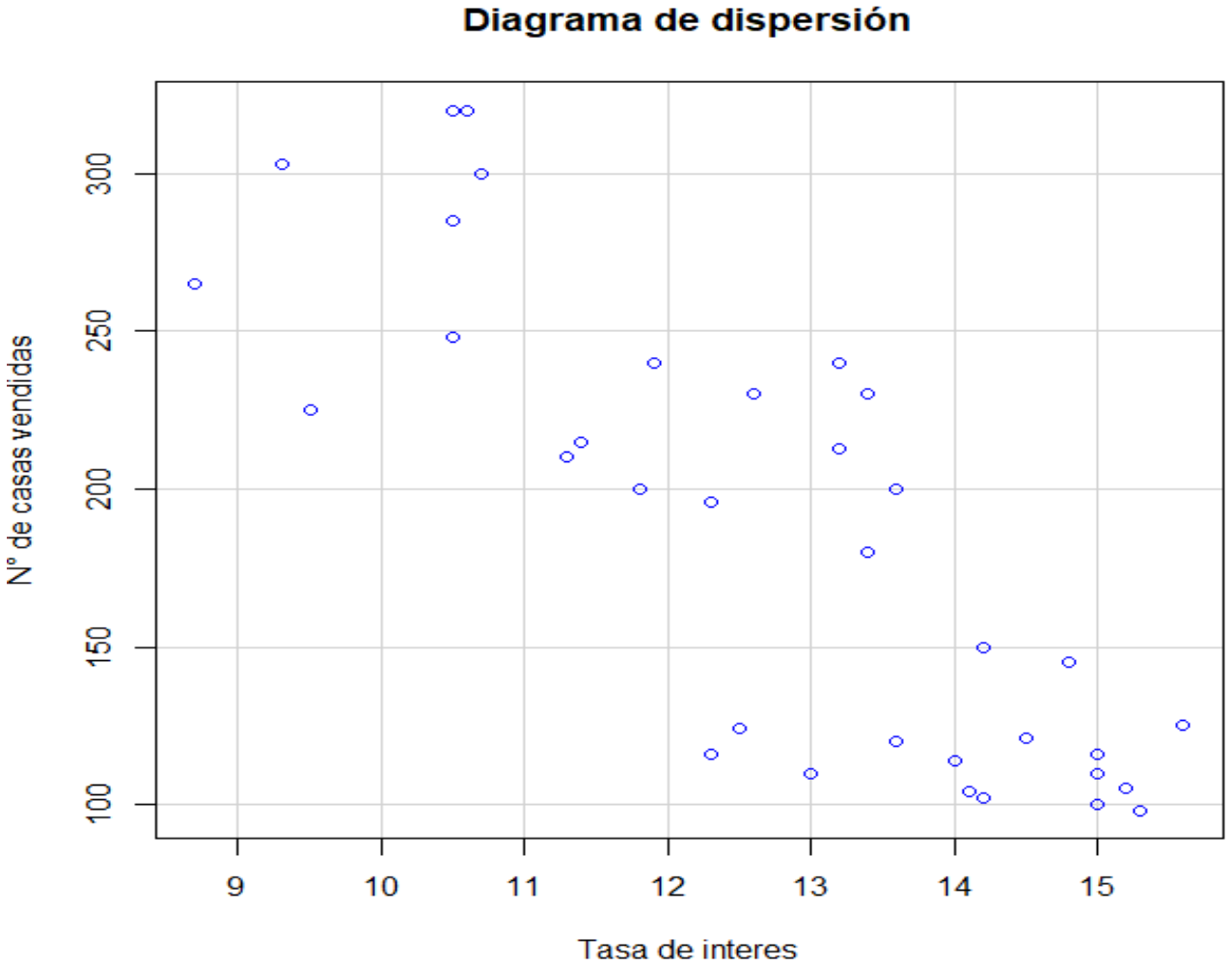
$$\frac{r-0}{\sqrt{\frac{1-r^2}{n-2}}} = T$$ ~ **t - student con ν = n-2 gl**

Donde $\sqrt{\frac{1-r^2}{n-2}} = \sqrt{V(r)}$, es el **error estándar** del coeficiente de correlación muestral **r**.

