(1994-
2024)
30 años de la
Consagración Constitucional
de la Autonomía y Autarquía
Universitaria en Argentina.

UNIVERSIDAD NACIONAL DEL LITORAL
LUX INDEFICIENS · 1919

FICH

ESTADÍSTICA

Clase #7 - Unidad 5 - *Estadística Descriptiva*

Ingenierías en: Recursos Hídricos, Ambiental y Agrimensura

Año 2024

Prof. Juan Pablo Taulamet

consultas: taulamet@unl.edu.ar

CLASIFICACIÓN

- TENDENCIA CENTRAL
- VARIABILIDAD
- FORMA
  - ASIMETRÍA
  - CURTOSIS

CARACTERÍSTICAS

- TENDENCIA CENTRAL
  - PROMEDIOS
  - UBICACIÓN

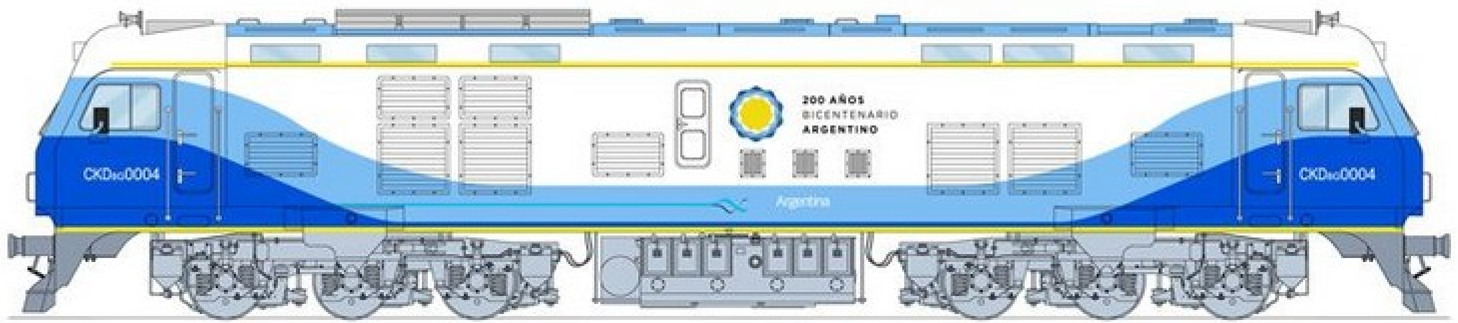
CARACTERÍSTICAS

- TENDENCIA CENTRAL
  - PROMEDIOS
    - ESPERANZA
  - UBICACIÓN
    - MEDIANA

Centro de masa de la distribución

Mitad de la frecuencia acumulada

consultas: taulamet@unl.edu.ar



Durante los últimos 10 días de junio un tren llegó tarde a su destino en los siguientes tiempos (en minutos)*:

-3	6	4	10	-4	124	2	-1	4	1

En cada caso, obtener los valores e interpretar:

a) Si el administrador del ferrocarril le contratara para demostrar que el tren está proporcionando un buen servicio, ¿Qué medidas utilizaría?

b) Si por el contrario se le contratara para demostrar que el ferrocarril está proporcionando un mal servicio, ¿Qué medida usaría?

c) Si quisiera buscar la objetividad en cuanto al desempeño del ferrocarril, ¿Qué medidas utilizaría?

* Considere que un número negativo significa que el tren llegó adelantado ese número de minutos.

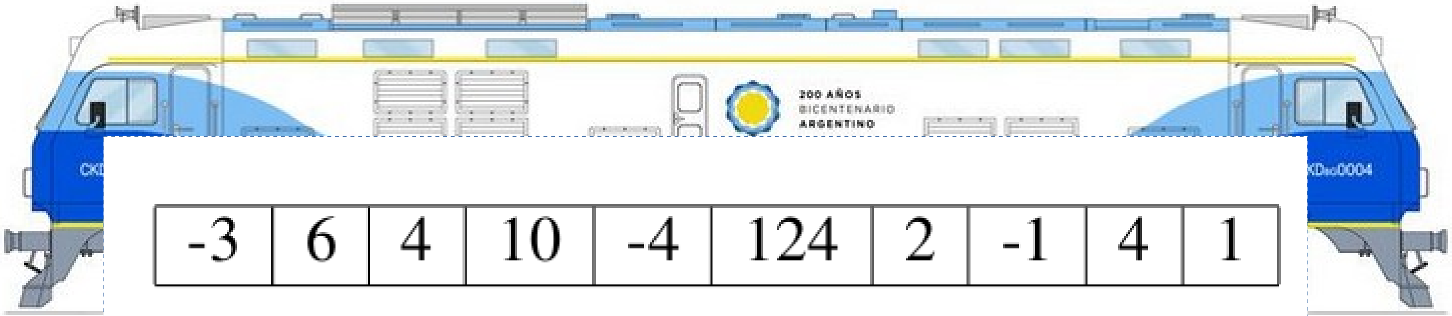
consultas: taulamet@unl.edu.ar

Estadística Descriptiva

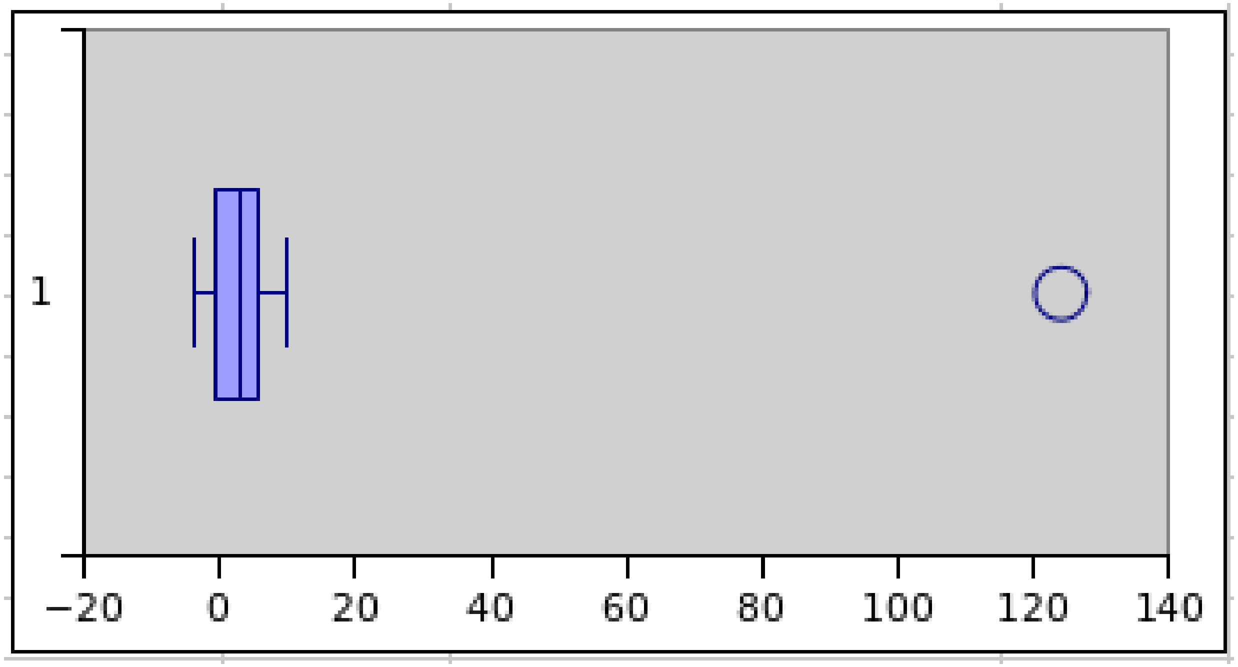
−4
−3
−1
1
2
4
4
6
10
124

Media	14,3
Error estándar	12,2610947
Moda	4
Mediana	3
Primer cuartil	-0,5
Tercer cuartil	5,5
Varianza	1503,34444
Desviación estándar	38,772986
Curtosis	9,67868027
Asimetría	3,09329907
Intervalo	128
Mínimo	-4
Máximo	124
Suma	143
Recuento	10

Box Plot



-3	6	4	10	-4	124	2	-1	4	1



ESTADÍSTICA
DIAPOSITIVAS DE EJEMPLO #5

Unidad 5 – *Estadística Descriptiva*

Ingeniería en Informática

Año 2022

Prof. Juan Pablo Taulamet

consultas: taulamet@unl.edu.ar

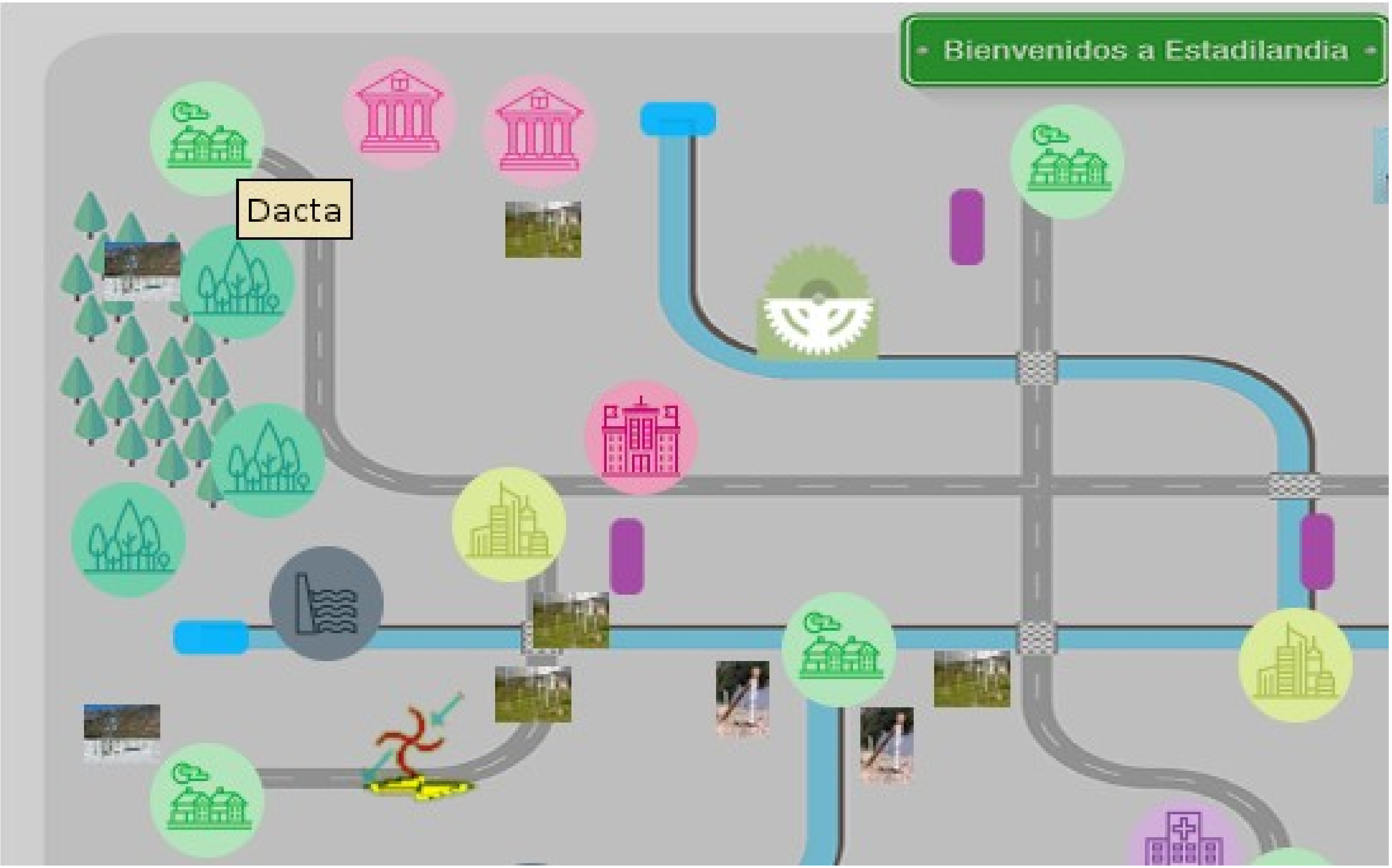
Análisis de Dacta

Se desea estudiar el comportamiento de las temperaturas en la población Dacta.

- Hallar la media muestral y el desvío.
- Encontrar los cuartiles e interpretar.
- Repetir los cálculos luego de eliminar la observación más pequeña. ¿Cómo se modifican las medidas?
- Construir un BoxPlot con los datos y comentar sobre los valores atípicos.

Buscamos los datos en Estadilandia

Mapa



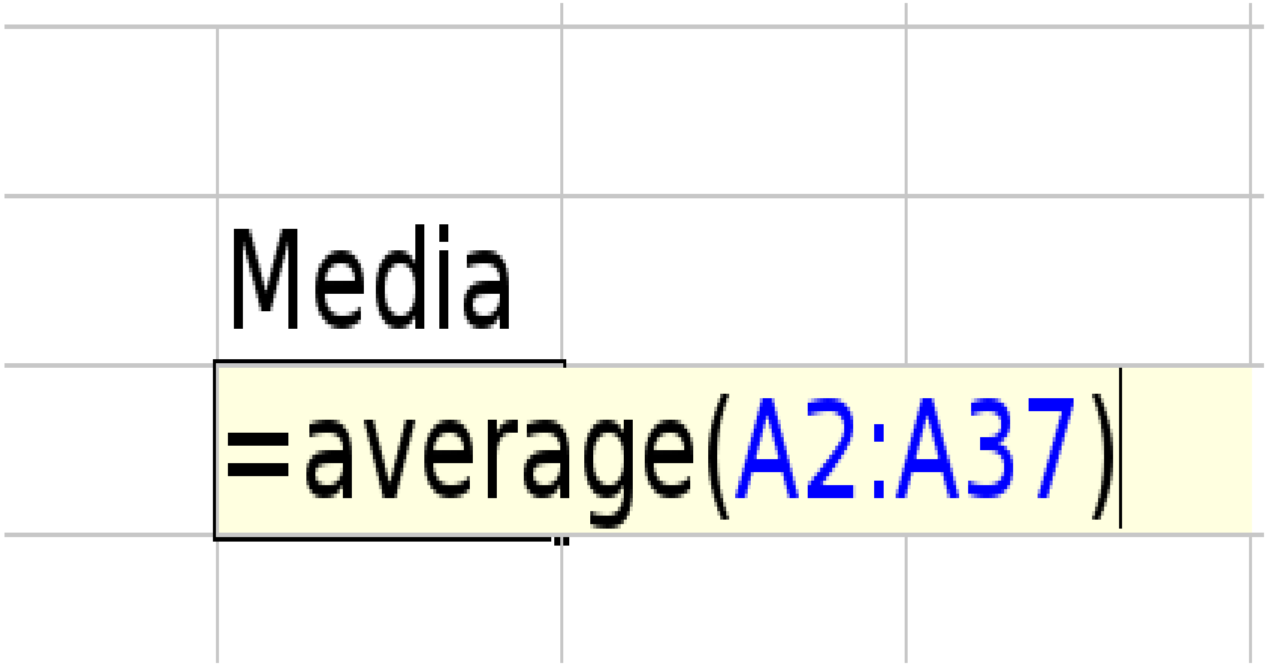
Descargando...

Dacta

Descargas [+]

[GNUMERIC] Temperaturas

Media



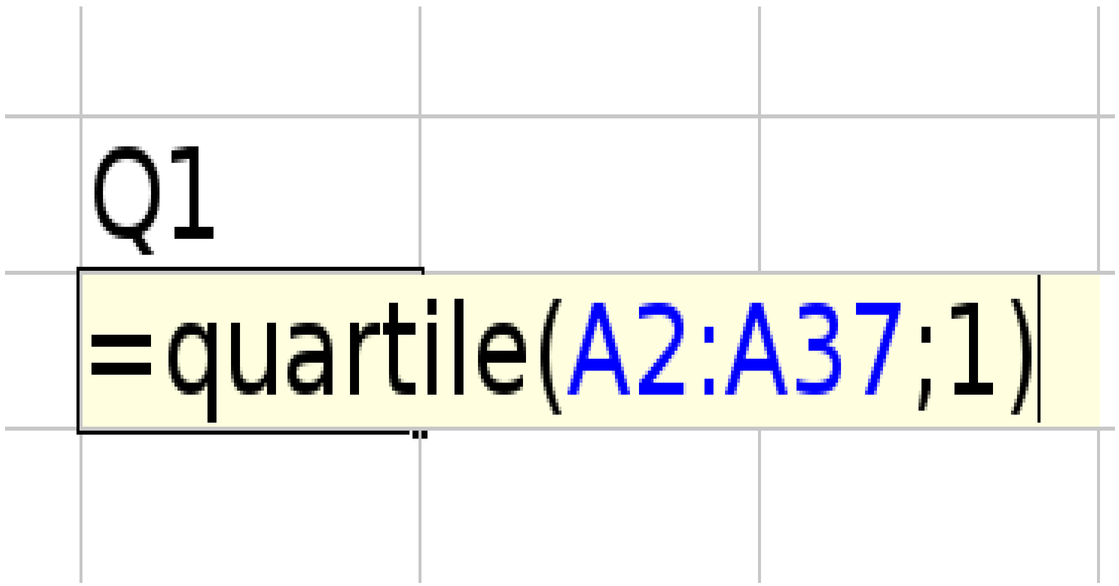
consultas: taulamet@unl.edu.ar

Desvío Estándar

Probar:

=stdev

Cuartiles



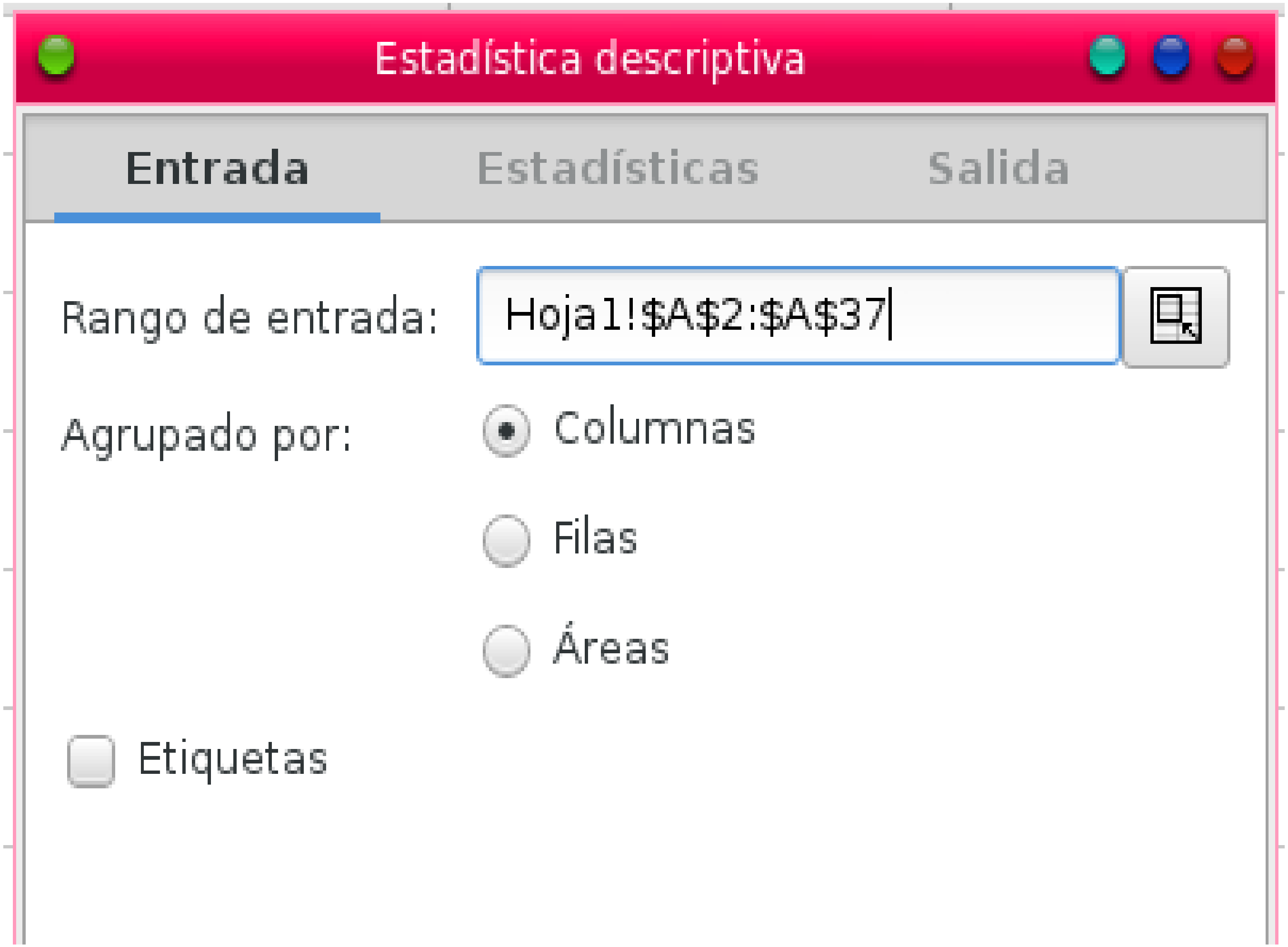
Para comentar…

(Reemplazando el mínimo por la media)

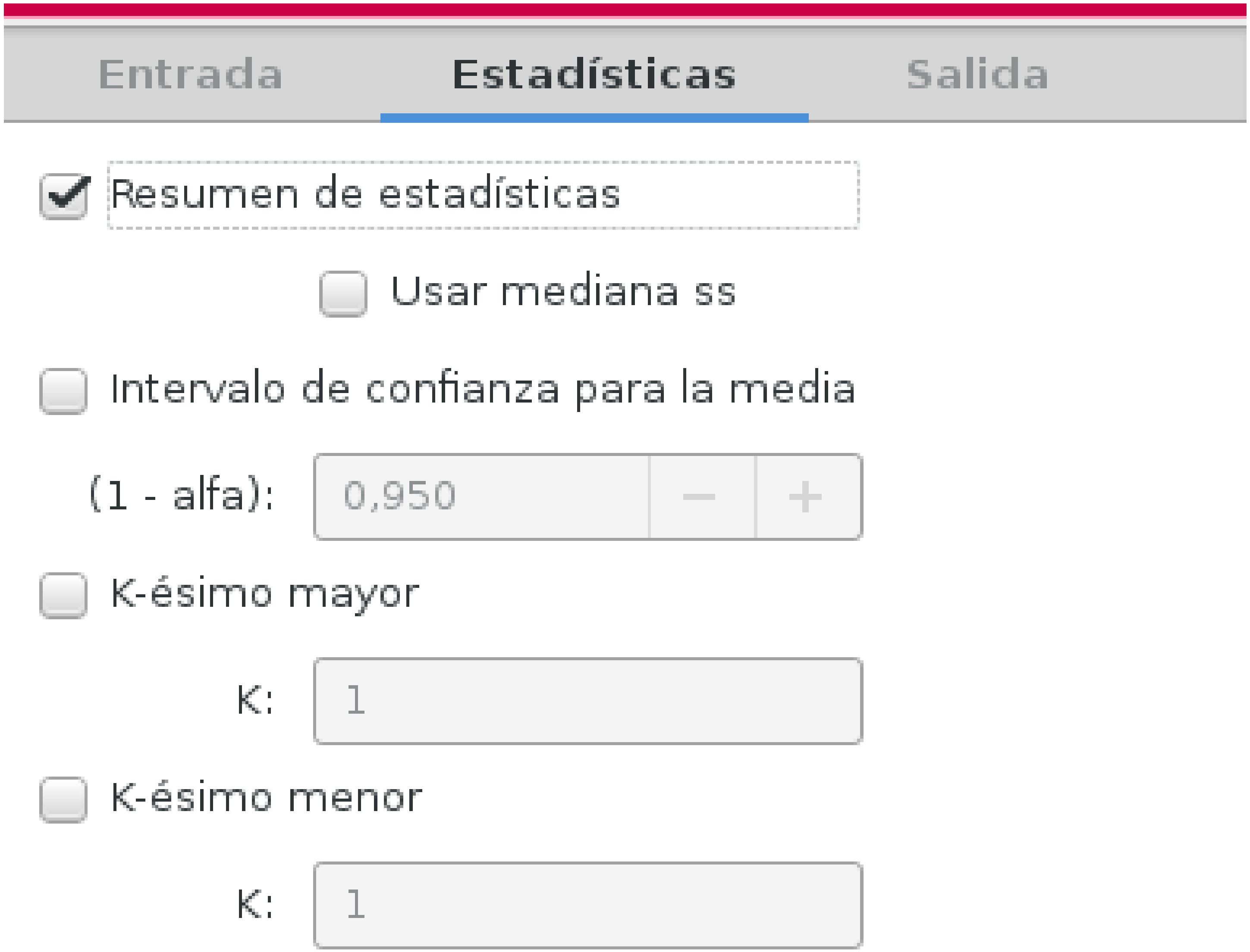
	Antes	Después	Dif.
Media	18,80	19,34	0,54
Desvío	6,76	5,89	−0,87
Q1	15,30	15,98	0,68
Q2	19,70	19,70	0,00
Q3	23,9	23,9	0,00
IRQ	8,6		
IRQ*1,5	12,9		

Descriptiva...

Descriptiva...



Descriptiva...

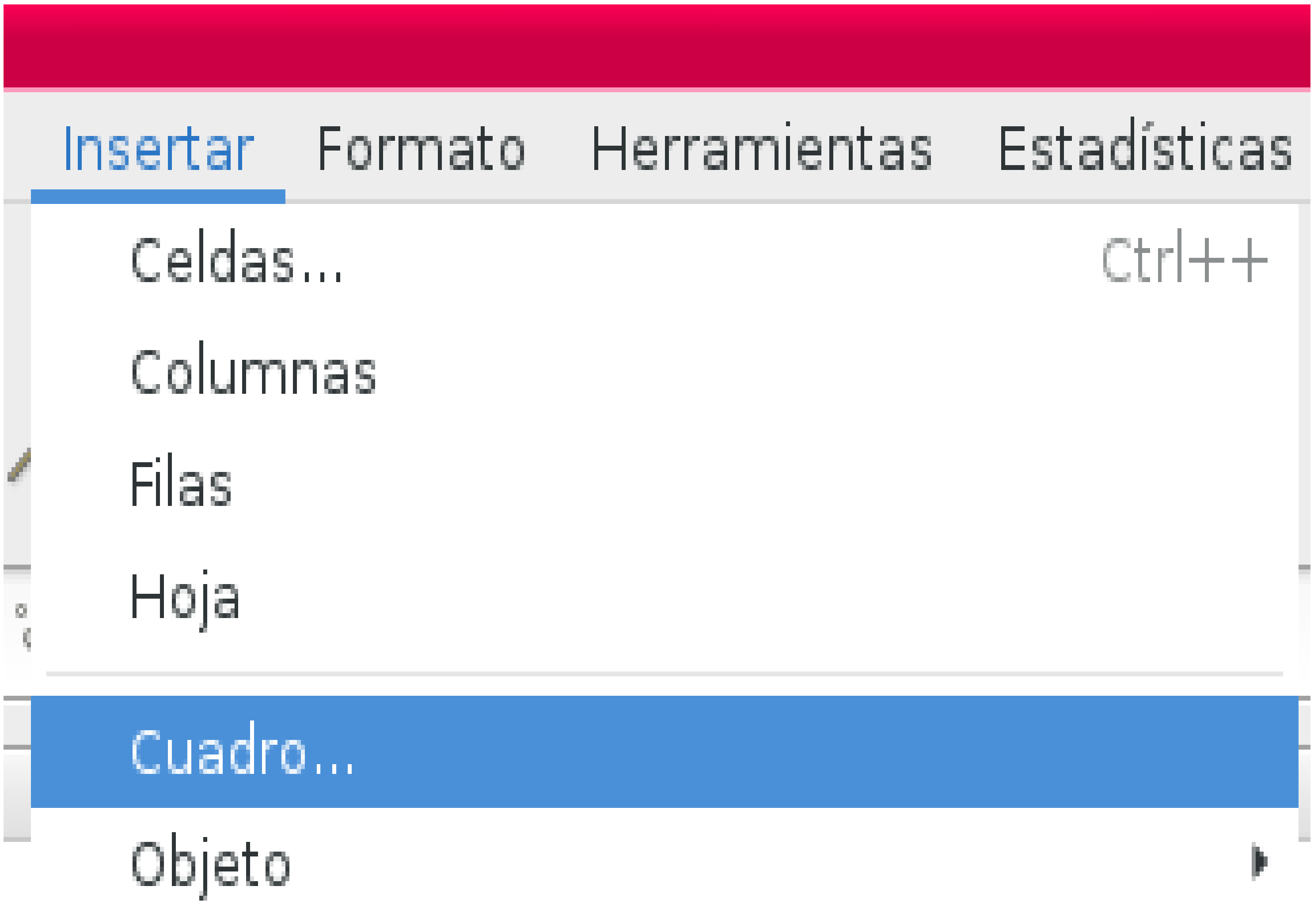


Descriptiva...

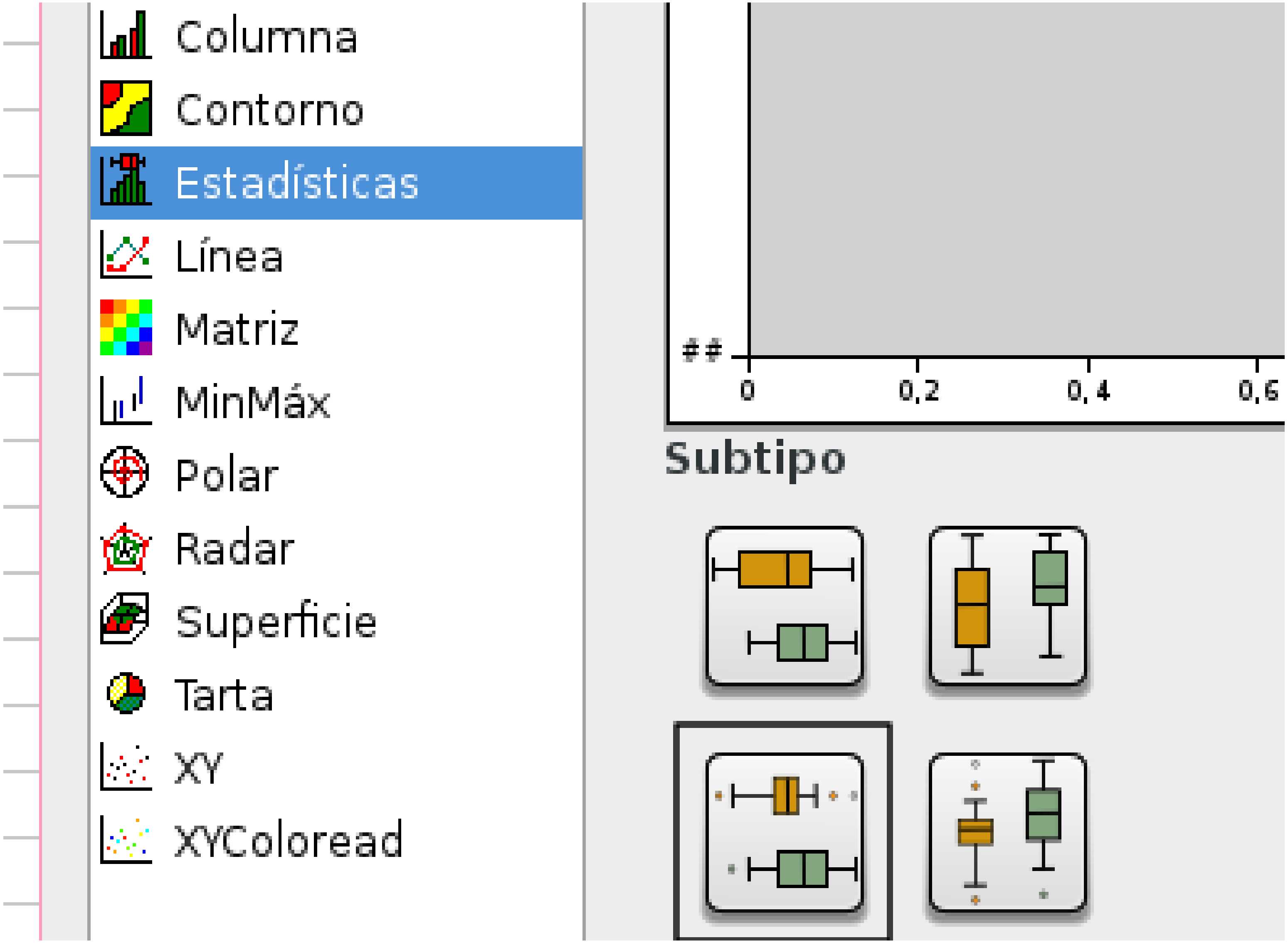
Media	18,8
Error estándar	1,127154
Mediana	19,7
Moda	19,4
Desviación estándar	6,7629241
Varianza de la muestra	45,737143
Curtosis	0,8495281
Desviación Ojo	−0,919714
Rango	29,5
Mínimo	−0,6
Máximo	28,9
Suma	676,8
Cuenta	36

consultas: taulamet@unl.edu.ar

Box Plot...


Insertar
Formato
Herramientas
Estadísticas
Celdas...
Ctrl++
Columnas
Filas
Hoja
Cuadro...
Objeto

Box Plot...


Columna
Contorno
Estadísticas
Línea
Matriz
MinMáx
Polar
Radar
Superficie
Tarta
XY
XYColoread
##
0
0,2
0,4
0,6
Subtipo

Box Plot...

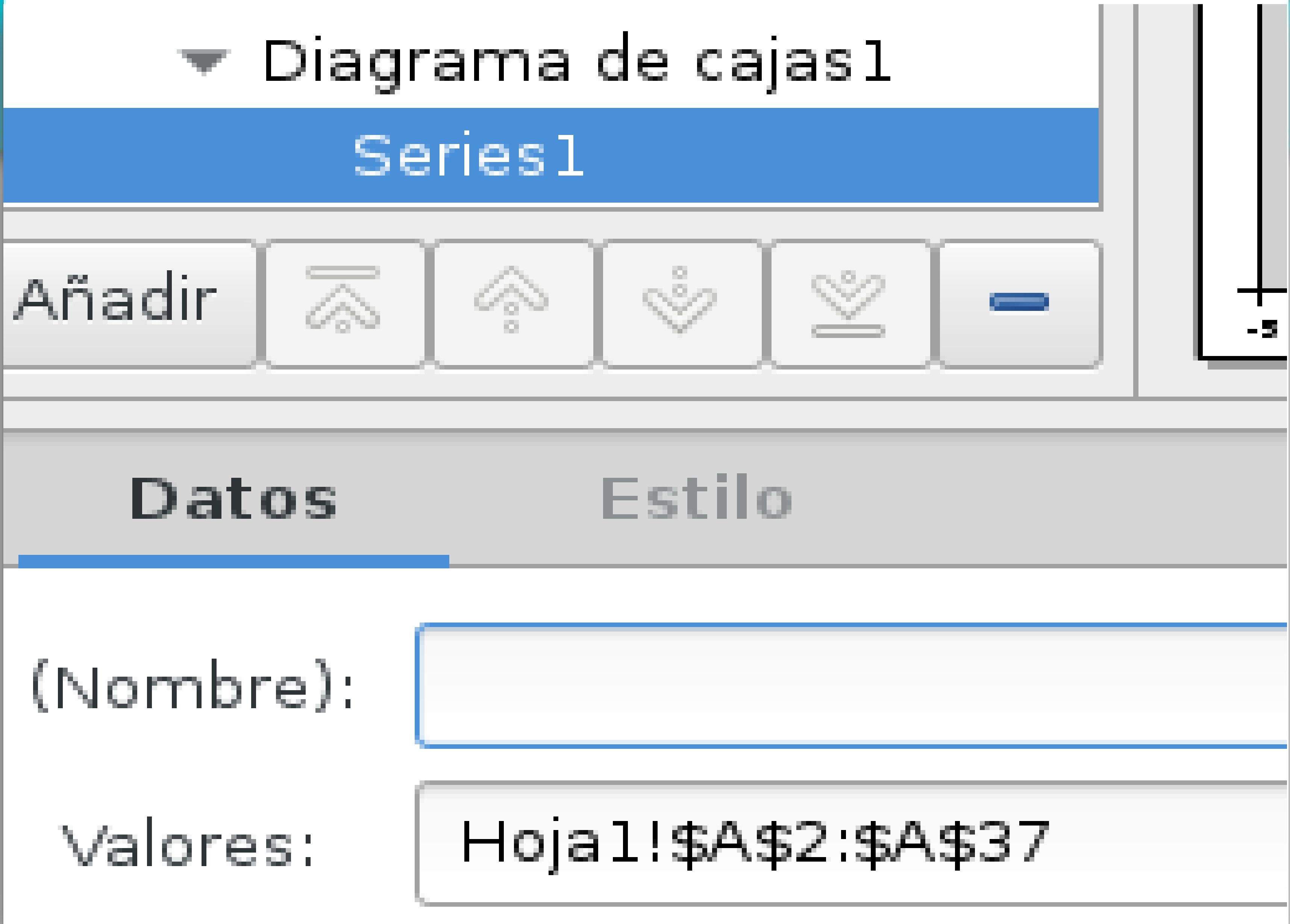
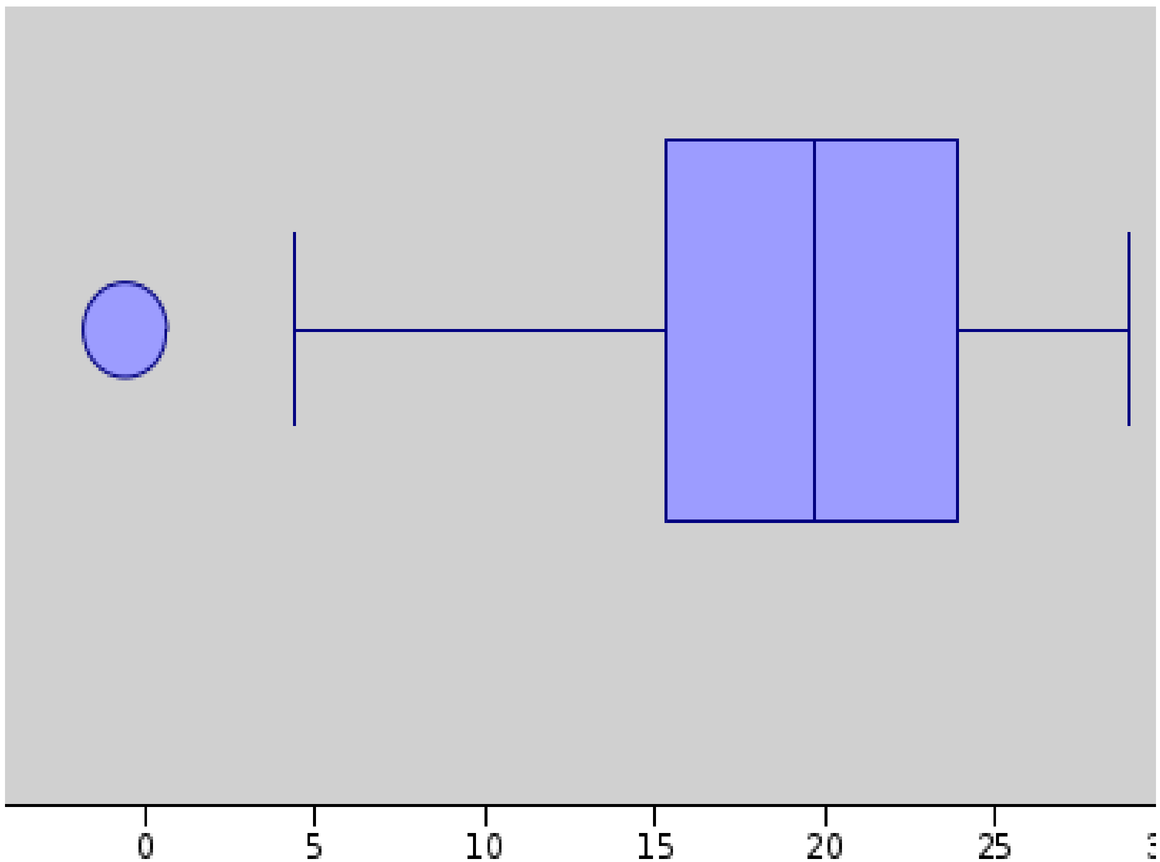


Diagrama de cajas1
Series1
Añadir
Datos
Estilo
(Nombre):
Valores:
Hoja1!$A$2:$A$37

Box Plot



Primer Parcial

Un profesor agrupa las calificaciones exámenes parciales corregidos, separándolos en teoría y práctica de la siguiente forma:

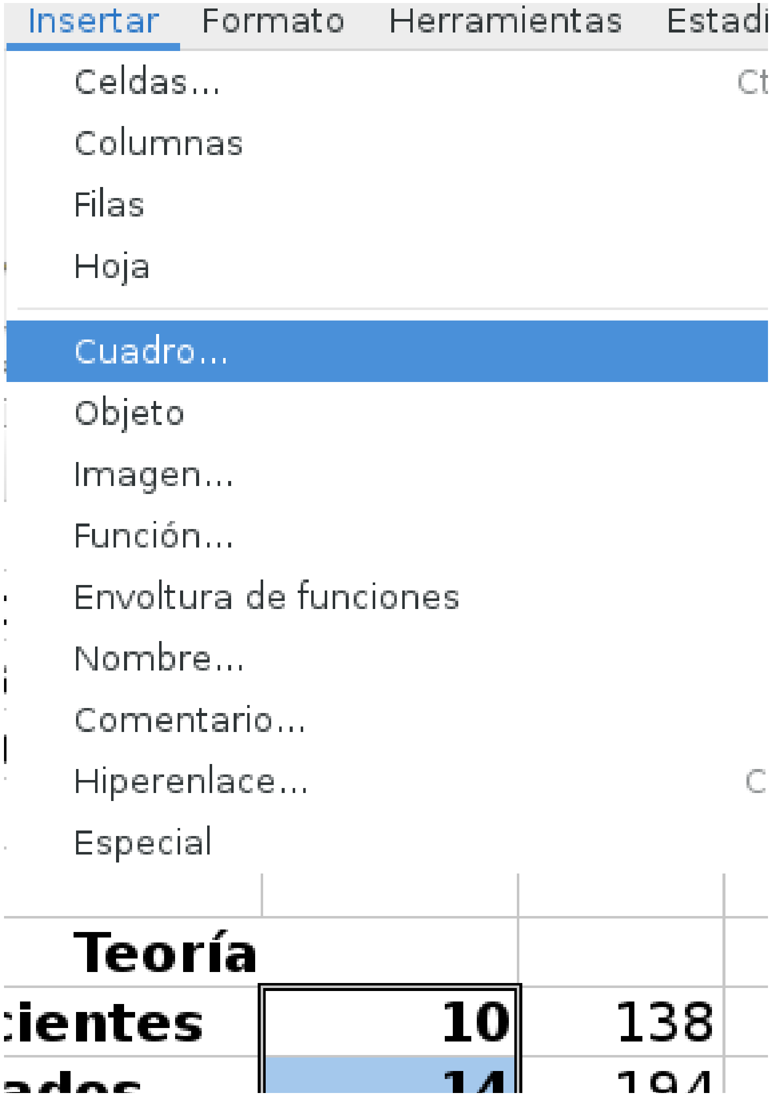
Práctica		Teoría	
Insuficientes	7	Insuficientes	10
Aprobados	12	Aprobados	14
Promocionados	4	Promocionados	2

Realizar un gráfico de sectores.

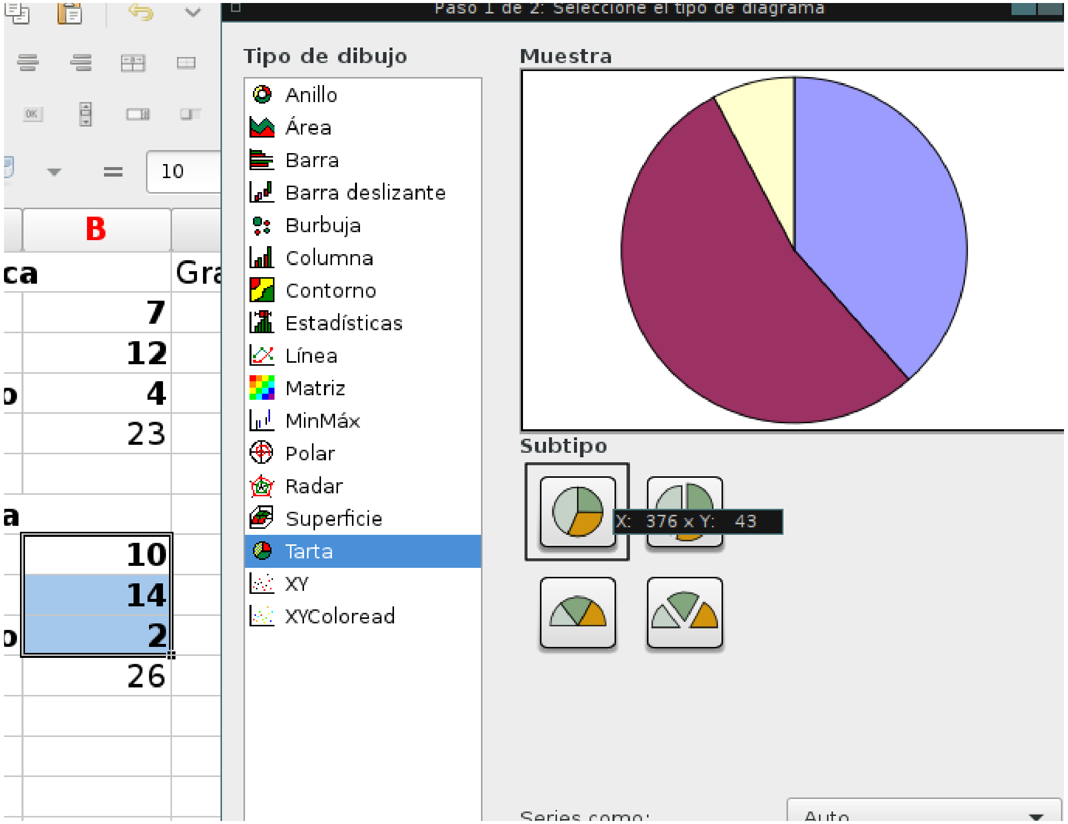
En Gnumeric

	A	B	C
1	**Práctica**		Grados
2	**Insuficientes**	**7**	110
3	**Aprobados**	**12**	188
4	**Promocionado**	**4**	63
5	Total	23	360
6			
7	**Teoría**		
8	**Insuficientes**	**10**	138
9	**Aprobados**	**14**	194
10	**Promocionado**	**2**	28
11	Total	26	360
12			

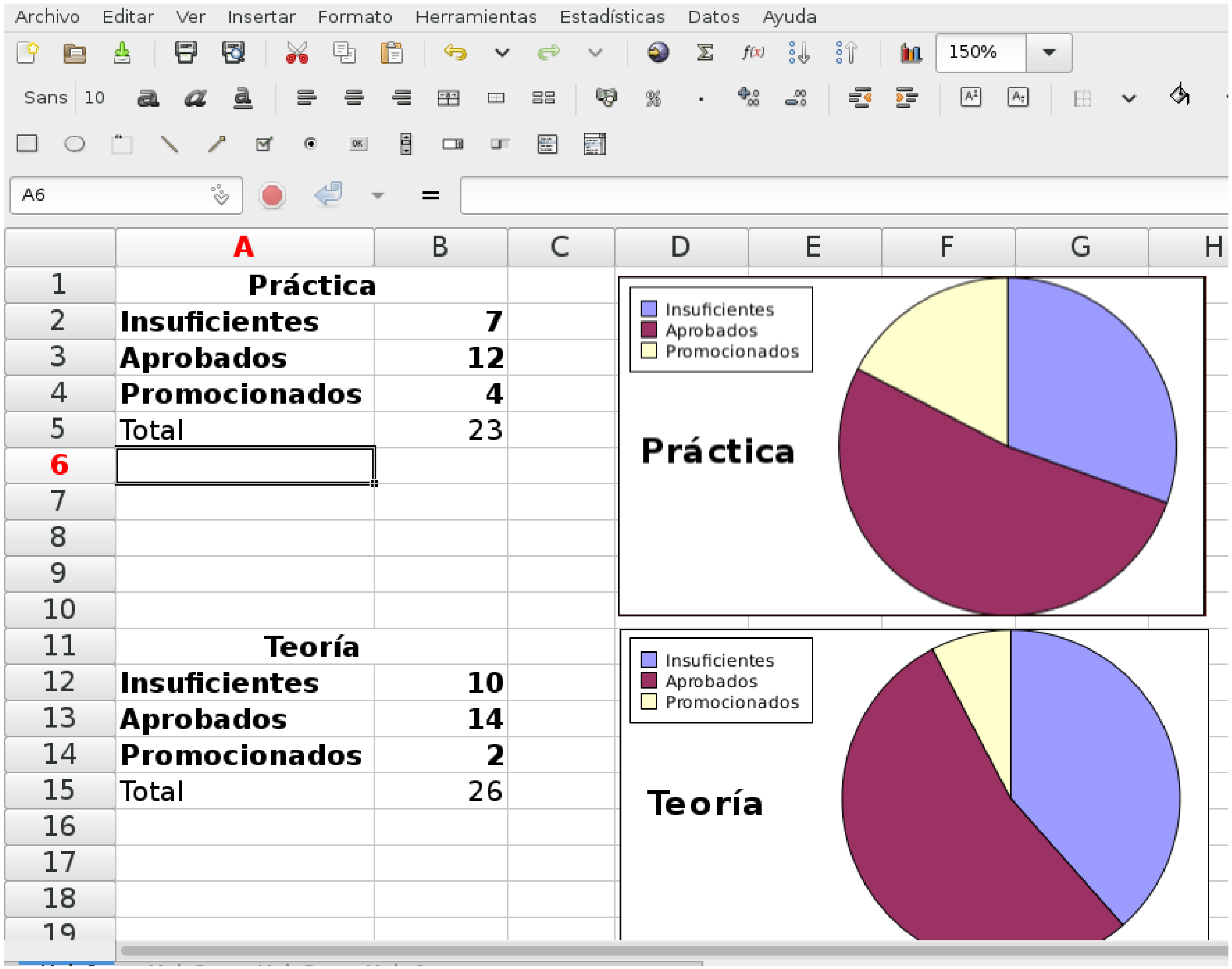
En Gnumeric


Insertar
Formato
Herramientas
Estadí
Celdas...
Columnas
Filas
Hoja
Cuadro...
Objeto
Imagen...
Función...
Envoltura de funciones
Nombre...
Comentario...
Hiperenlace...
Especial
Teoría
cientes
10
138

En Gnumeric


Paso 1 de 2: Seleccione el tipo de diagrama
Tipo de dibujo
Anillo
Área
Barra
Barra deslizante
Burbuja
Columna
Contorno
Estadísticas
Línea
Matriz
MinMáx
Polar
Radar
Superficie
Tarta
XY
XYColoread
Muestra
Subtipo
X: 376 x Y: 43
B
10
7
12
4
23
10
14
2
26

En Gnumeric



consultas: taulamet@unl.edu.ar

Problema

El gobierno de la provincia está estudiando la cantidad de hijos por familia en Santa Fe y cuenta con los siguientes datos provenientes de una muestra y necesita obtener una gráfica que le ayude a interpretar éstos datos:

2	4	2	3	1	2	4	2	3	0	2	2	2
3	2	6	2	3	2	2	3	2	3	3	4	3
3	4	5	2	0	3	2	1	2	3	2	2	3
1	4	2	3	2	4	3	3	2	2	1		

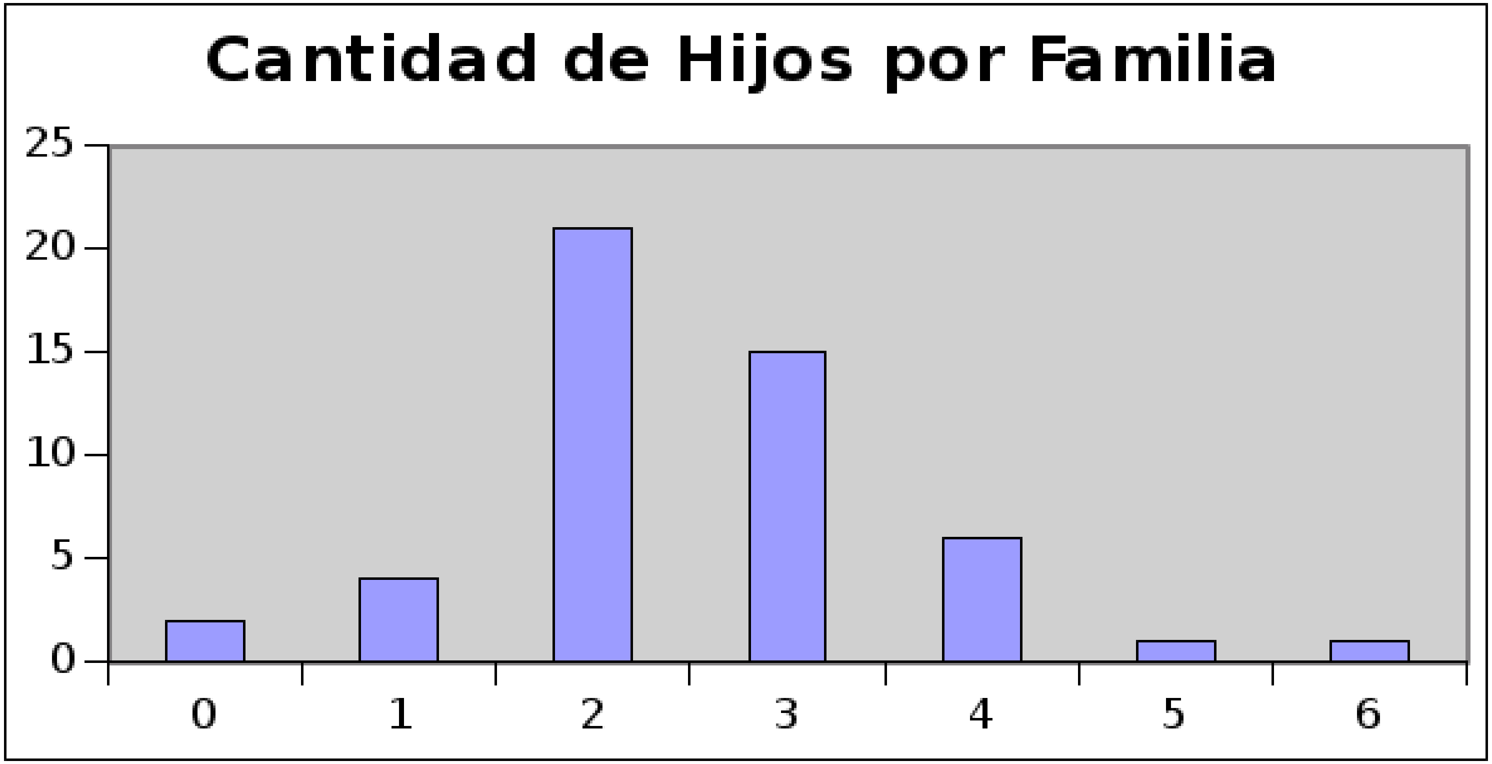
consultas: taulamet@unl.edu.ar

En Gnumeric

A	B	C	D	E
Datos		Agrupados		
2		0	=countif(A$2:A$51;C2)	
4		1	4	
2		2	21	
3		3	15	
1		4	6	
2		5	1	
4		6	1	
2			50	
3				
0				
2				

En Gnumeric

Para crear el Diagrama de Barras:
Menú Insertar, Cuadro, Diagrama de Columnas



Problema

Muchas de las personas que invierten en bolsa lo hacen para conseguir beneficios rápidos, por ello el tiempo en que mantienen las acciones es relativamente breve.

De una muestra de 40 inversores habituales sobre el tiempo en meses que han mantenido sus últimas inversiones se recogieron los siguientes datos:

Problema

10,5	12,7	3,8	9,1	13,6
11,2	11,4	10,5	13,4	14,7
9,9	11,6	11,7	12,3	11,5
15	6,2	8,4	5,9	11,5
11,4	7,9	12,5	11,4	10,9
12,7	8,3	11,2	8,8	9,8
16,5	10,9	9,1	7,4	12,9
10,1	8,1	10,4	8,6	9,9

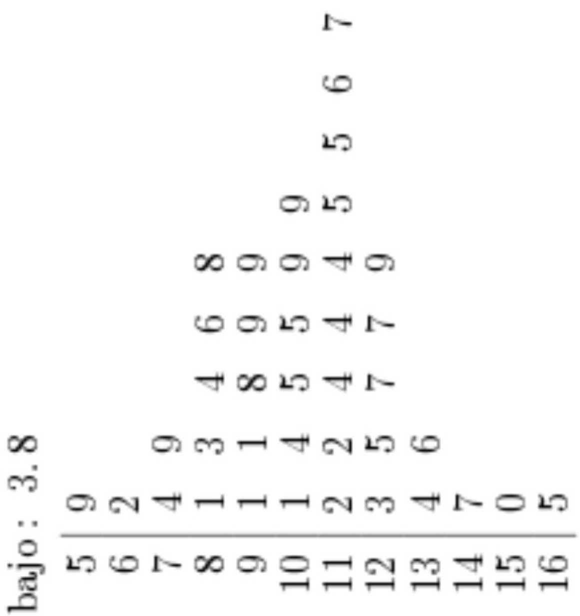
Represente los datos en un Diagrama de Tallo y Hoja

En GeoGebra

Tallos: Unidades
Hojas: Decimales

bajo : 3.8

5	9
6	2
7	4 9
8	1 3 4 6 8
9	1 1 8 9 9
10	1 4 5 5 9 9
11	2 2 4 4 4 5 5 6 7
12	3 5 7 7 9
13	4 6
14	7
15	0
16	5



Problema

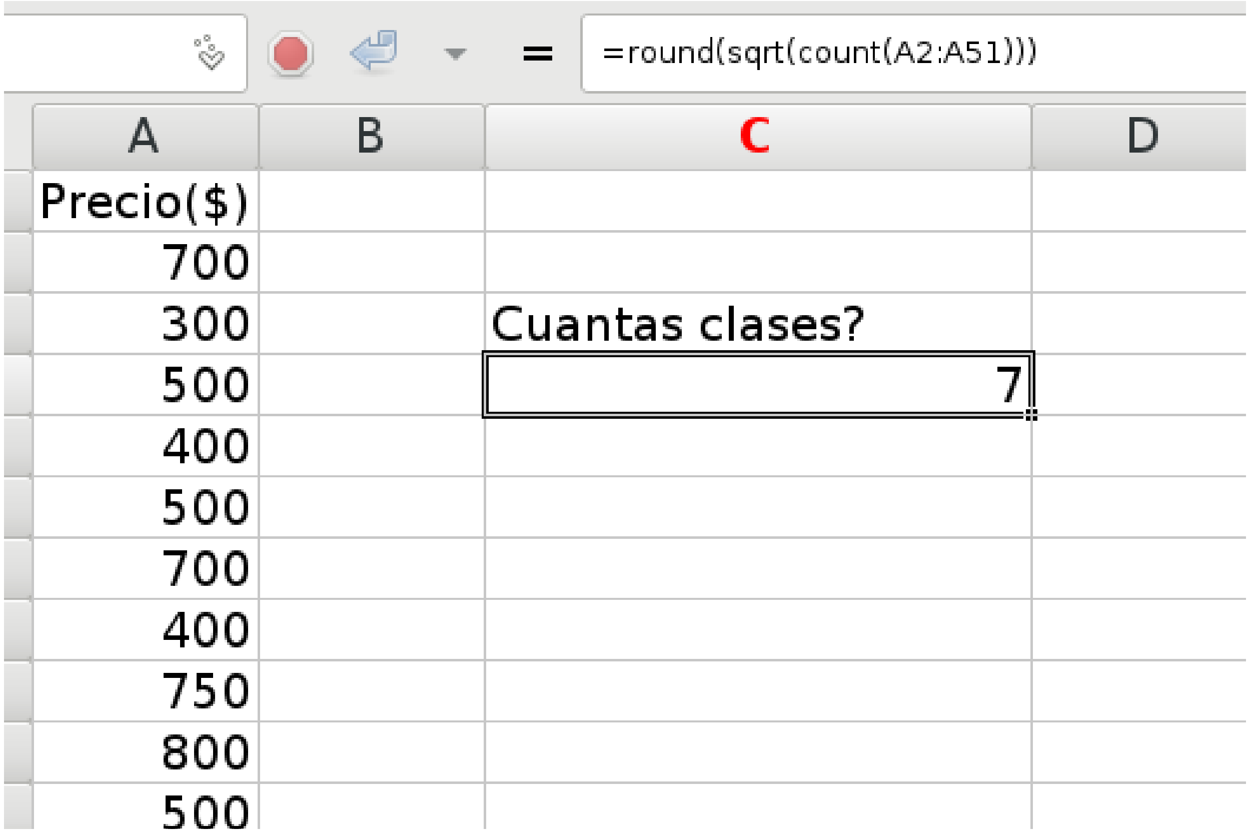
Construir un histograma de los precios ($) por habitación por noche de 50 hoteles de una ciudad:

700	300	500	400	500
500	750	300	700	1000
400	500	300	500	1000
300	400	700	400	700
700	800	750	700	750
700	400	750	800	500
1500	500	750	1200	800
300	400	500	700	500
500	400	700	1000	750
800	700	700	1200	800

consultas: taulamet@unl.edu.ar

En Gnumeric

- Cargar los datos en una columna
- Obtener la cantidad de clases como la raíz cuadrada de n redondeada:



En Gnumeric

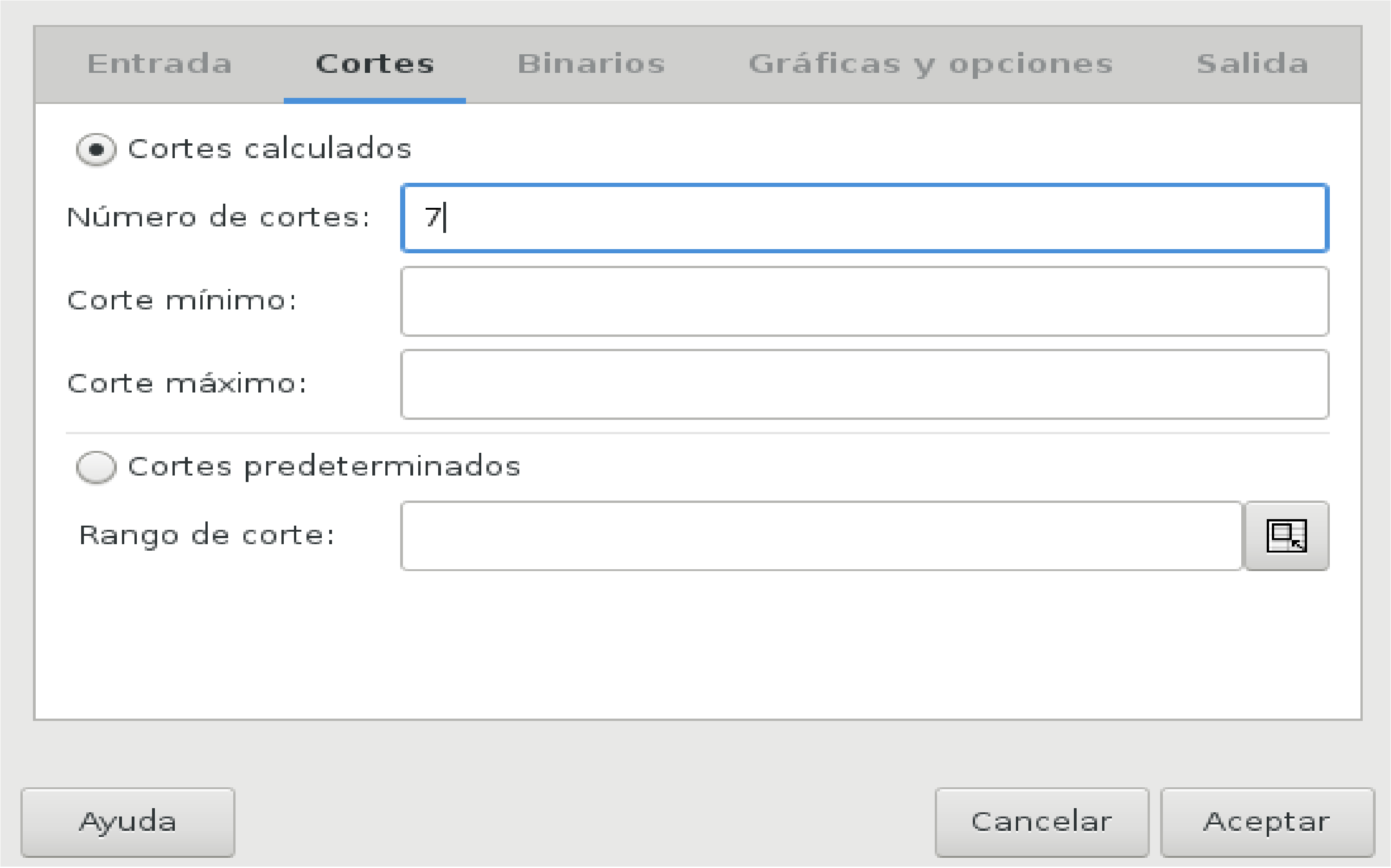
Obtenemos las Frecuencias Absolutas
Menús:

Estadística
→ Estadística Descriptiva
  → Tablas de Frecuencia
    → Histograma

Completamos
- Rango de Entrada
- Cortes Calculados
- Binarios

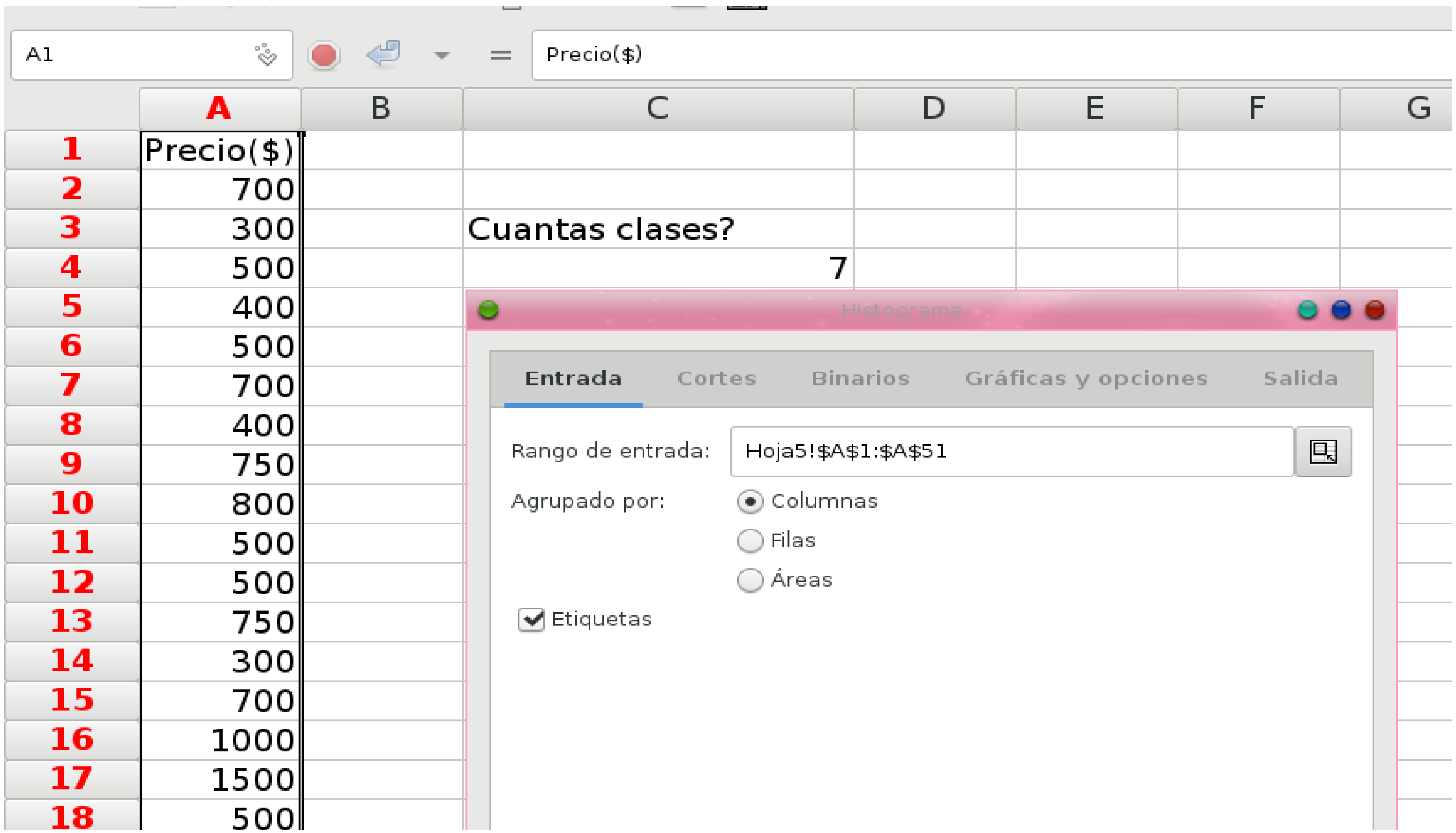
En Gnumeric

Obtenemos las Frecuencias Absolutas



En Gnumeric

Obtenemos las Frecuencias Absolutas



En Gnumeric

Obtenemos las Frecuencias Absolutas

consultas: taulamet@unl.edu.ar

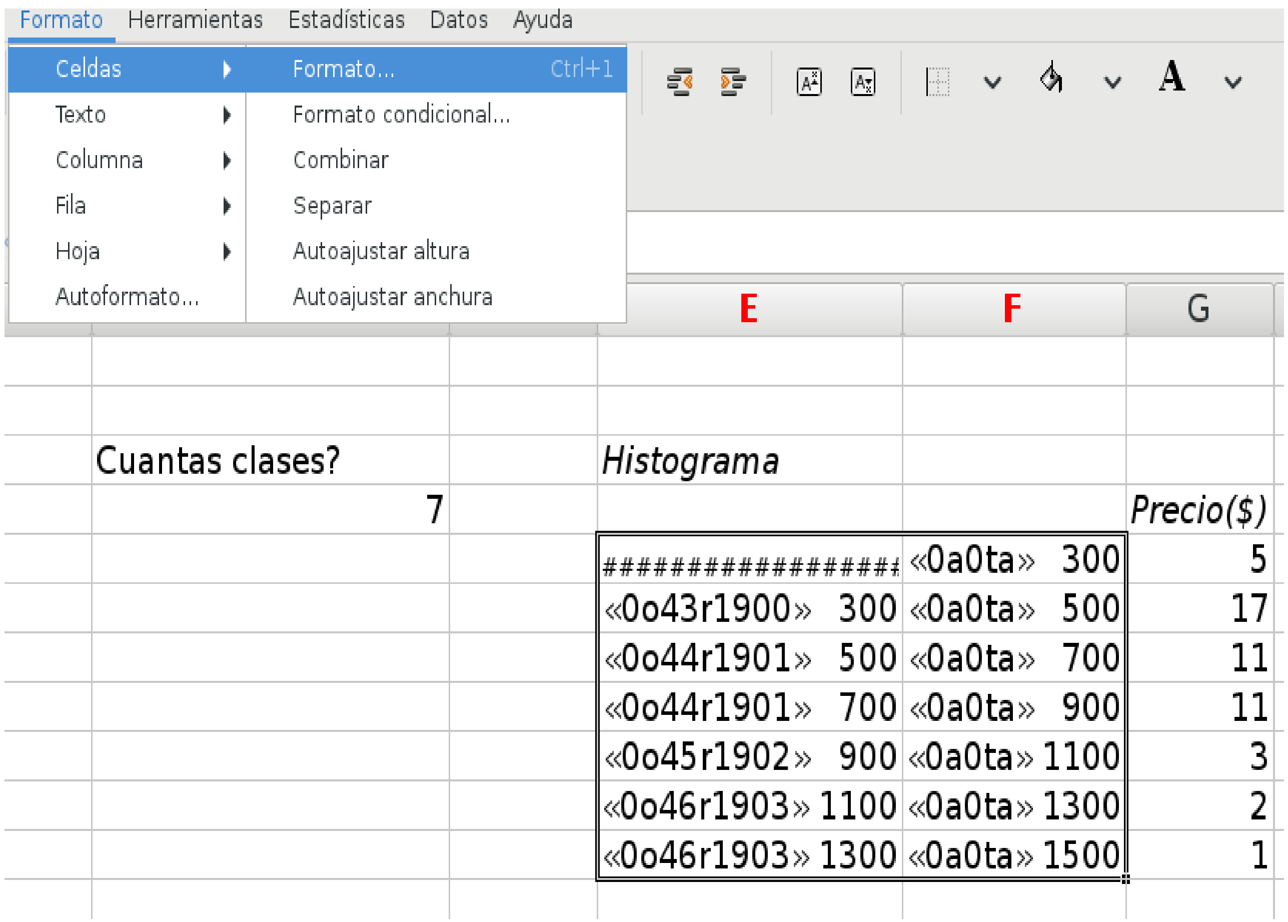
En Gnumeric

Obtenemos las Frecuencias Absolutas

	Histograma			
			Precio($)	
	##################	«0a0ta» 300	5	
	«0o43r1900» 300	«0a0ta» 500	17	
	«0o44r1901» 500	«0a0ta» 700	11	
	«0o44r1901» 700	«0a0ta» 900	11	
	«0o45r1902» 900	«0a0ta» 1100	3	
	«0o46r1903» 1100	«0a0ta» 1300	2	
	«0o46r1903» 1300	«0a0ta» 1500	1	

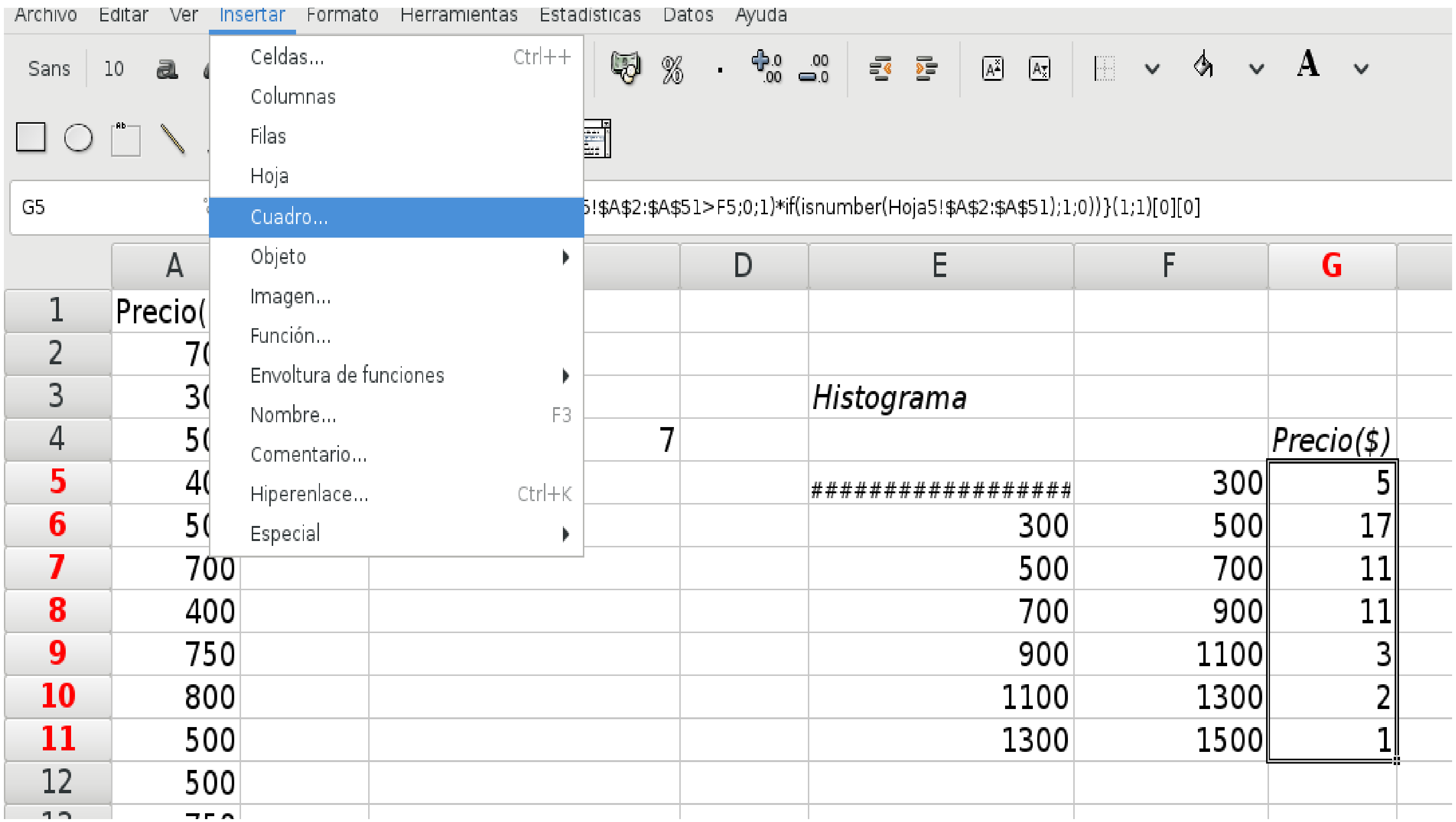
En Gnumeric

Mejoramos el formato a número general...



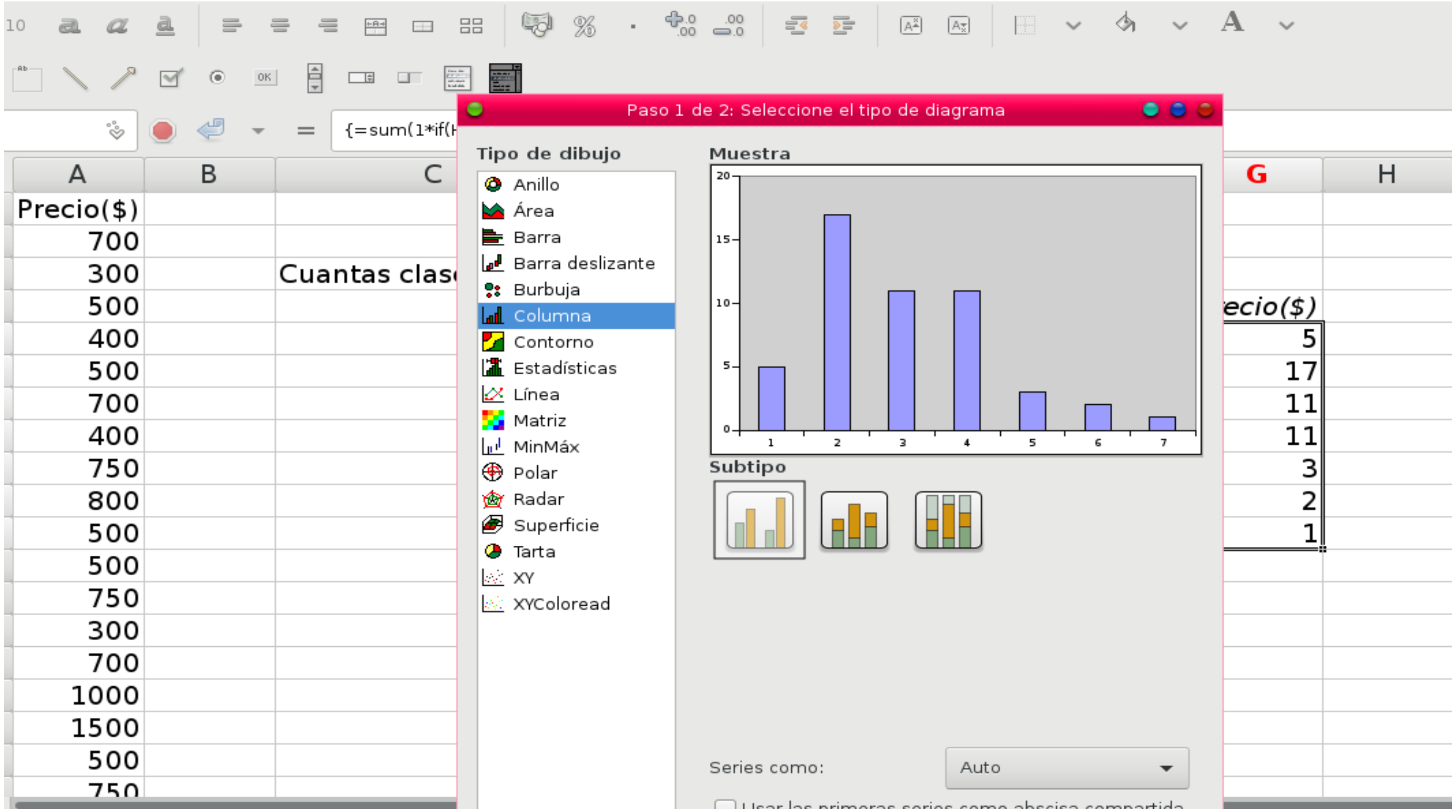
En Gnumeric

Seleccionamos las frecuencias y elegimos un diagrama



En Gnumeric

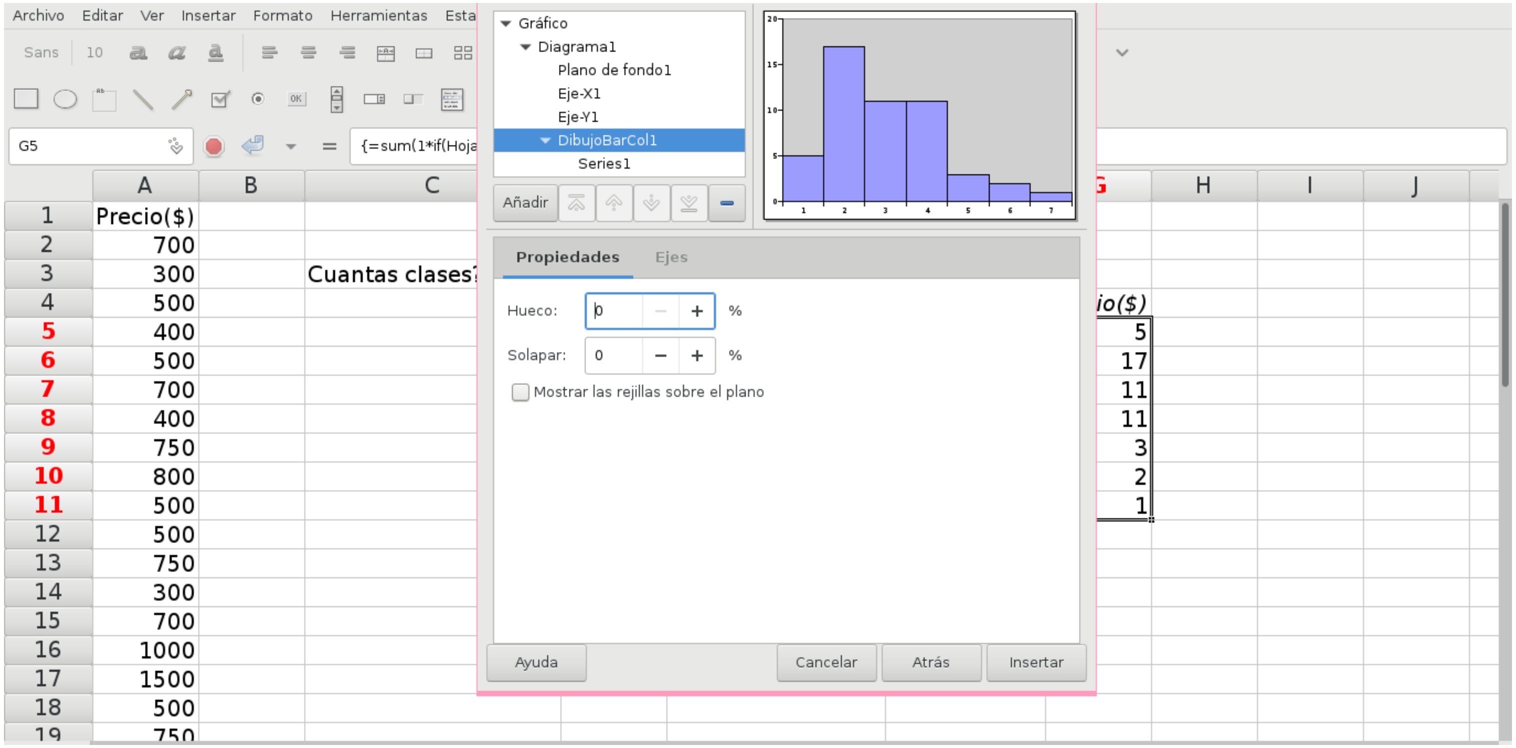
Insertamos el gráfico de columna



consultas: taulamet@unl.edu.ar

En Gnumeric

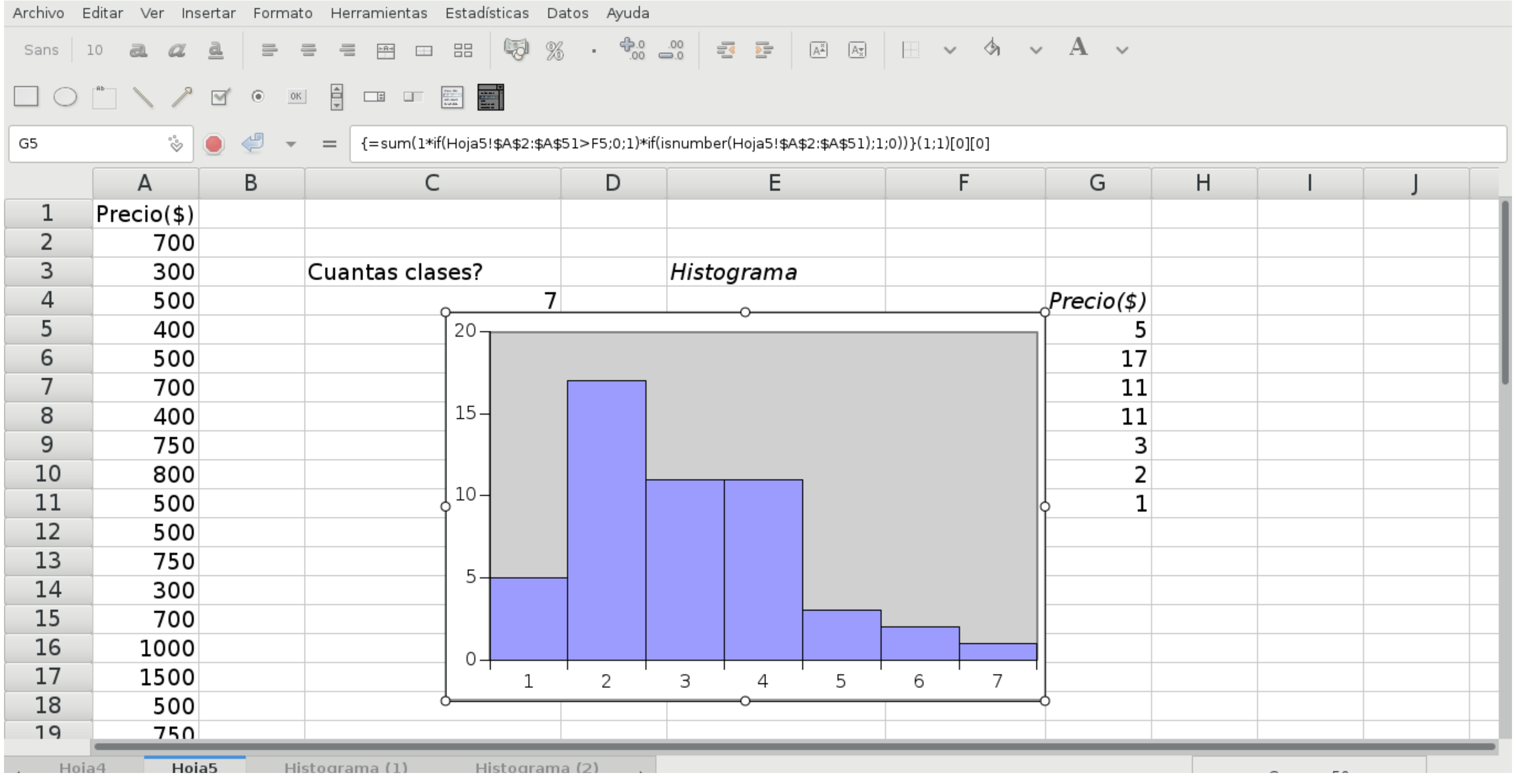
Quitamos los huecos desde las propiedades del gráfico



consultas: taulamet@unl.edu.ar

En Gnumeric

Y terminamos nuestro Histograma



consultas: taulamet@unl.edu.ar

(1994-2024)
30 años de la Consagración Constitucional de la Autonomía y Autarquía Universitaria en Argentina.

UNIVERSIDAD NACIONAL DEL LITORAL
LUX INDEFICIENS · 1919

FICH

ESTADÍSTICA
PROBABILIDAD Y ESTADÍSTICA

Clase #8 - Unidad 6 - *Estadística Inferencial*
Distribución en el muestreo - Intervalos de confianza

Ingenierías en: Recursos Hídricos, Ambiental y Agrimensura

Año 2024

Prof. Juan Pablo Taulamet

consultas: taulamet@unl.edu.ar

Problema

Dada la siguiente población de elementos:

$$X = \{22, 24, 26\}$$

a) Enumerar todas las muestras posibles de tamaño dos, escogidas mediante muestreo aleatorio simple.

b) Obtener la esperanza y varianza de la población y de la media muestral.

consultas: taulamet@unl.edu.ar

Planteo

a) Diferenciaremos los dos casos de muestreo: con y sin reposición.

b) Se pide:

$$E(X), V(X), E(\bar{X}), V(\bar{X})$$

Es decir:

$$\mu, \sigma^2, \mu_{\bar{X}}, \sigma^2_{\bar{X}}$$

Planteo

Por último intentaremos confirmar que:

$$\mu_{\bar{X}} = \mu$$

$$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n} \text{ (Con Rep.)}$$

$$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n} * \frac{N-n}{N-1} \text{ (Sin Rep.)}$$

Esperanza de X

Los valores de la población tienen igual probabilidad de ocurrencia por tanto su función de cuantía es:

X	22	24	26
$f(X)$	$1/3$	$1/3$	$1/3$

Luego:

$$\mu = \alpha_1 = \sum_{\forall x_i} x_i^1 * f(x_i) =$$

$$= \frac{22}{3} + \frac{24}{3} + \frac{26}{3} = 24$$

Varianza de X

$$\sigma^2 = \alpha_2 - \alpha_1^2 =$$

$$[\sum_{\forall x_i} x_i^2 * f(x_i)] - 24^2 =$$

$$= [\frac{22^2}{3} + \frac{24^2}{3} + \frac{26^2}{3}] - 576 =$$

$$\frac{8}{3} \approx 2.67$$

Esperanza de la media muestral

(Caso con reposición)

M1: 22, 24 $\bar{X}_1 = 23$
M2: 22, 26 $\bar{X}_2 = 24$
M3: 24, 26 $\bar{X}_3 = 25$

M4: 22, 22 $\bar{X}_4 = 22$
M5: 24, 24 $\bar{X}_5 = 24$
M6: 26, 26 $\bar{X}_6 = 26$

M7: 24, 22 $\bar{X}_7 = 23$
M8: 26, 22 $\bar{X}_8 = 24$
M9: 26, 24 $\bar{X}_9 = 25$

X	22	23	24	25	26
$f(X)$	1/9	2/9	3/9	2/9	1/9

$$\mu_{\bar{X}} = \sum_{\forall \bar{x}_i} \bar{X}_i * f(\bar{X}_i) =$$

$$= \frac{22*1}{9} + \frac{23*2}{9} + \frac{24*3}{9} + \frac{25*2}{9} + \frac{26*1}{9} = 24$$

Varianza de media muestral

(Caso <u>con</u> reposición)

$$\sigma^2 = \alpha_2 - \alpha_1^2 =$$

$$[\sum_{\forall \bar{x}_i} x_i^2 * f(\bar{x}_i)] - 24^2 =$$

$$= [\frac{22^2 * 1}{9} + \frac{23^2 * 2}{9} + \frac{24^2 * 3}{9} +$$

$$\frac{25^2 * 2}{9} + \frac{26^2 * 1}{9}] - 576 = \frac{4}{3} \approx 1.34$$

Esperanza de la media muestral

(Caso <u>sin</u> reposición)

M1: 22, 24 $\overline{X}_1 = 23$
M2: 22, 26 $\overline{X}_2 = 24$

M3: 24, 26 $\overline{X}_3 = 25$
M4: 24, 22 $\overline{X}_4 = 23$

M5: 26, 22 $\overline{X}_5 = 24$
M6: 26, 24 $\overline{X}_6 = 25$

X	23	24	25
$f(X)$	$2/6$	$2/6$	$2/6$

$$\mu_{\bar{X}} = \sum_{\forall \bar{x}_i} \bar{X}_i * f(\bar{X}_i) =$$

$$= \frac{23*2}{6} + \frac{24*2}{6} + \frac{25*2}{6} = 24$$

Varianza de media muestral

(Caso sin reposición)

$$\sigma^2 = \alpha_2 - \alpha_1^2 =$$

$$[\sum_{\forall \bar{x}_i} x_i^2 * f(\bar{x}_i)] - 24^2 =$$

$$= [\frac{23^2 * 2}{6} + \frac{24^2 * 2}{6} + \frac{25^2 * 2}{6}] - 576 =$$

$$\frac{2}{3} \approx 0.67$$

Conclusiones

$$\mu = 24 \qquad \sigma^2 = \frac{8}{3}$$

Caso con rep.

$\mu_{\bar{X}} = 24$
$\sigma^2_{\bar{X}} = 4/3$
n = 2

Se verifica que:

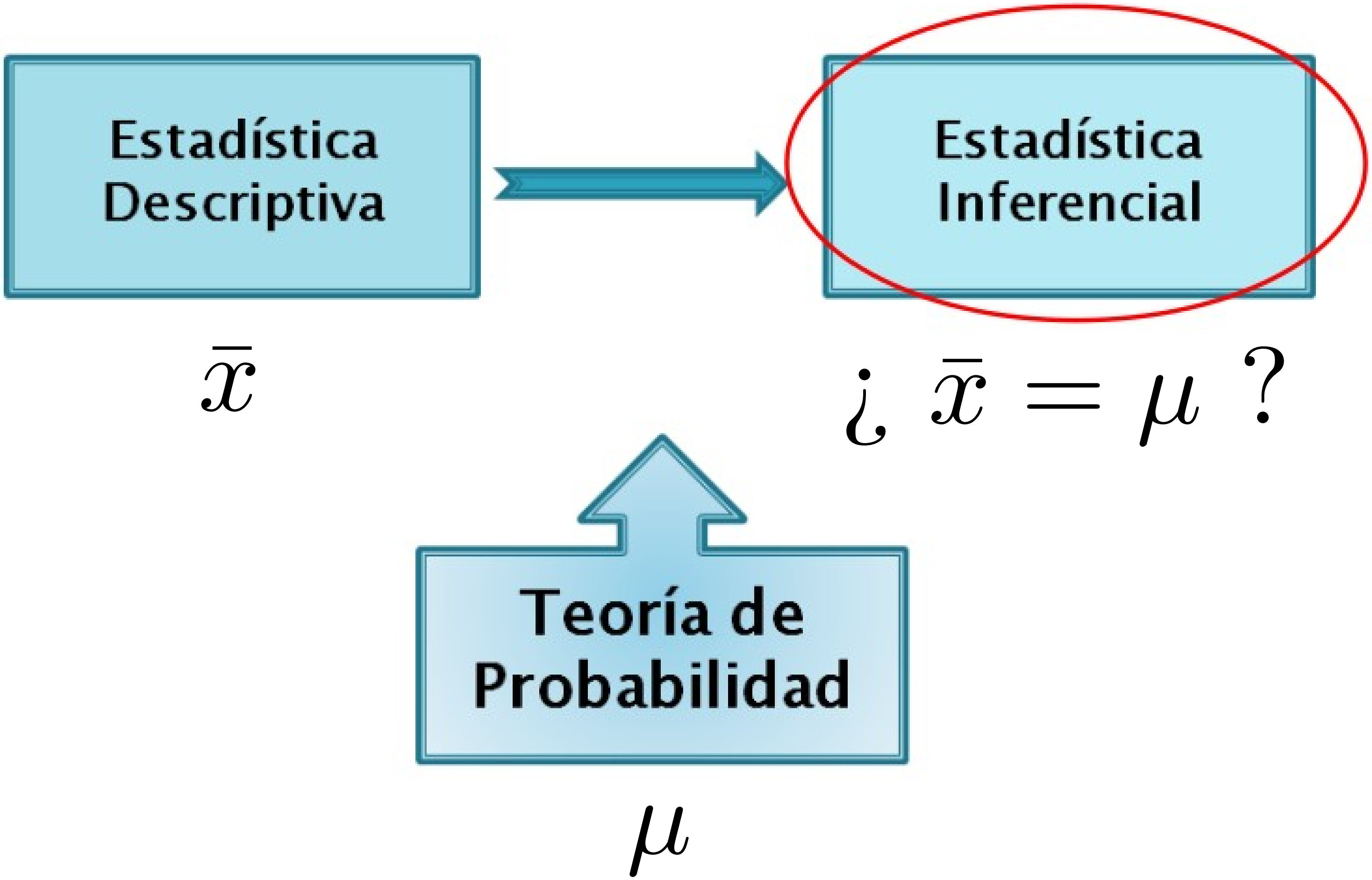
$\mu_{\bar{X}} = \mu$
$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$

Caso sin rep.

$\mu_{\bar{X}} = 24$
$\sigma^2_{\bar{X}} = 2/3$
n = 2
$\frac{N-n}{N-1} = \frac{3-2}{3-1} = \frac{1}{2}$

Se verifica que:

$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n} * \frac{N-n}{N-1}$


Estadística Descriptiva
Estadística Inferencial
$\bar{x}$
¿ $\bar{x} = \mu$?
Teoría de Probabilidad
μ

ESTIMACIÓN PUNTUAL

Consiste en utilizar el valor de una característica que depende de la muestra para estimar un el valor de un parámetro de la población.

Genéricamente:

$$\hat{\theta} = \theta$$

Por ejemplo:

$$\bar{x} = \mu$$

ESTIMACIÓN PUNTUAL

Genéricamente: $\hat{\theta} = \theta$

En nuestro cursado: ¡Aparecen nuevas VA!

$$\bar{x} = \mu$$

$$S_x^2 = \sigma^2$$

$$p = \pi$$

$$\bar{x} - \bar{y} = \mu_x - \mu_y$$

$$\frac{S_x^2}{S_y^2} = \frac{\sigma_x^2}{\sigma_y^2}$$

$$p_x - p_y = \pi_x - \pi_y$$

ESTIMADOR PUNTUAL

Es es una función de valores observados (muestra) que no depende de ningún parámetro desconocido.

Suponiendo que $\hat{\theta}$ es un estimador de θ, será considerado un “buen estimador” si cumple con las siguientes **propiedades**:

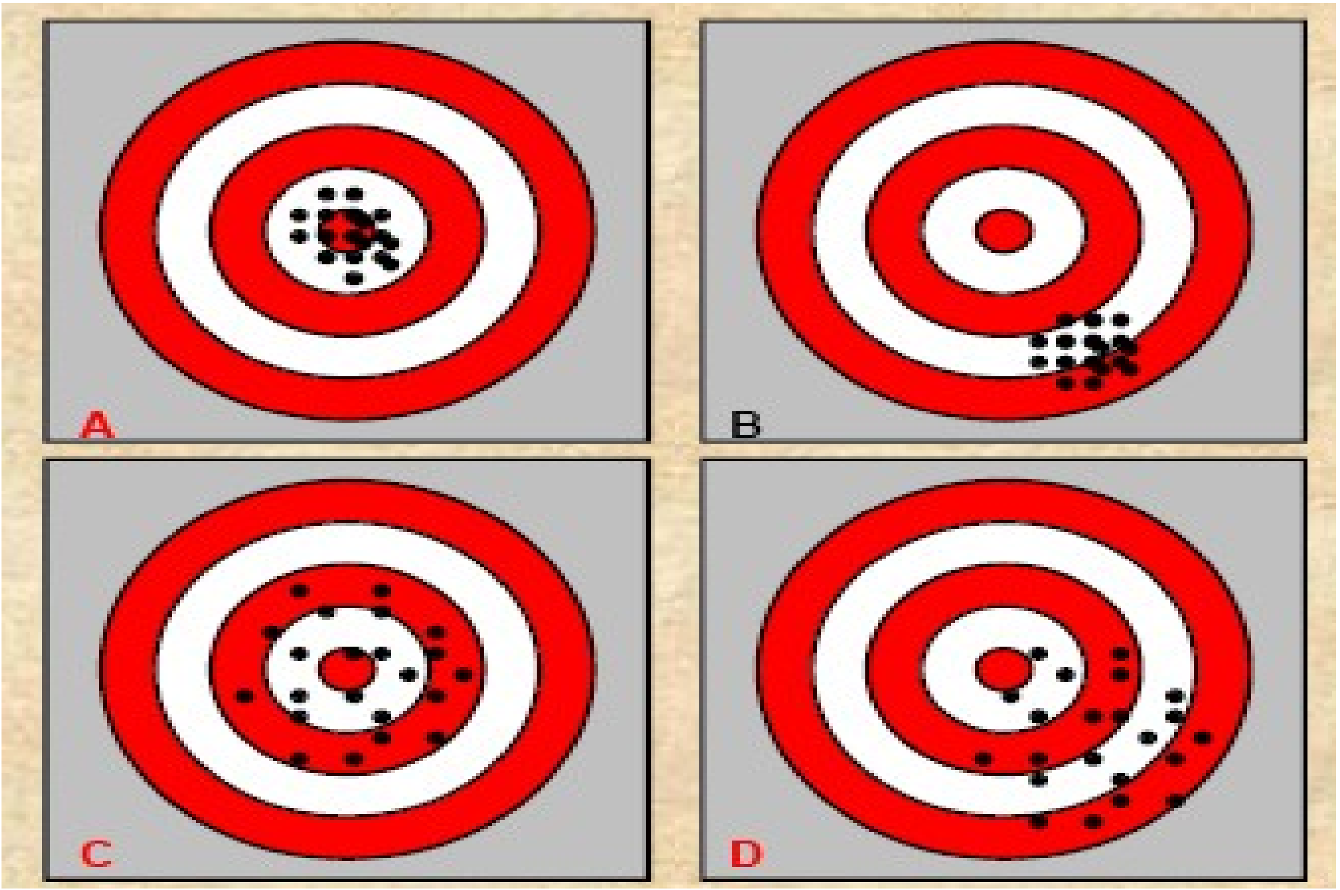
Insesgado: La esperanza del estimador coincide con el parámetro a estimar.

$$E(\hat{\theta}) = \theta$$

Eficiente: Su varianza es lo menor posible.

Consistente: Cuando el tamaño de la muestra crece arbitrariamente, se aproxima al valor del parámetro.

GRÁFICAMENTE



A: Estimador centrado y eficiente;
B: Estimador sesgado y eficiente
C: Estimador centrado e ineficiente;
D: Estimador sesgado e ineficiente

INTERVALOS DE CONFIANZA

El método de estimación por intervalos de confianza se basa en el valor de la estimación puntual pero considera además la distribución en el muestreo del estimador.

Dada una variable aleatoria con distribución normal, podremos estimar ***con una cierta confianza*** un parámetro desconocido de su planteando la siguiente probabilidad:

$$\mathrm{P}\ (\ \mathrm{a} \leq \theta \leq b) = confianza$$

consultas: taulamet@unl.edu.ar

INTERVALOS DE CONFIANZA

Si definimos la confianza como: $1 - \alpha$

Tendremos: $\mathrm{P}\ (\ \mathrm{a} \leq \theta \leq b) = 1 - \alpha$

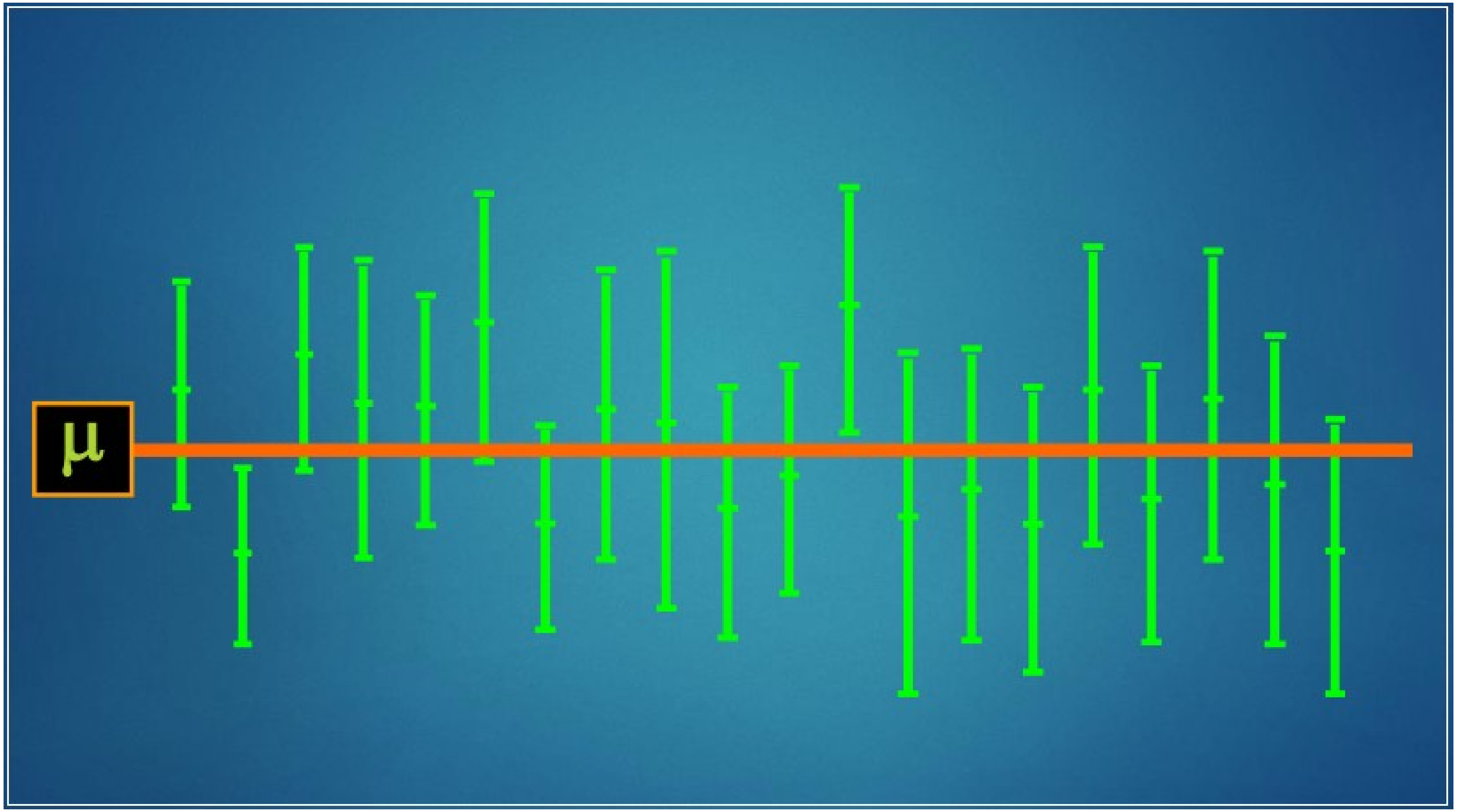
Lo que podría expresarse como:

$$P(|\theta - \hat{\theta}| \leq k\sigma_{\hat{\theta}}) = 1 - \alpha$$

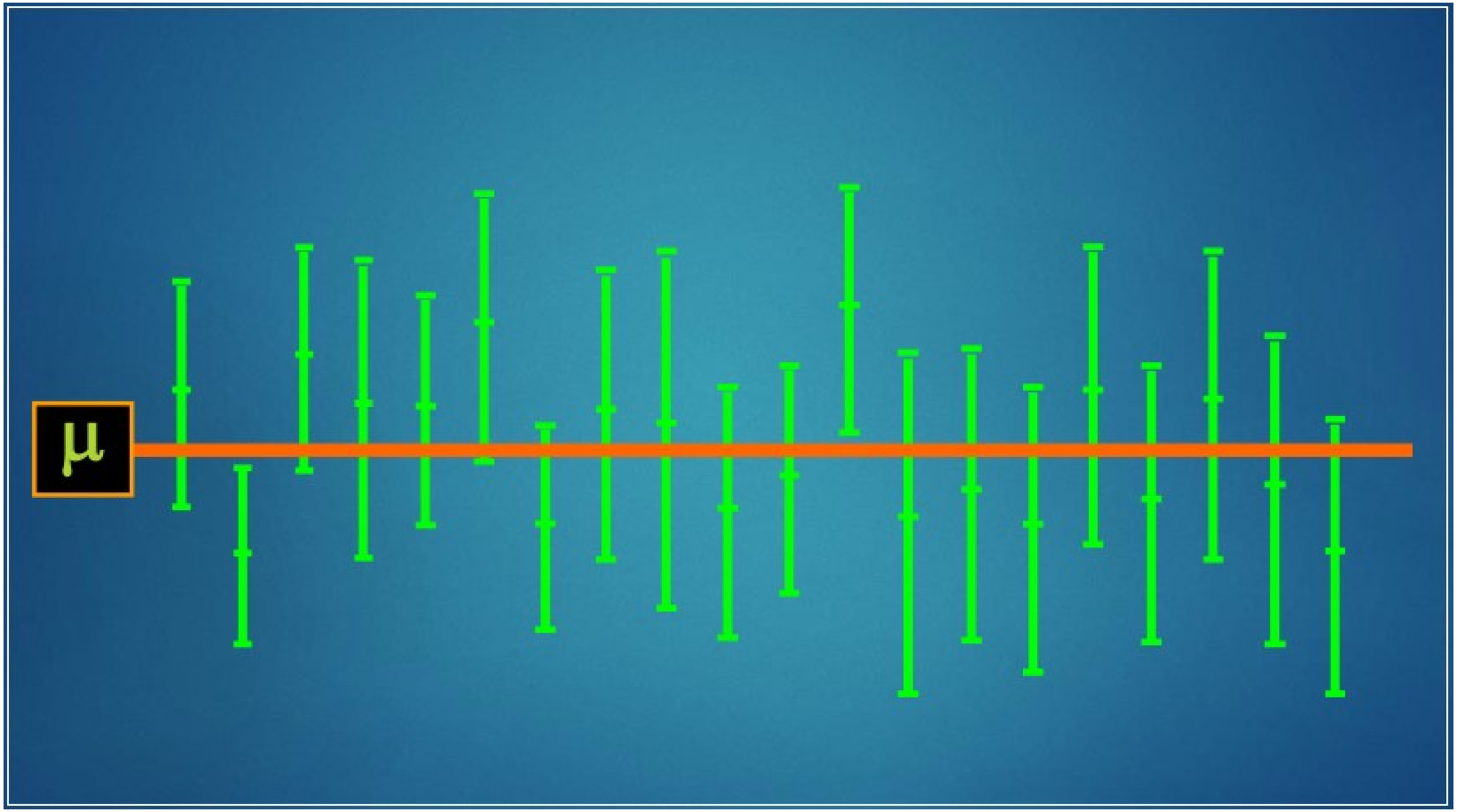
$$P(-k\sigma_{\hat{\theta}} \leq \theta - \hat{\theta} \leq k\sigma_{\hat{\theta}}) = 1 - \alpha$$

Siendo,
Error de estimación: $k\sigma_{\hat{\theta}}$
Error del estimador: $\sigma_{\hat{\theta}}$
Porcentil de la confianza: k
Amplitud del intervalo: $2k\sigma_{\hat{\theta}}$

LA NOCIÓN DE CONFIANZA

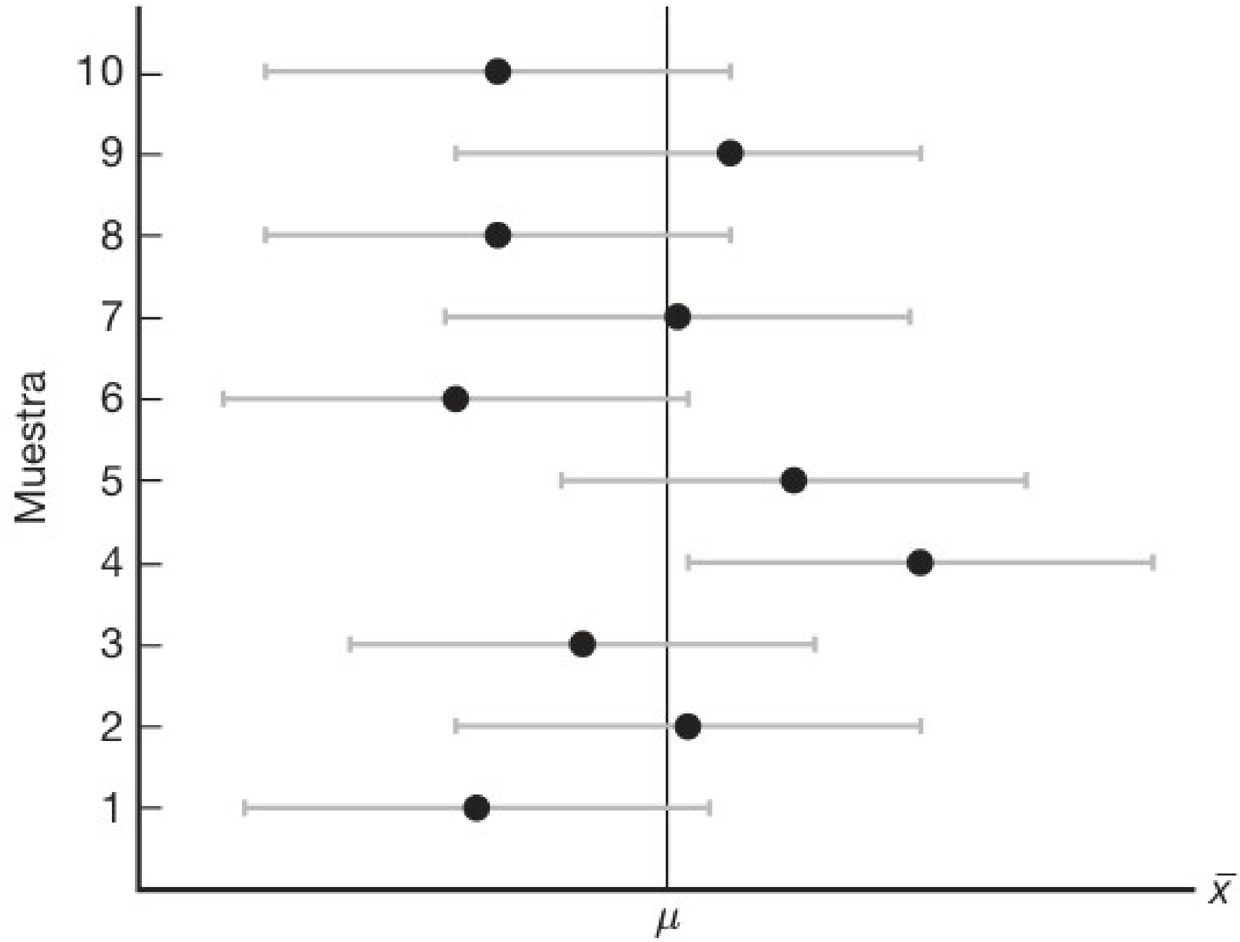


LA NOCIÓN DE CONFIANZA

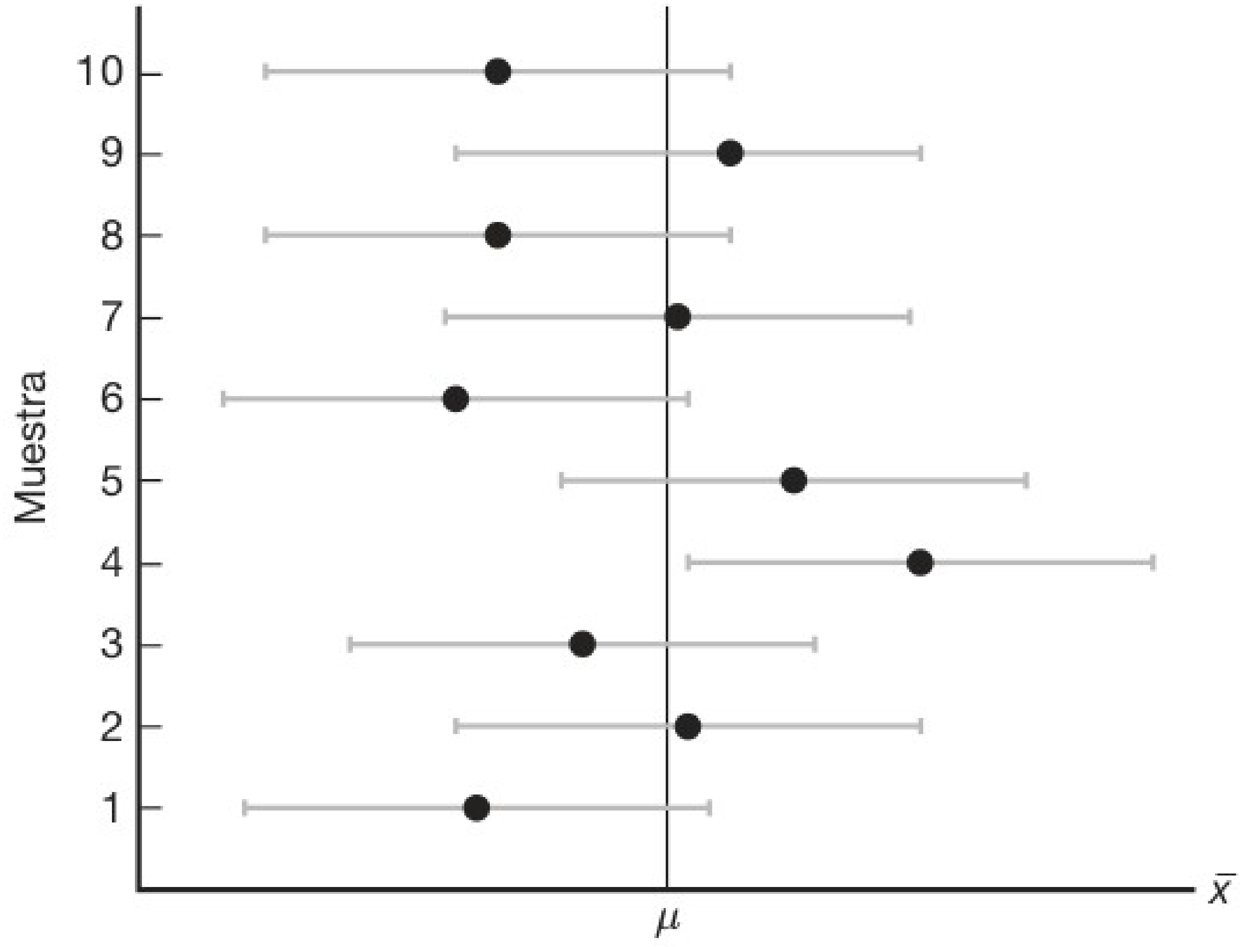


$$1 - \alpha = \frac{18}{20}$$

LA NOCIÓN DE CONFIANZA

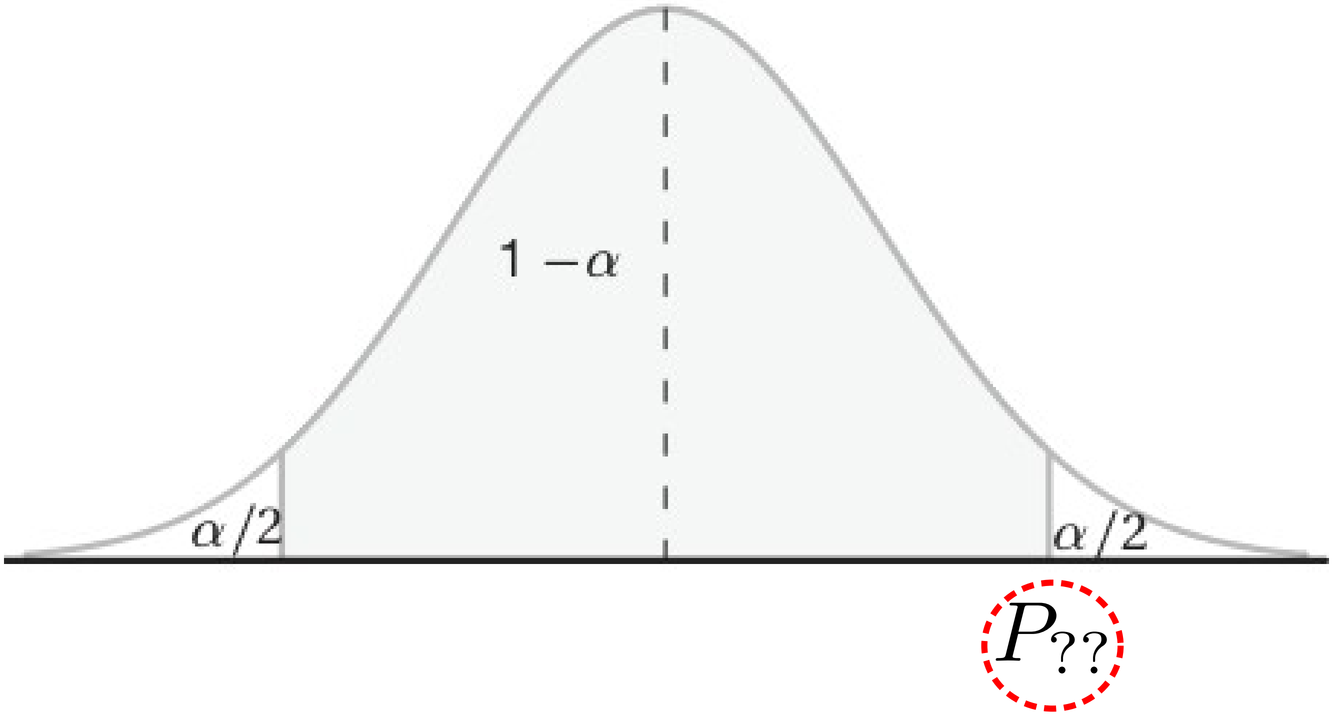


LA NOCIÓN DE CONFIANZA

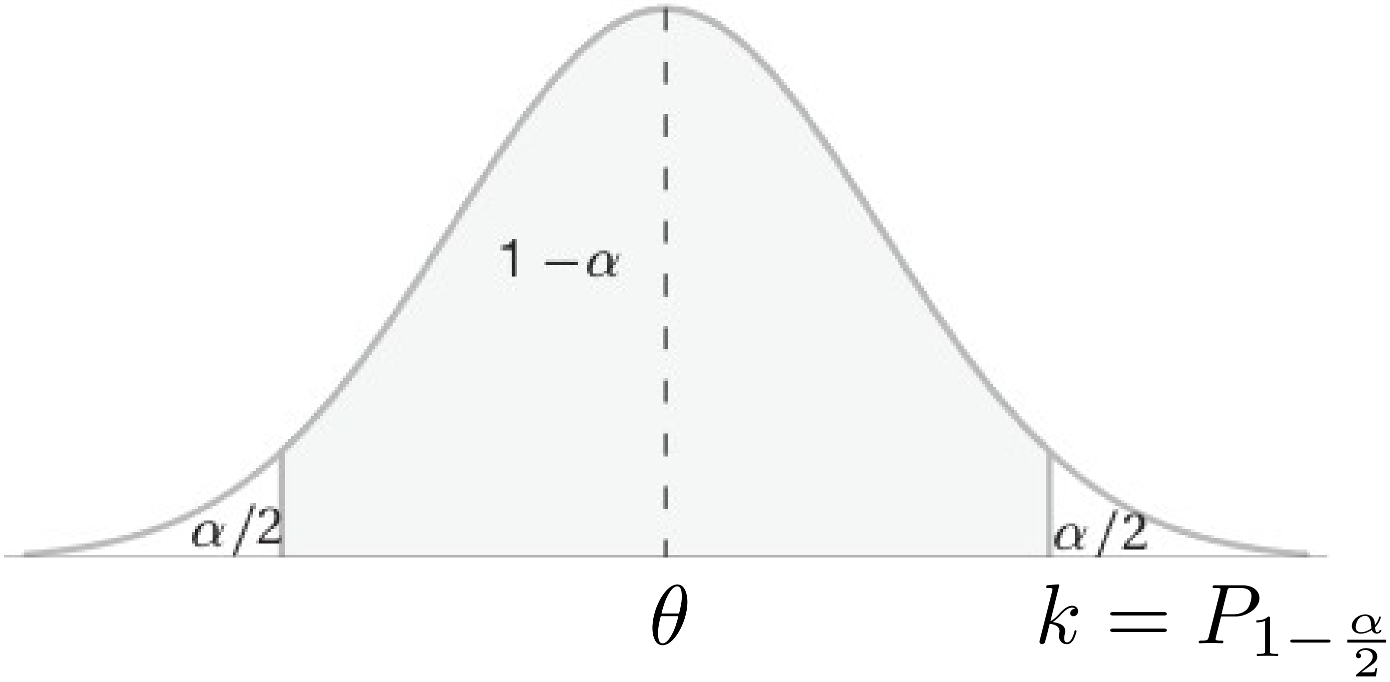


$$1-\alpha=\frac{9}{10}$$

LA NOCIÓN DE CONFIANZA



INTERVALOS DE CONFIANZA



Genéricamente:

$$(\hat{\theta} - k\sigma_{\hat{\theta}}; \hat{\theta} + k\sigma_{\hat{\theta}})$$

CASOS CONCRETOS

Parámetro	Supuesto	Estimador
μ	σ *conocido*	$\bar{x}$

Intervalo de confianza
$\left(\bar{x} \pm \left\| Z_{\left(1-\frac{\alpha}{2}\right)} \right\| \frac{\sigma}{\sqrt{n}} \right)$

CASOS CONCRETOS

Parámetro	Supuesto	Estimador
μ	σ *desconocido*, $n > 30$	$\bar{x}$

Intervalo de confianza
$\left(\bar{x} \pm \left\lvert Z_{\left(1-\frac{\alpha}{2}\right)} \right\rvert \dfrac{S}{\sqrt{n}}\right)$

CASOS CONCRETOS

Parámetro	Supuesto	Estimador
μ	σ *desconocido*, $n < 30$	$\bar{x}$

Intervalo de confianza
$\left(\bar{x} \pm \left| t_{\left(1-\frac{\alpha}{2};n-1\right)} \right| \frac{S'}{\sqrt{n}} \right)$

CASOS CONCRETOS

Parámetro	Supuesto	Estimador
σ^2	Población normal	S^2

Intervalo de confianza
$\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}$

CASOS CONCRETOS

Parámetro	Supuesto	Estimador
π	Población normal	p

Intervalo de confianza
$\left(p \pm \left\lvert Z_{\left(1-\frac{\alpha}{2}\right)} \right\rvert \sqrt{\frac{\pi(1-\pi)}{n}} \right)$

CASOS CONCRETOS

Parámetro	**Supuesto**	**Estimador**
$\mu_x - \mu_y$	σ_x, σ_y *conocidas*	$\overline{x} - \overline{y}$

Intervalo de confianza
$\left(\left(\overline{x} - \overline{y}\right) \pm \left\lvert Z_{\left(1-\frac{\alpha}{2}\right)} \right\rvert \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}} \right)$

consultas: taulamet@unl.edu.ar

CASOS CONCRETOS

Parámetro	Supuesto	Estimador
$\mu_x - \mu_y$	σ_x, σ_y *desconocidas* $n > 30$	$\bar{x} - \bar{y}$

Intervalo de confianza
$\left((\bar{x} - \bar{y}) \pm \lvert Z_{\left(1-\frac{\alpha}{2}\right)} \rvert \sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}} \right)$

CASOS CONCRETOS

Parámetro	Supuesto	Estimador
$\mu_x - \mu_y$	σ_x, σ_y *desconocidas* *pero iguales* $n < 30$	$\bar{x} - \bar{y}$

Intervalo de confianza
$\left((\bar{x} - \bar{y}) \pm \left\| t_{\left(1-\frac{\alpha}{2}\right)} \right\| \cdot S_w \cdot \sqrt{\frac{1}{n_x} + \frac{1}{n_y}} \right)$

$$Sw = \sqrt{\frac{n_x \cdot S_x^2 + n_y \cdot S_y^2}{n_x + n_y - 2}}$$

CASOS CONCRETOS

Parámetro	**Supuesto**	**Estimador**
$\mu_x - \mu_y$	σ_x, σ_y *desconocidas* *y distintas* $n < 30$	$\bar{x} - \bar{y}$

Intervalo de confianza
$\left((\bar{x} - \bar{y}) \pm \lvert t_{\left(1-\frac{\alpha}{2}\right), \upsilon} \rvert \sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}} \right)$

CASOS CONCRETOS

Parámetro	**Supuesto**	**Estimador**
$\pi_1 - \pi_2$	*Poblaciones normales*	$p_1 - p_2$

Intervalo de confianza
$\left(\Delta p \;\pm\; \left\| Z_{\left(1-\frac{\alpha}{2}\right)} \right\| \;(I) \right)$

$$(I) = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$$

consultas: taulamet@unl.edu.ar

CASOS CONCRETOS

Parámetro	Supuesto	Estimador
$\dfrac{\sigma_x^2}{\sigma_y^2}$	*Poblaciones normales*	$\dfrac{S_x^2}{S_y^2}$

Intervalo de confianza
$\left(\dfrac{S_x^2}{S_y^2}(I) \le \dfrac{\sigma_x^2}{\sigma_y^2} \le \dfrac{S_x^2}{S_y^2}(II)\right)$

$$(I) = \frac{1}{F_{\left(\frac{\alpha}{2};n_x-1;n_y-1\right)}}$$

$$(II) = F_{\left(\frac{\alpha}{2};n_x-1;n_y-1\right)}$$

Durante los últimos 10 días de junio un tren llegó tarde a su destino en los siguientes tiempos (en minutos; un número negativo significa que el tren llegó temprano ese número de minutos):

-3	6	4	10	-4	124	2	-1	4	1

Obtener un intervalo de confianza para estimar la esperanza poblacional de los tiempos de de llegada tarde a destino del tren en minutos, con un 90% de confianza.

$$\left(\bar{x} \pm \left| t_{\left(1-\frac{\alpha}{2};n-1\right)} \right| \frac{S'}{\sqrt{n}} \right)$$

consultas: taulamet@unl.edu.ar

$$\left(\bar{x} \pm \left|t_{\left(1-\frac{\alpha}{2};n-1\right)}\right| \frac{S'}{\sqrt{n}}\right)$$

Media	14,3
Error estándar	12,2611

CONF	90,00%
T95	1,83
Error	22,48
LI	-8,2
LS	34,7

Respuesta: Con un 90% de confianza, el promedio poblacional de los tiempos de demora en llegar a destino del tren se econtrará entre -8,2 y 34,7, en símbolos:

$$\text{IC p/ } \mu, 1-\alpha = 90\% : (-8,2; 34,7)$$

(1994-2024)
30 años de la Consagración Constitucional de la Autonomía y Autarquía Universitaria en Argentina.

UNIVERSIDAD NACIONAL DEL LITORAL · LUX INDEFICIENS · 1919

FICH

ESTADÍSTICA
PROBABILIDAD Y ESTADÍSTICA

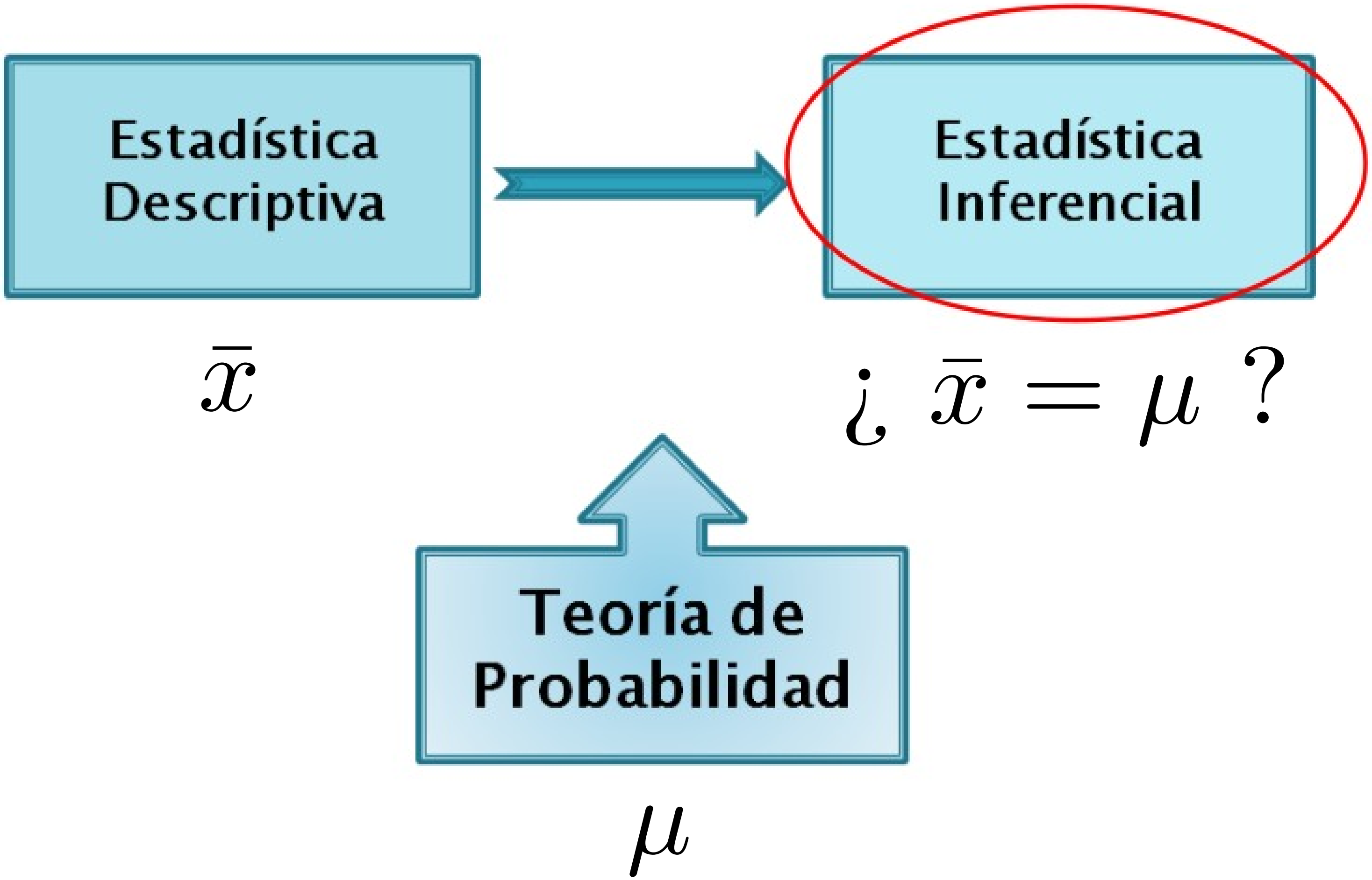
Clase Pública 23/10 - Unidad 6 - *Estadística Inferencial*
Introducción a las Pruebas de hipótesis

Ingenierías en: Recursos Hídricos, Ambiental y Agrimensura

Año 2024

Prof. Juan Pablo Taulamet

consultas: taulamet@unl.edu.ar



consultas: taulamet@unl.edu.ar

ESTIMACIÓN PUNTUAL

Consiste en utilizar el valor de una característica que depende de la muestra para estimar un el valor de un parámetro de la población.

Genéricamente:

$$\hat{\theta} = \theta$$

Por ejemplo:

$$\bar{x} = \mu$$

ESTIMACIÓN PUNTUAL

Genéricamente: $\hat{\theta} = \theta$

En nuestro cursado: ¡Aparecen nuevas VA!

$$\bar{x} = \mu$$

$$S_x^2 = \sigma^2$$

$$p = \pi$$

$$\bar{x} - \bar{y} = \mu_x - \mu_y$$

$$\frac{S_x^2}{S_y^2} = \frac{\sigma_x^2}{\sigma_y^2}$$

$$p_x - p_y = \pi_x - \pi_y$$

consultas: *taulamet@unl.edu.ar*

ESTIMADOR PUNTUAL

Es es una función de valores observados (muestra) que no depende de ningún parámetro desconocido.

Suponiendo que $\hat{\theta}$ es un estimador de θ, será considerado un "buen estimador" si cumple con las siguientes **propiedades**:

Insesgado: La esperanza del estimador coincide con el parámetro a estimar.

$$E(\hat{\theta}) = \theta$$

Eficiente: Su varianza es lo menor posible.

Consistente: Cuando el tamaño de la muestra crece arbitrariamente, se aproxima al valor del parámetro.

PRUEBAS DE HIPÓTESIS

- *Paramétricas*
  - *Afirmaciones acerca de un parámetro*
    - *Unilaterales*
    - *Bilaterales*
- *No Paramétricas*
  - *Afirmaciones acerca de la distribución de los datos*

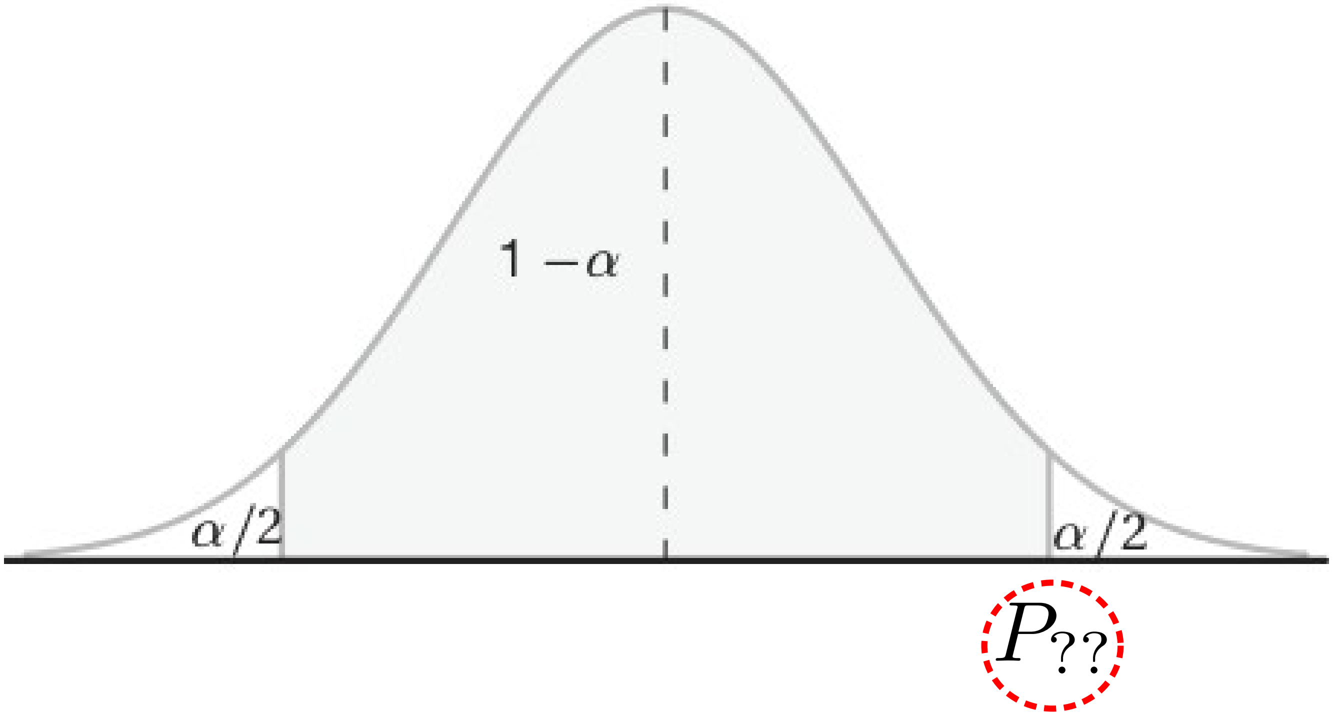
PRUEBAS DE HIPÓTESIS

- *Paramétricas*
  - *Afirmaciones acerca de un parámetro*
    - *Hipótesis antagónicas (M.E.)*
      - *Hipótesis Nula*
      - *Hipótesis Alternativa*
- *Ejemplo 1:*

$$H_0 : \mu = 42$$

$$H_1 : \mu > 42$$

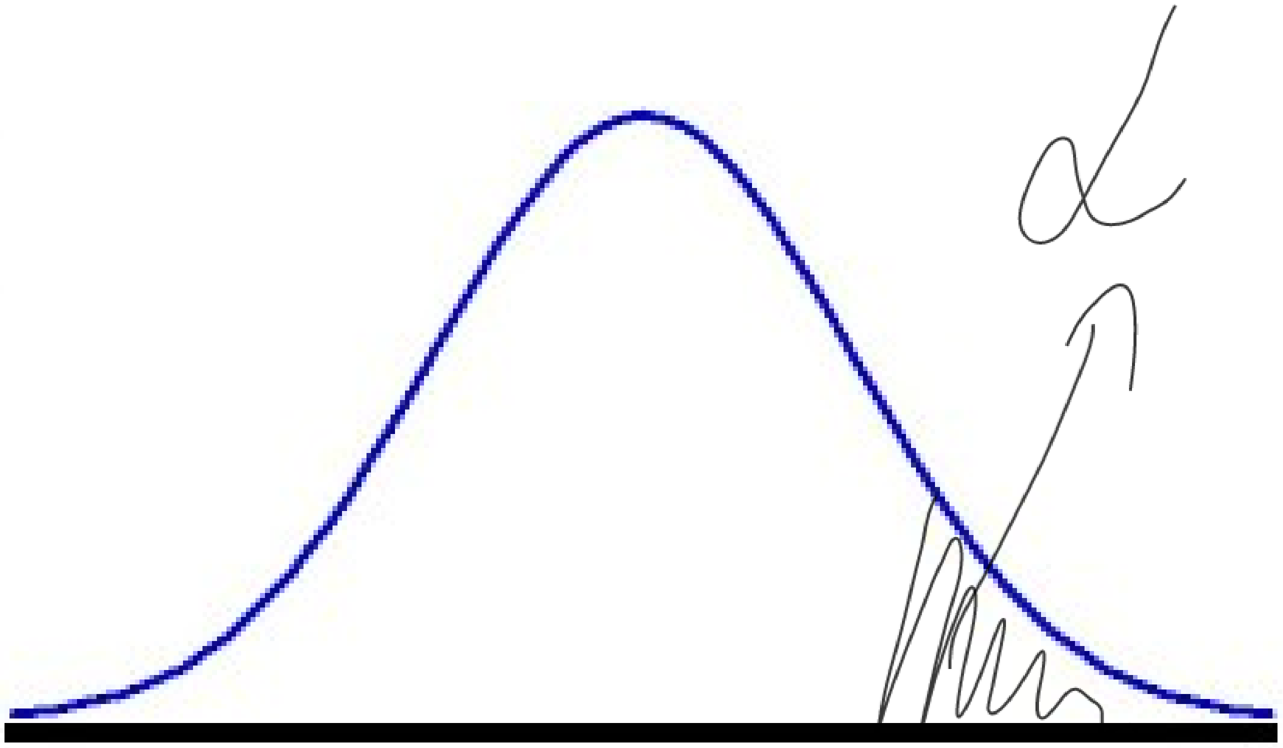
LA NOCIÓN DE CONFIANZA



ESQUEMA DE UNA PRUEBA

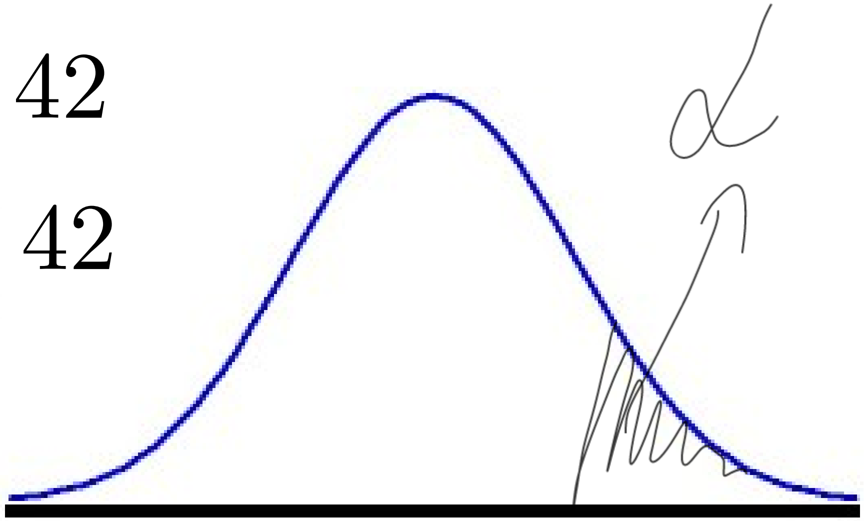
$$H_0 : \mu = 42$$

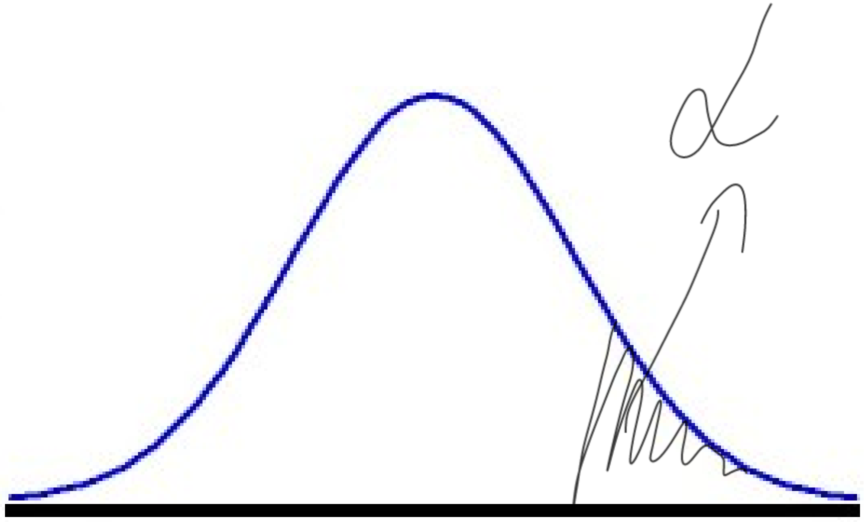
$$H_1 : \mu > 42$$



$$H_0 : \mu = 42$$

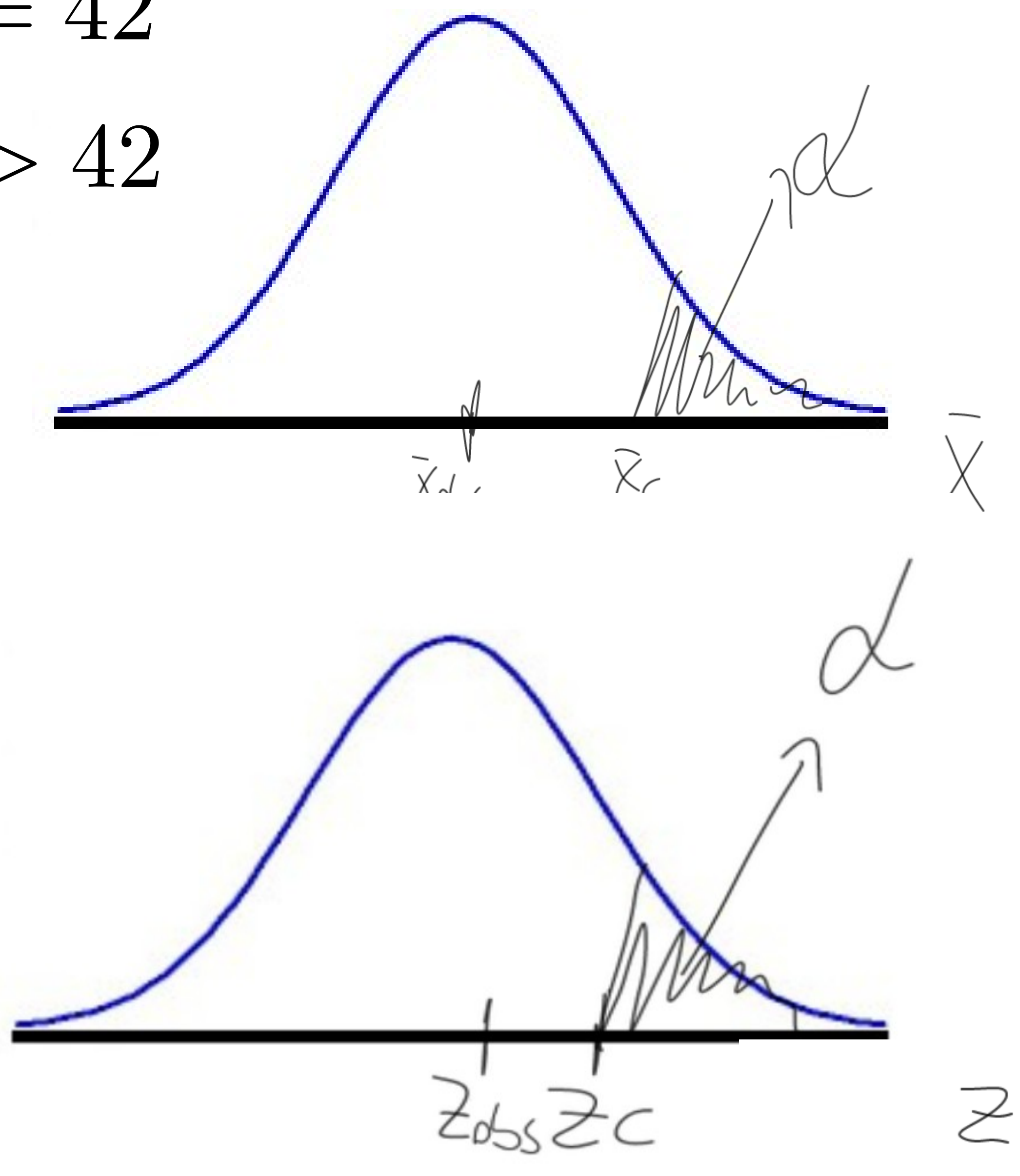
$$H_1 : \mu > 42$$





$$H_0 : \mu = 42$$

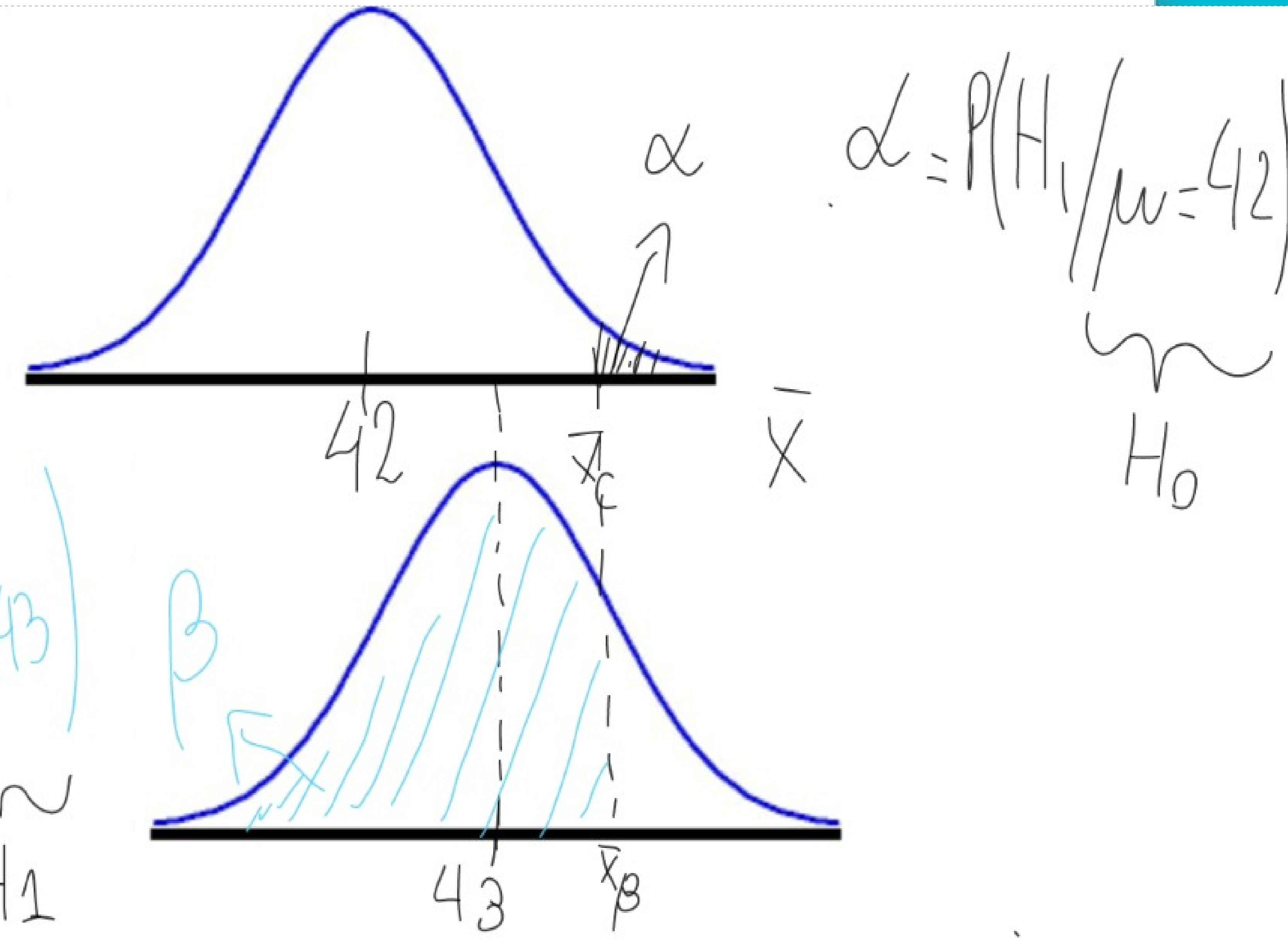
$$H_1 : \mu > 42$$



REPASO Y POTENCIA

$H_0 : \mu = 42$

$H_1 : \mu > 42$



(1994–2024)
30 años de la Consagración Constitucional de la Autonomía y Autarquía Universitaria en Argentina.

UNIVERSIDAD NACIONAL DEL LITORAL
LUX INDEFICIENS · 1919

FICH

ESTADÍSTICA
PROBABILIDAD Y ESTADÍSTICA

Semana#11 - Unidad 7 - *Regresión y Correlación*

Ingenierías en: Recursos Hídricos, Ambiental y Agrimensura

Año 2024

Prof. Juan Pablo Taulamet

consultas: taulamet@unl.edu.ar

ESTUDIO DE ASOCIACIÓN ENTRE VARIABLES

REGRESIÓN

ENCONTRAR UN MODELO

CORRELACIÓN

EXACTITUD DEL MODELO

ANÁLISIS DE REGRESIÓN

Permite encontrar un modelo que vincula a dos o más variables, brindando un mecanismo de pronóstico.

ANÁLISIS DE CORRELACIÓN

Determina la medida del grado de exactitud de la relación entre las variables

consultas: taulamet@unl.edu.ar

La siguiente tabla muestra valores de precipitación y escurrimiento en una cuenca urbana.

x (mm)	y (mm)
50	60
60	60
70	80
80	70
90	80
100	100
110	100
120	130
130	110
140	120
150	150
160	140

La siguiente tabla muestra valores de precipitación y escurrimiento en una cuenca urbana.

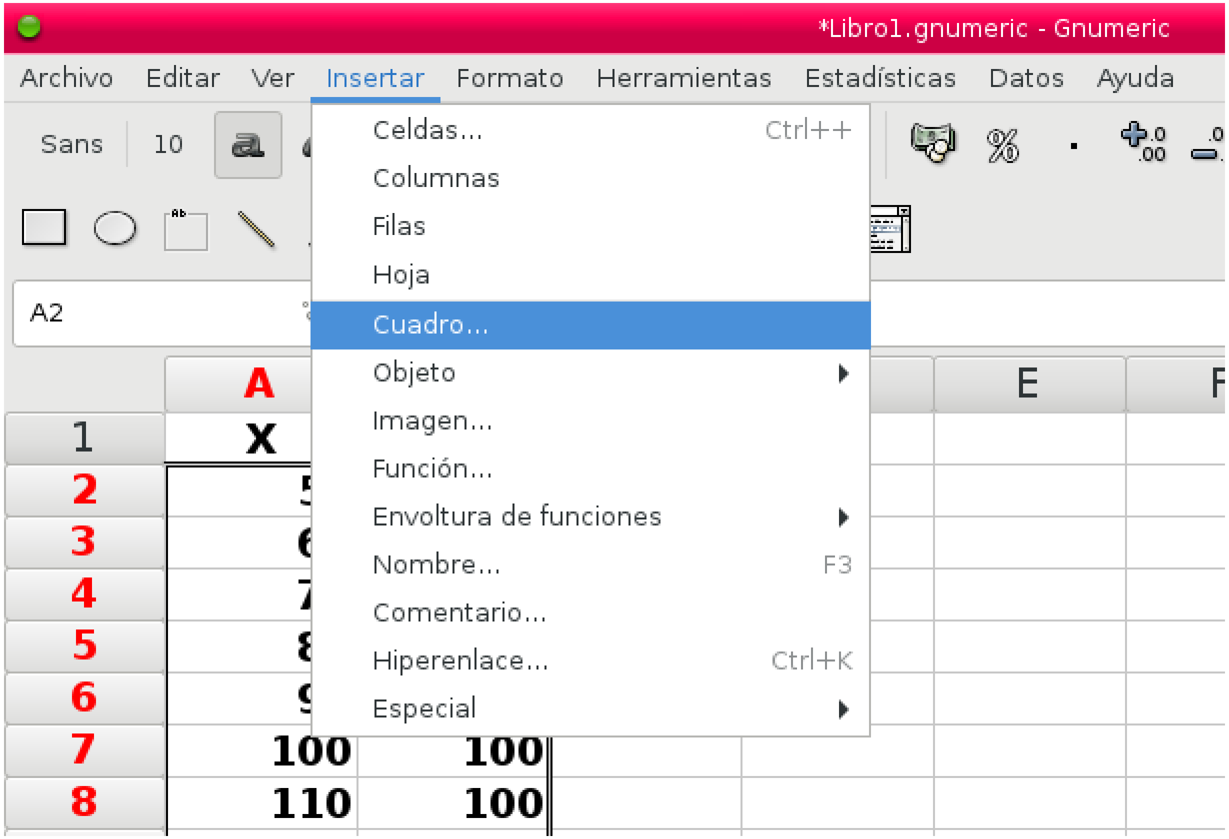
x (mm)	y (mm)
50	60
60	60
70	80
80	70
90	80
100	100
110	100
120	130
130	110
140	120
150	150
160	140

¿Existe relación entre X e Y?

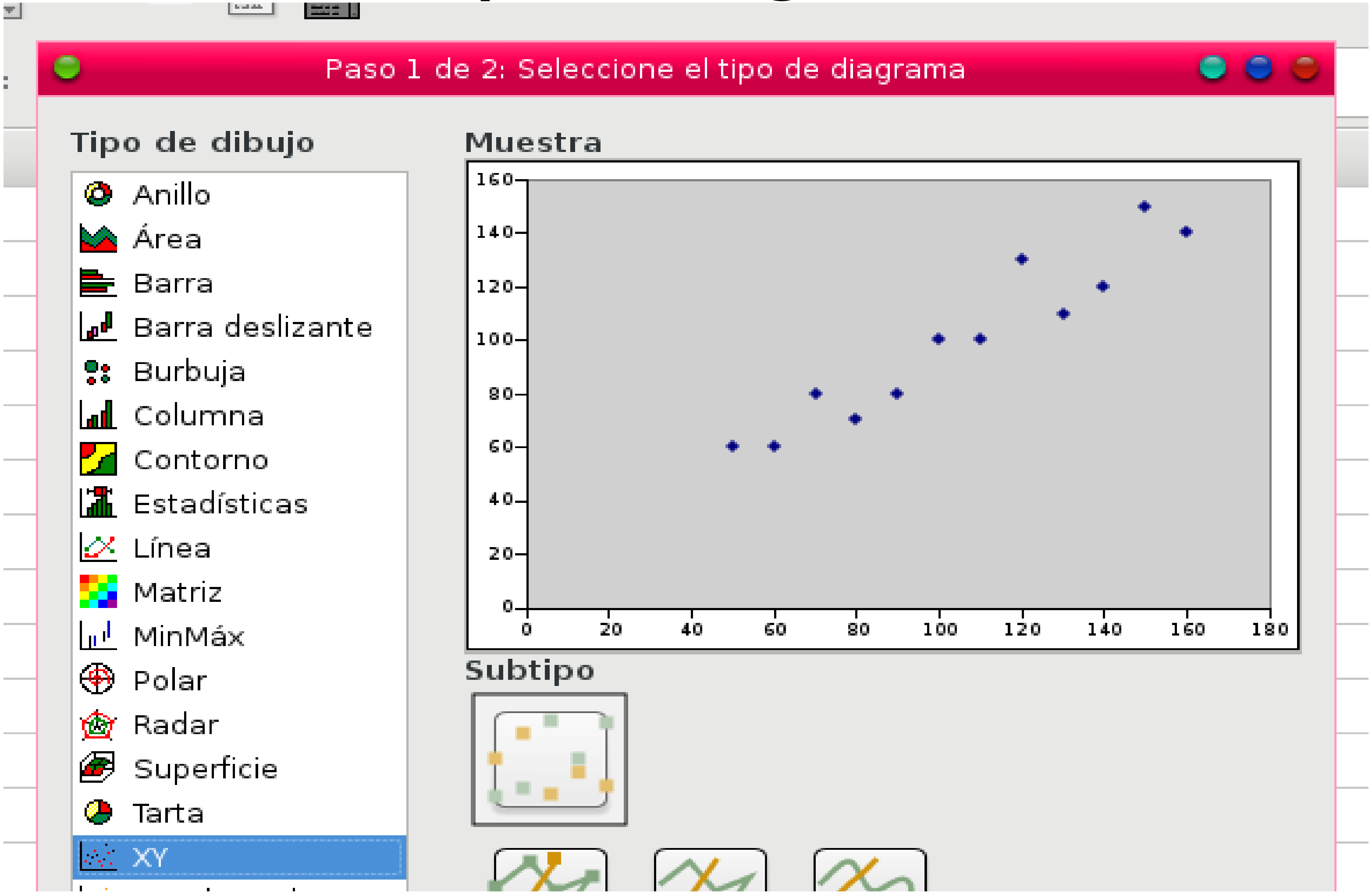
En Gnumeric

	A	B	C	D
1	**X**	**Y**		
2	**50**	**60**		
3	**60**	**60**		
4	**70**	**80**		
5	**80**	**70**		
6	**90**	**80**		
7	**100**	**100**		
8	**110**	**100**		
9	**120**	**130**		
10	**130**	**110**		
11	**140**	**120**		
12	**150**	**150**		
13	**160**	**140**		
14				

Insertamos un Dispersiograma

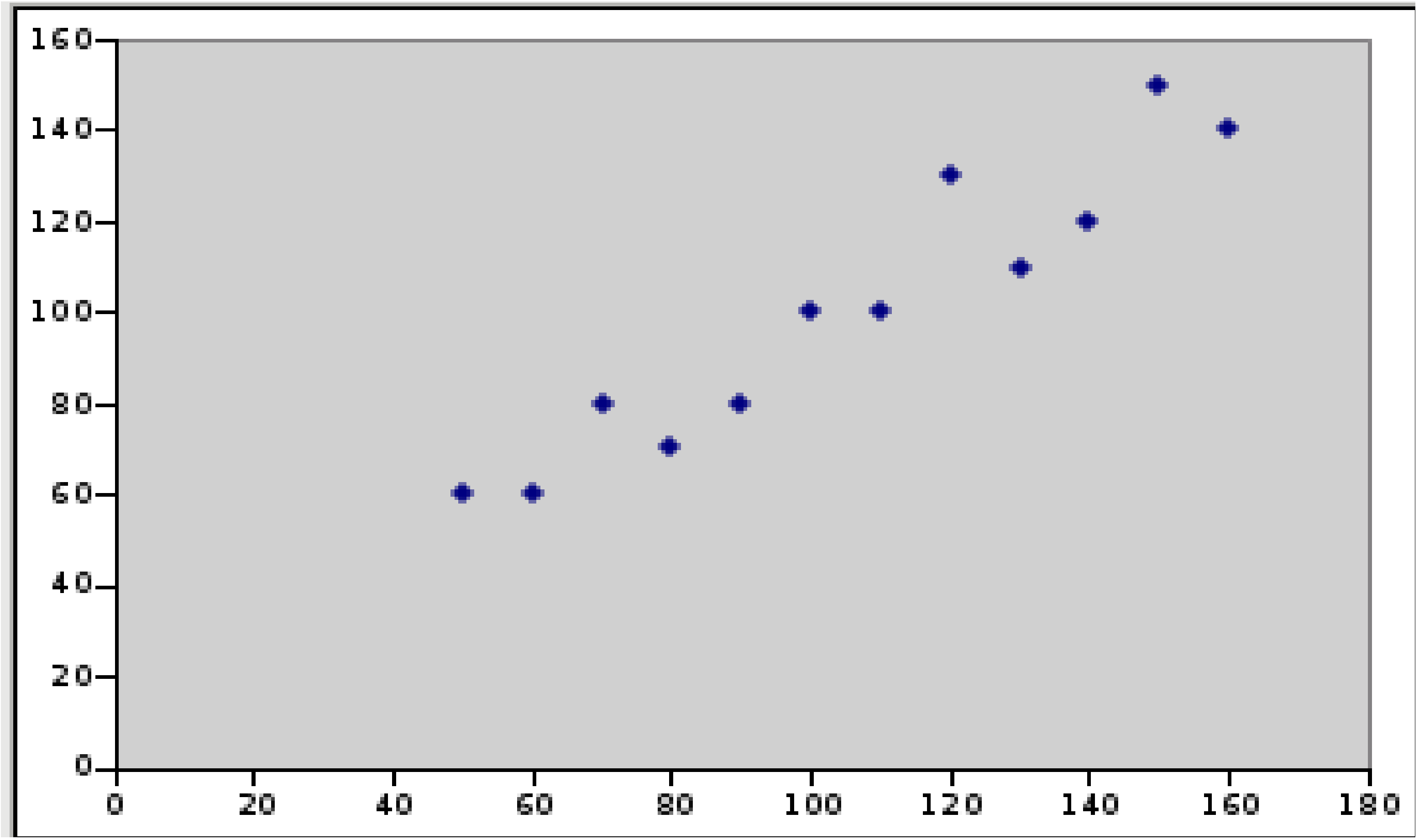


Insertamos un Dispersiograma



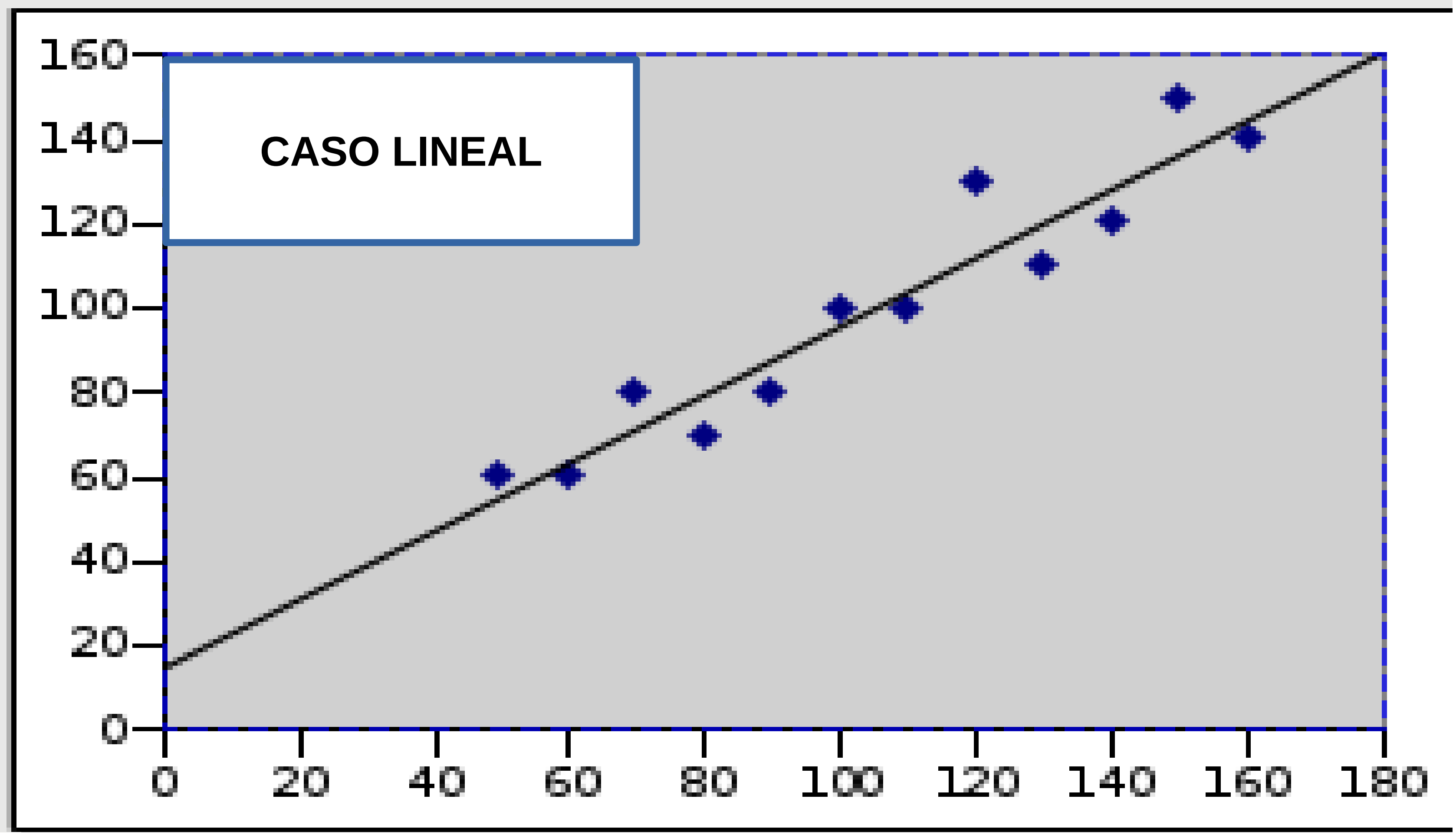
Dispersiograma

¿Existe relación entre X e Y? ¿Cómo es?



MODELO DE REGRESIÓN BIVARIADO

$$E(Y/x) = \mu_{Y/x} = \alpha + \beta X_i$$



MODELO DE REGRESIÓN LINEAL SIMPLE

PARÁMETROS DEL MODELO

VARIABLE EXPLICATIVA INDEPENDIENTE

ERROR ALEATORIO

VARIABLE DEPENDIENTE

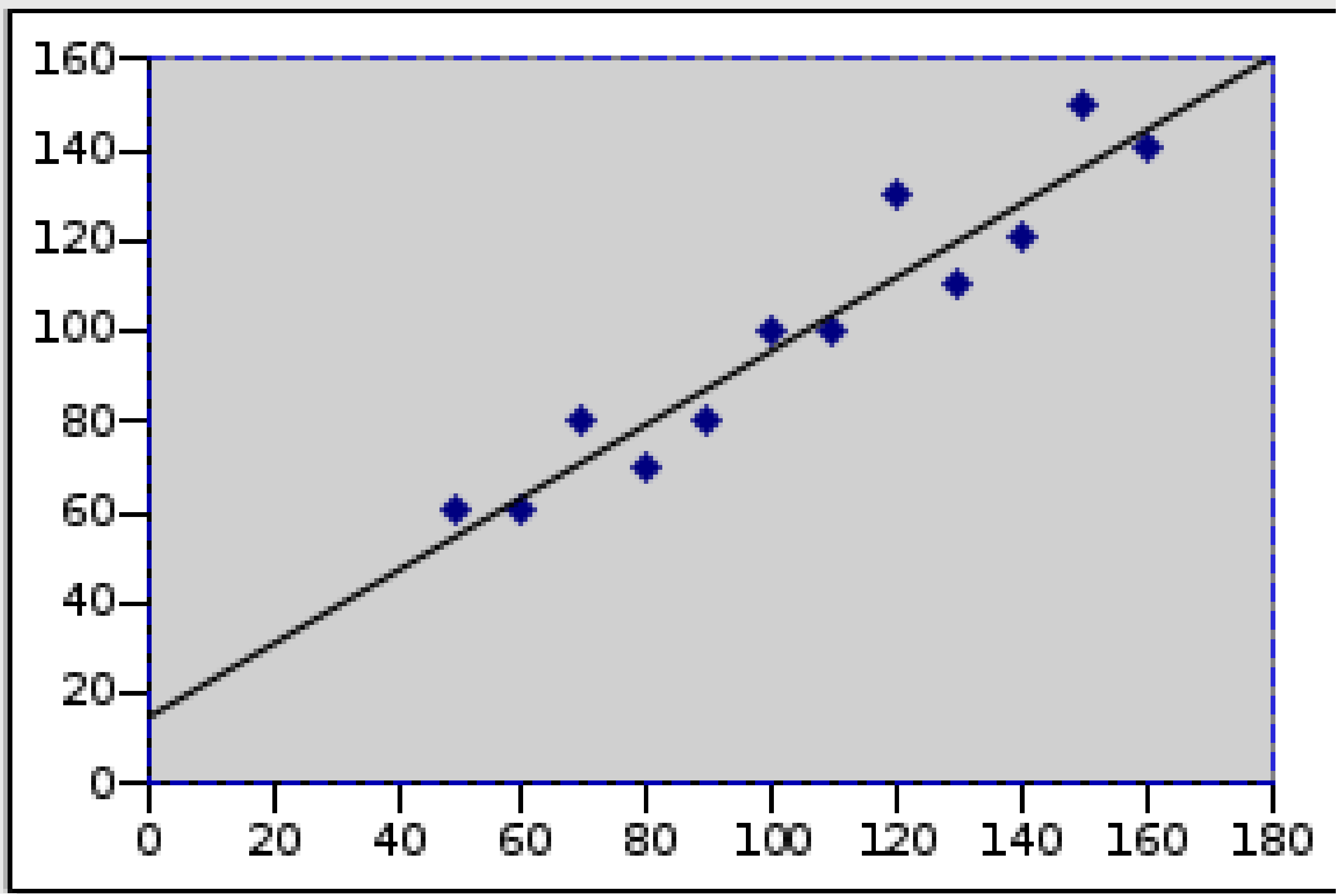
$$Y_i = \underset{I}{\underline{\alpha + \beta X_i}} + \underset{II}{\underline{\in_i}}$$

PARTE SISTEMÁTICA

PARTE ESTOCÁSTICA

consultas: taulamet@unl.edu.ar

PENSANDO ESTA RELACIÓN ESTADÍSTICA



consultas: taulamet@unl.edu.ar

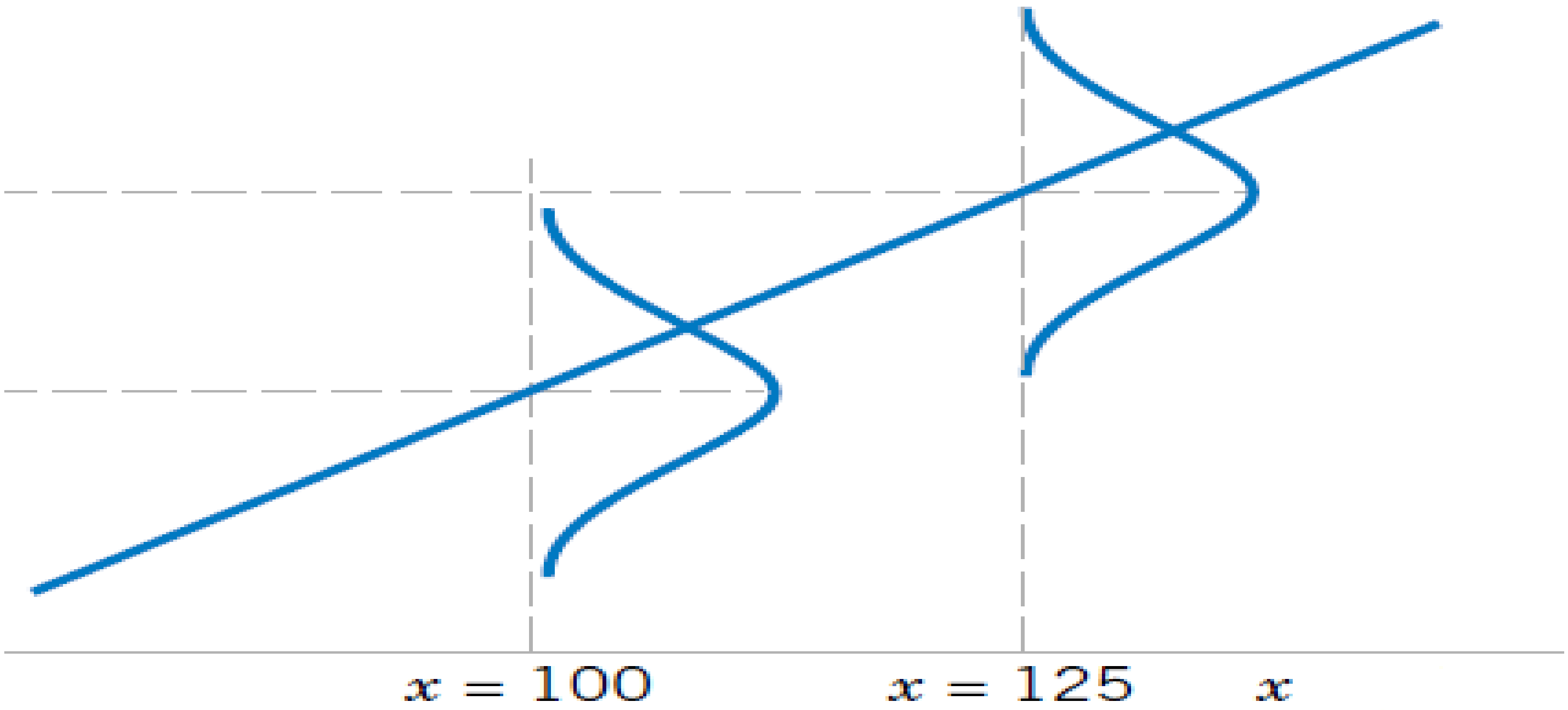
PENSANDO ESTA
RELACIÓN ESTADÍSTICA

La variable dependiente tiende a variar con la variable independiente de una manera sistemática.

Existe dispersión de los puntos entorno a la curva que relaciona las variables.

PENSANDO ESTA RELACIÓN ESTADÍSTICA

Existe una distribución de probabilidades de la variable dependiente (Y) para cada valor de X. Existen grupos de valores de Y para cada valor de X que llamaremos sub-poblaciones. El modelo de regresión es el que *relaciona los valores medios* de dichas sub-poblaciones:



consultas: taulamet@unl.edu.ar

MODELOS DE REGRESIÓN

LINEAL
NO-LINEAL
(Polinomial, exponencial, etc.)

SIMPLE
(2 Variables)
MÚLTIPLE
(Más de 2 Variables)

MODELO DE REGRESIÓN LINEAL SIMPLE

$$Y_i = \underbrace{\alpha + \beta X_i}_{I} + \underbrace{\in_i}_{II}$$

Para conocer más del modelo supondremos que la X se fija en un valor y hallaremos la esperanza y varianza del mismo.

MODELO DE REGRESIÓN LINEAL SIMPLE

Supongamos que la media y la varianza de $\in$ son 0 y σ^2

$$E(Y/x) = E(\alpha + \beta x + \varepsilon) = \alpha + \beta x + E(\varepsilon) = \alpha + \beta x$$

$$Var(Y/x) = Var(\alpha + \beta x + \varepsilon) = Var(\alpha + \beta x) + Var(\varepsilon) = 0 + \sigma^2 = \sigma^2$$

ESTIMACIÓN POR MÍNIMOS CUADRADOS

SUPUESTOS:

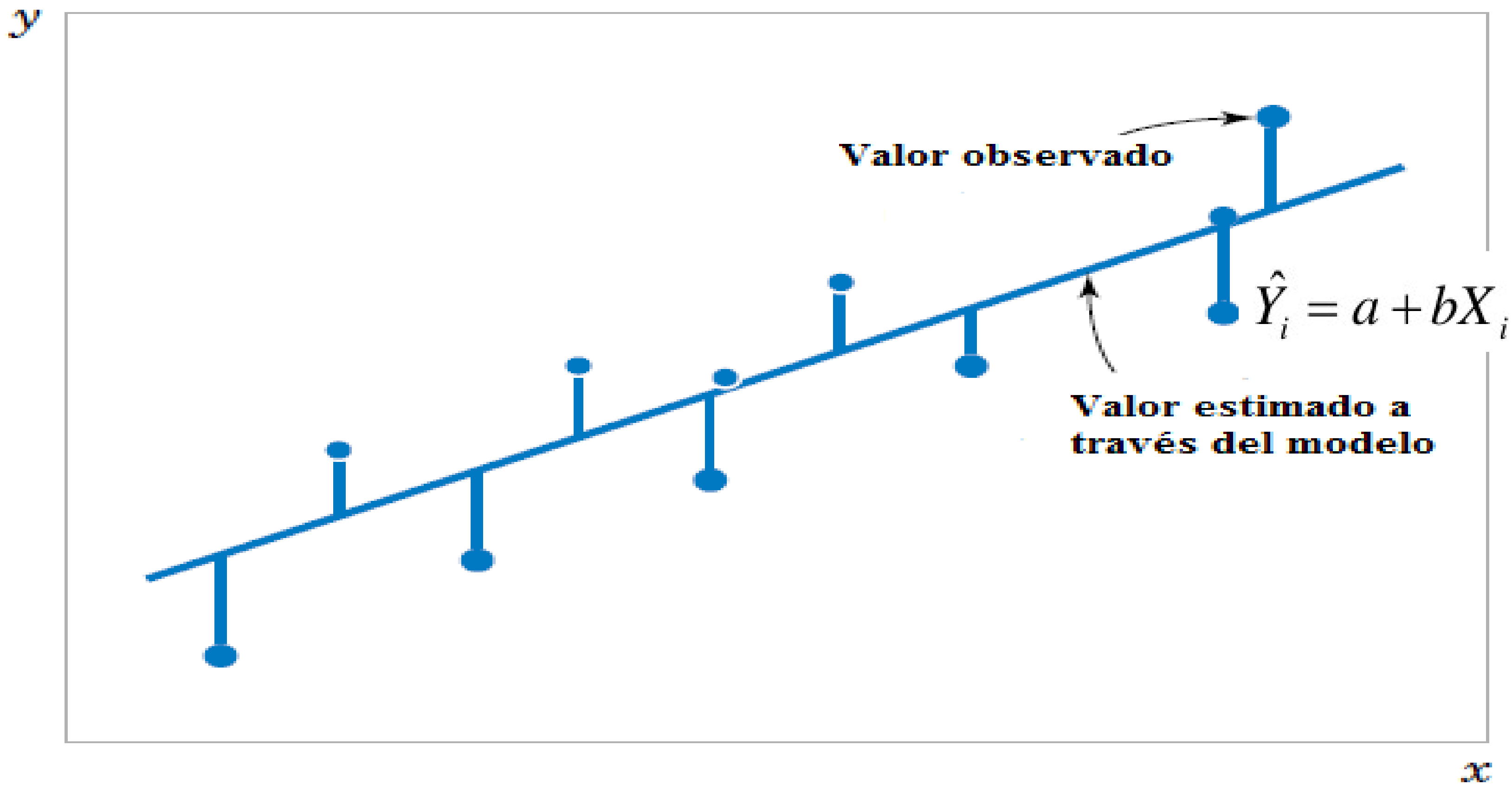
**La VA del error es independiente.*

**La distribución del error es normal*

**X es un conjunto de números fijos*

consultas: taulamet@unl.edu.ar

ESTIMACIÓN POR MÍNIMOS CUADRADOS



$$S = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \qquad \frac{\partial S}{\partial \alpha} = 0 \quad \text{y} \quad \frac{\partial S}{\partial \beta} = 0$$

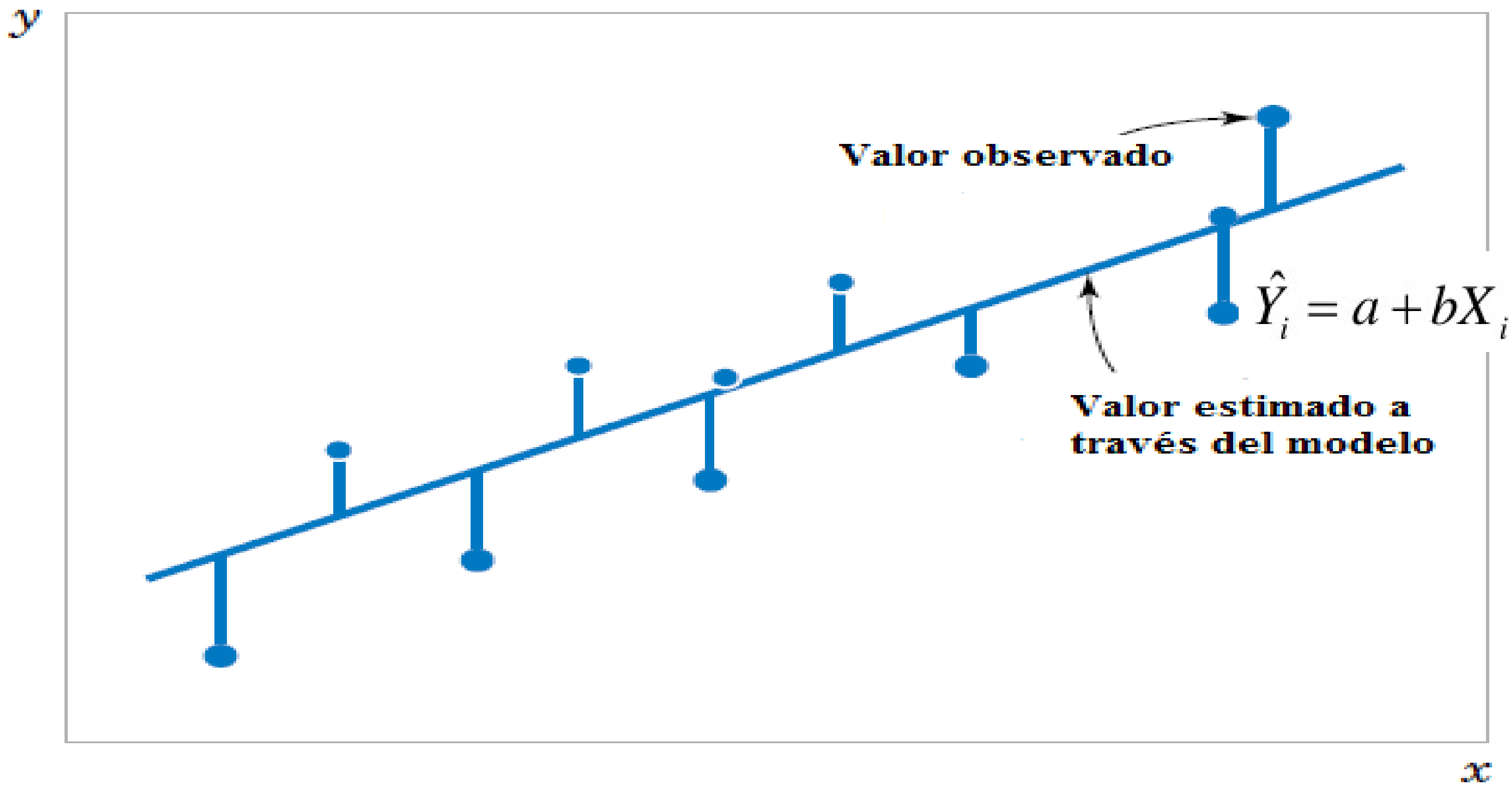
ESTIMACIÓN POR MÍNIMOS CUADRADOS

$$\hat{\beta} = b = \frac{m_{1,1}}{S_x^2} = \frac{\text{cov}}{S_x^2}$$

$$\hat{\alpha} = a = \overline{Y} - \hat{\beta}\overline{X}$$

Este método minimiza las distancias verticales

VARIANZA DE LA REGRESIÓN



$$S^2_{y/x} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-2}$$

ANÁLISIS DE CORRELACIÓN

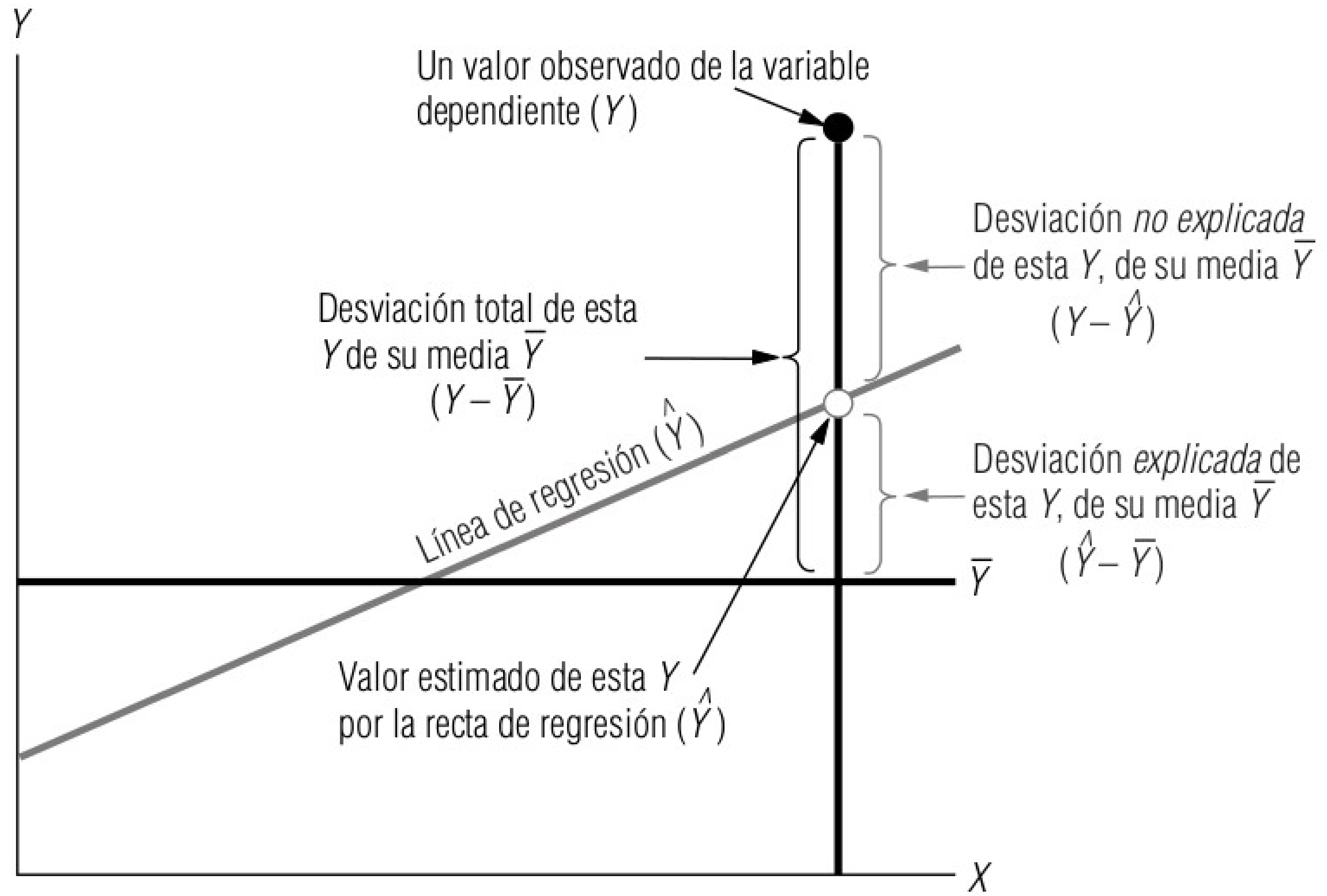
CASO LINEAL
Se utiliza el coeficiente de correlación r:

$$r = \hat{\rho}\,\frac{\sum_i (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_i (x_i-\bar{x})^2}\,\sqrt{\sum_i (y_i-\bar{y})^2}} = \frac{m_{1,1}}{S_x S_y}$$

$$-1 \leq r \leq 1$$

Cuanto más cerca se encuentre del 0 menor relación lineal habrá entre la X y la Y. En la medida que se acerque a 1 o a -1, expresará un mayor grado de asociación lineal ya sea relación directa o inversamente proporcional.

PARTICIÓN DE LOS CUADRADOS



ANÁLISIS DE CORRELACIÓN

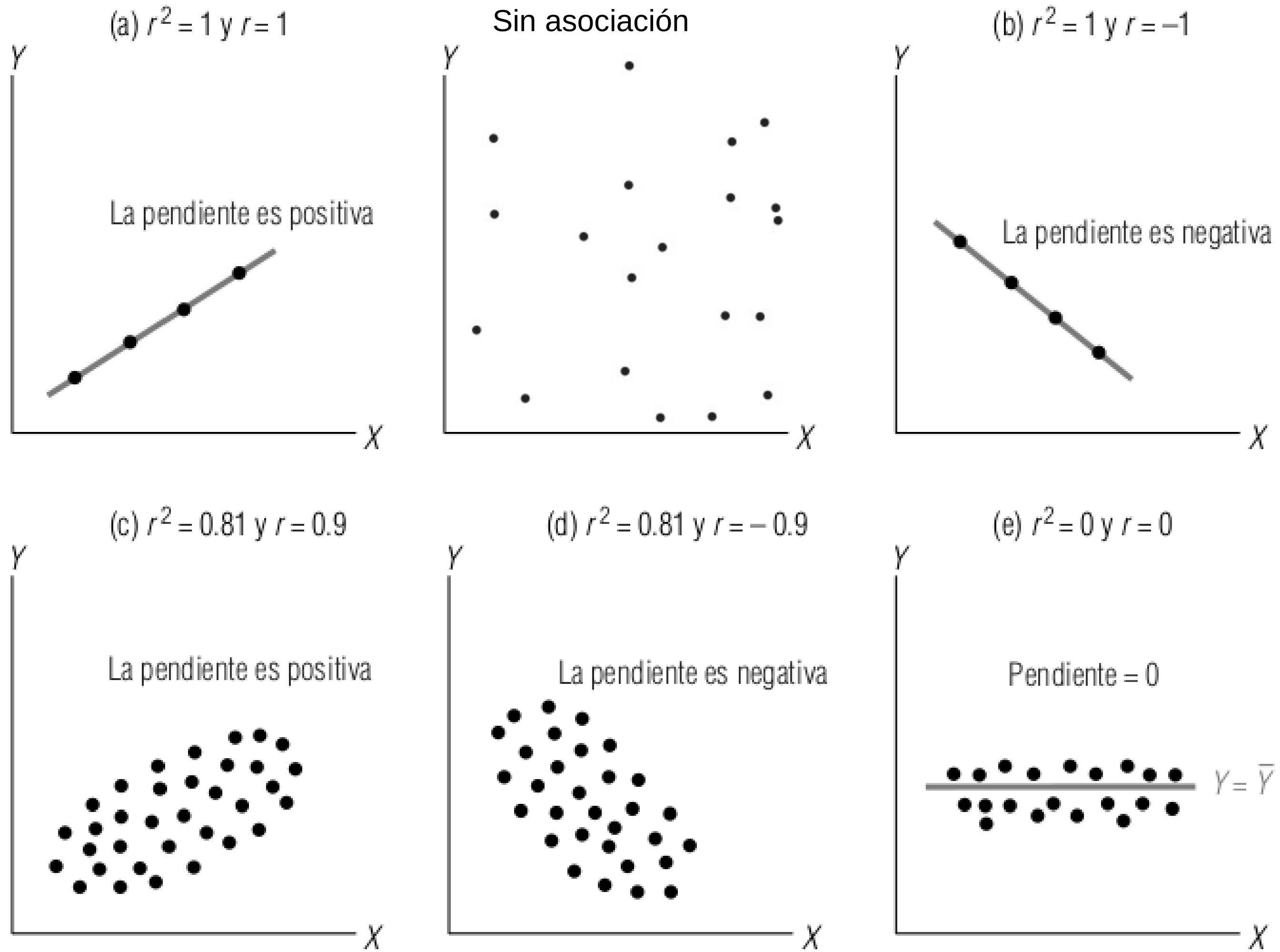
CASO NO LINEAL
Se utiliza el coeficiente de determinación r^2:

$$r^2 = 1 - \frac{\Sigma(Y - \hat{Y})^2}{\Sigma(Y - \bar{Y})^2}$$

$$-0 \leq r^2 \leq 1$$

Cuanto más se acerque a 1, mayor será la proporción de suma de cuadrados explicados por la regresión en relación a la suma de cuadrados totales.

EJEMPLOS DE DISPERSIOGRAMAS



(1994–2024)

30 años de la Consagración Constitucional de la Autonomía y Autarquía Universitaria en Argentina.

UNIVERSIDAD NACIONAL DEL LITORAL · LUX INDEFICIENS · 1919

FICH

ESTADÍSTICA
PROBABILIDAD Y ESTADÍSTICA

Semana#12 - Unidad 8 – *Series Cronológicas*

Ingenierías en: Recursos Hídricos, Ambiental y Agrimensura

Año 2024

Prof. Juan Pablo Taulamet

consultas: taulamet@unl.edu.ar

- *Repaso general*
  - *Unidad 5: Estadística Descriptiva*
  - *Unidad 6: Estadística Inferencial*
  - *Unidad 7: Regresión y Correlación*

- *Estadística Descriptiva*
  - *Análisis Exploratorio de datos*
    - *Resumen de estadísticas*
    - *Cuantiles y otros*
    - *Gráficas*

- *Estadística Descriptiva*

- *Estadística Inferencial*
  - *Distribución Muestreo*
  - *Estimación por intervalos de confianza*
  - *Pruebas de hipótesis*
    - *Paramétricas*
    - *No Paramétricas*

- *Estadística Descriptiva*

- *Estadística Inferencial*

- *Regresión y correlación*

confianza

  - *Análisis de regresión*
  - *Análisis de correlación*
  - *Análisis de varianza*
  - *Análisis de residuos*

RECAPITULANDO CON FIBRONES

U1

A: " . . . . . . "

$P(A \cup B) = ?$

U2

X: V.A.

$P(X=2)=f(2)$

U3

$$E(x) = \int x \cdot f(x)\,dx$$

consultas: taulamet@unl.edu.ar