Ejemplo:

Un banco local que se especializa en créditos para vivienda intenta analizar la relación entre *las tasas de interés mensual (%)* y el *número de casas vendidas* por mes en cierta área de la ciudad. Se compilaron los datos para un período de 35 meses elegidos al azar. La información obtenida se muestra a continuación:

Interés	N° de casas		Interés	N° de casas
12,3	196		10,7	300
10,5	285		13	110
15,6	125		14,1	104
9,5	225		15	110
10,5	248		13,2	240
9,3	303		13,6	200
8,7	265		13,4	230
14,2	102		13,2	213
15,2	105		12,5	124
14	114		11,9	240
15	116		10,5	320
13,6	120		13,4	180
14,5	121		15	100
12,3	116		15,3	98
11,3	210		14,8	145
11,8	200		14,2	150
11,4	215		12,6	230
			10,6	320



Ejemplo:

¿Están correladas *las tasas de interés mensual (%)*, **Y** y el *número de casas vendidas mensualmente*, **X**?

Interés	N° de casas		Interés	N° de casas
12,3	196		10,7	300
10,5	285		13	110
15,6	125		14,1	104
9,5	225		15	110
10,5	248		13,2	240
9,3	303		13,6	200
8,7	265		13,4	230
14,2	102		13,2	213
15,2	105		12,5	124
14	114		11,9	240
15	116		10,5	320
13,6	120		13,4	180
14,5	121		15	100
12,3	116		15,3	98
11,3	210		14,8	145
11,8	200		14,2	150
11,4	215		12,6	230
			10,6	320

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y - \bar{y})^2}} = \frac{S_{XY}}{\sqrt{S_X{}^2 S_Y{}^2}}$$

$$= \frac{78877.3 - 3512.76 \bullet 185.14}{\sqrt{\left(5823.51 - 35 \bullet 12.763^2\right) \bullet \left(1377222 - 35 \bullet 185.1429^2\right)}}$$

$$= \frac{-3826.959145}{4657.518467} = \mathbf{-0.82167342}$$

***Según el valor de r, el número de casas vendidas mensualmente* (X), y las tasas de interés mensual (Y), poseen *una* relación lineal negativa o inversa, fuerte.**

Prueba de Hipótesis: Coeficiente de Correlación

A un **nivel de significación** de **0.01,** ¿la **correlación** entre ***las tasas de interés mensual (%)*** y el ***número de casas vendidas mensualmente*** **es significativa o no?**

1. Planteo:

H_0: ρ_{XY}
H_1: ρ_{XY}

2. α **=**

3. Estadístico de Prueba bajo H_0 cierta:

$$\frac{-}{\sqrt{\qquad}} =$$

4. Región Crítica:

$$\mathbf{RC} = \{ \qquad \}$$

5. El valor del estadístico es:

De los n =

como **r** =

$$_{obs} = \frac{\quad}{\sqrt{\quad}} =$$

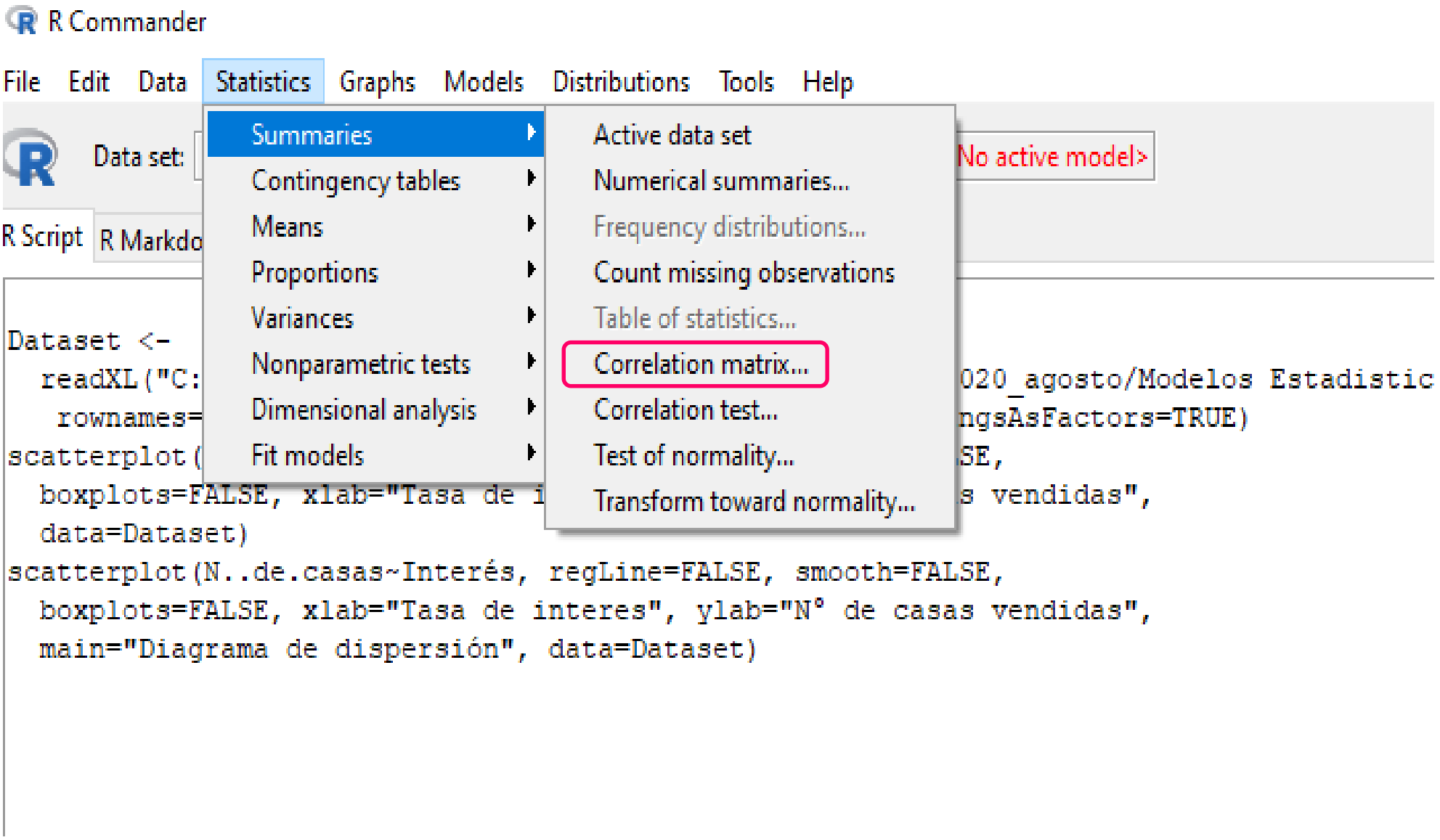
6. Comparación:

Como (la Región Crítica) entonces H_0.

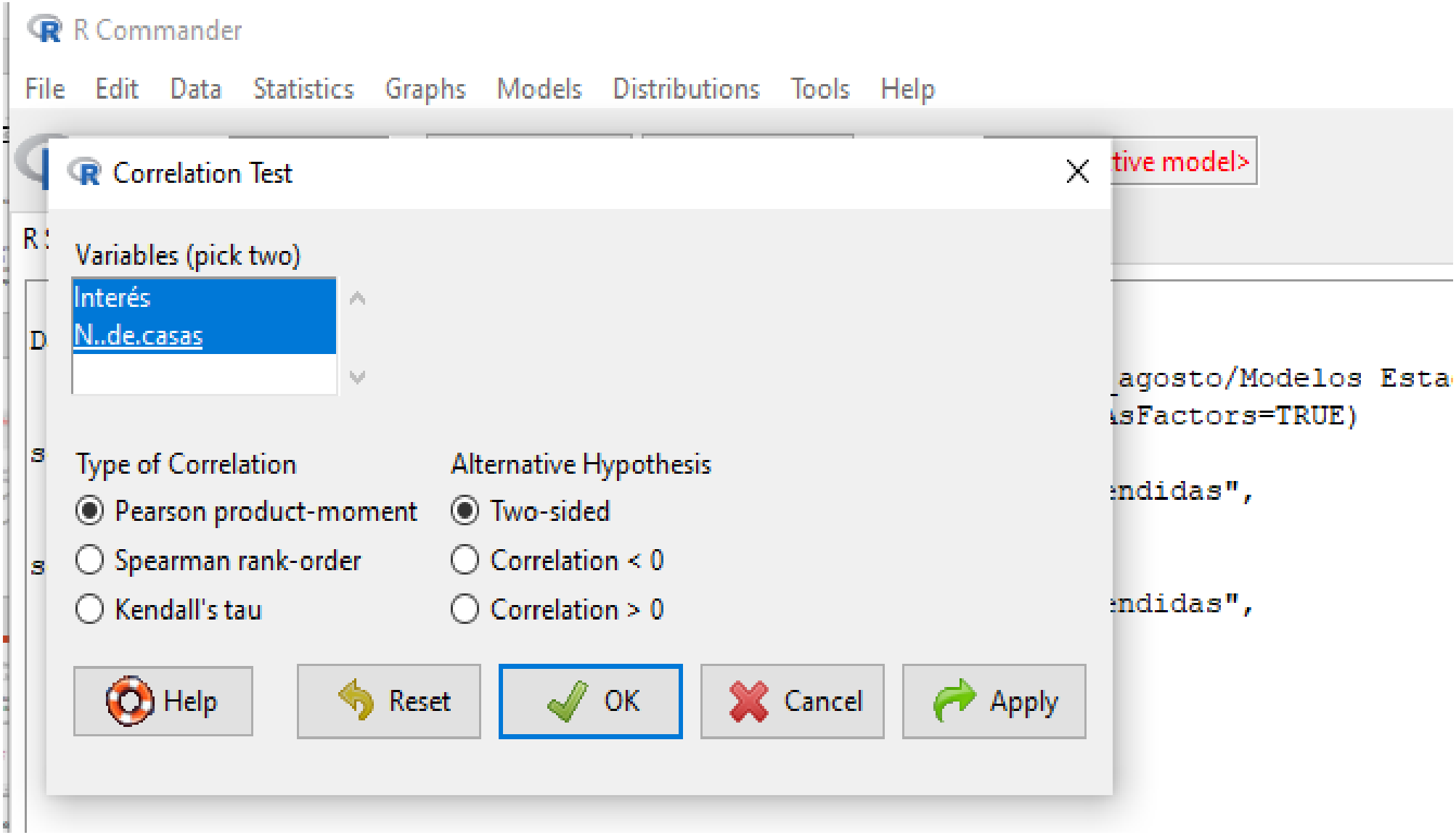
7. Conclusión:

Con , tengo evidencias suficientes para afirmar que

Ejemplo : Con Rcmdr



Ejemplo : Con Rcmdr



Ejemplo: Con Rcmdr

1. Planteo:

H_0: $\rho_{XY} = 0$ (No hay correlación entre la tasa de interés y el N° de casas vendidas en esa área.)
H_1: $\rho_{XY} \neq 0$ (Hay correlación entre la tasa de interés y el N° de casas vendidas en esa área.)

```
        Pearson's product-moment correlation

data:  Interés and N..de.casas
t = -8.262, df = 33, p-value = 0.000000001531
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9063121 -0.6715310
sample estimates:
       cor
-0.8210407
```

Valor p de la prueba

Estimación puntual de ρ_{XY}

Como **P = 0.000000001531** < **0.0001** **< 0.05**, entonces **Rechazo H_0**

Conclusión: Con una probabilidad de error menor a **0.0001**, tengo evidencias suficientes para afirmar que existe una **relación lineal** entre la tasa de interés y el N° de casas vendidas en esa área de la ciudad. Las v.a. están correlacionadas. **La relación es estadísticamente significativa (significa que es improbable que sea por azar).**

Ejercicio: Con Rcmdr

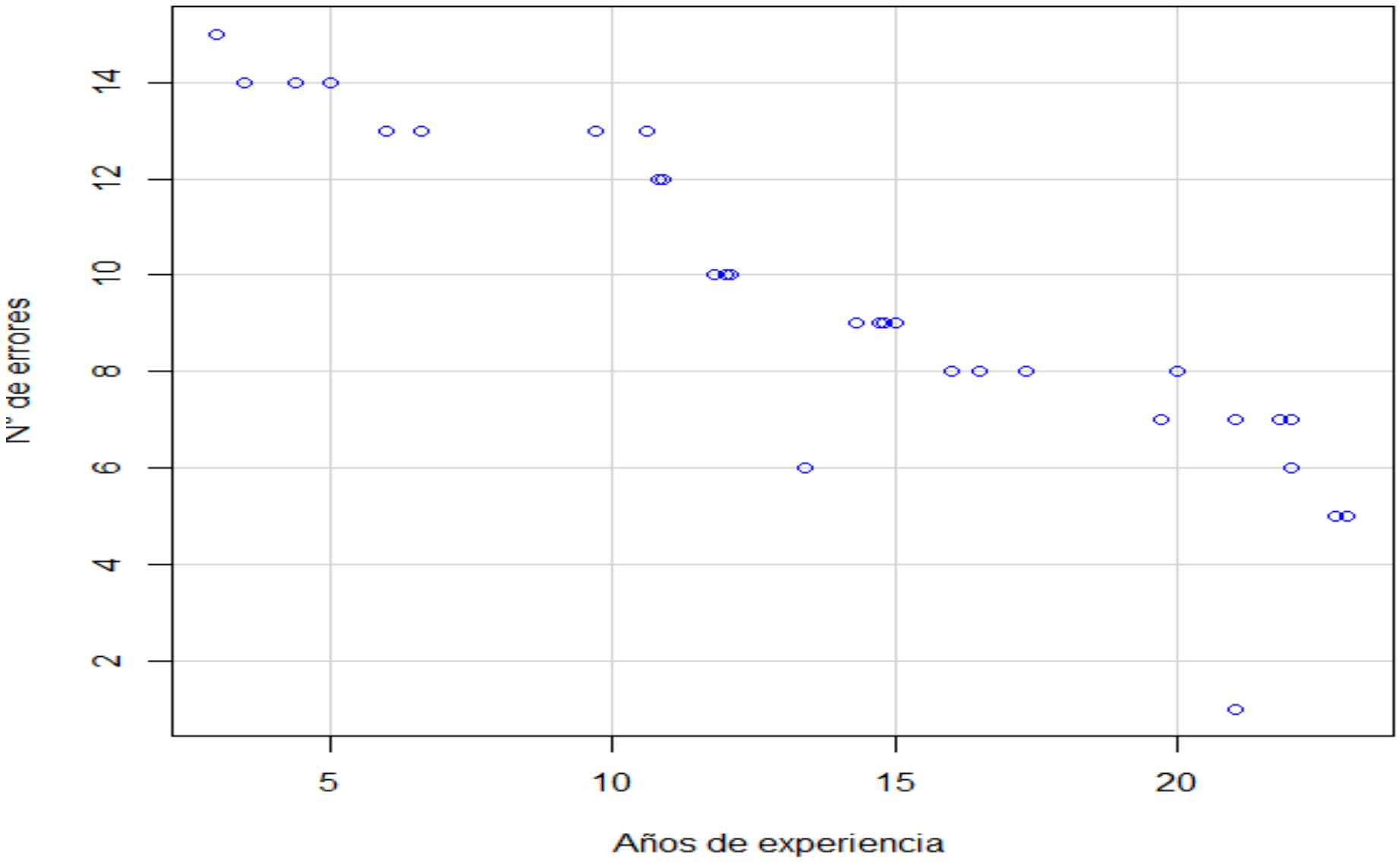
1. En la **Base_Ejer** se presentan los años de experiencia de un programador y el número de errores en el software creado por el mismo. El gerente de la empresa de software desea analizar cómo se relacionan estas dos variables aleatorias. Utilizando el **software Infostat:**

a) Construir el diagrama de dispersión.

b) ¿Las v.a. están correladas? Si lo están, interpretar el sentido de la relación en términos de las v.a.

c) ¿Es significativa la correlación entre ellas? Justificar la respuesta. No olvidarse de dar la conclusión en términos del problema.

a) Construir el **diagrama de dispersión**.



Ejercicio: Con Infostat

Las bodegas modernas utilizan vehículos guiados computarizados y automatizados para el manejo de materiales. En consecuencia, la disposición física de la bodega debe diseñarse con cuidado a modo de evitar el congestionamiento de los vehículos y optimar el tiempo de respuesta. En *The journal of Engineering for Industry* (agosto 1993) se estudió el diseño óptimo de una bodega automatizada. La disposición empleada supone que los vehículos no se bloquean entre sí cuando viajan dentro de la bodega, es decir, no haya congestionamiento. La validez de este supuesto se verificó simulando por computadora las operaciones de la bodega. En cada simulación se varió el número de vehículos y se registró el tiempo de congestionamiento (tiempo total que un vehículo bloquea a otro). Los datos se muestran en la tabla de abajo. Los investigadores están interesados en conocer la relación entre *el tiempo de congestionamiento* (**Y**) y el *número de vehículos* (**X**).

X	1	2	3	4	5	6	7	8	9	10
Y	0	0	0.02	0.01	0.01	0.01	0.03	0.03	0.02	0.04

a) Cuantificar la asociación lineal existente entre ambas variables.
b) ¿Es significativa la dependencia lineal entre las variables?. Usar el **valor P** para concluir.

Curiosidad !!!

Conjuntos de datos que son idénticos en varias propiedades estadísticas, pero producen gráficos diferentes, se utilizan con frecuencia para ilustrar la importancia de las representaciones gráficas cuando se exploraran datos. Se crean mediante el algoritmo **Simulated Anealing**.

Bases de datos:

- "DataSaurus" (first created by Alberto Cairo).
- "DataSaurus Dozen" (disponible en R package on GitHub).

Página web:

https://blog.revolutionanalytics.com/2017/05/the-datasaurus-dozen.html

Artículo:

Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing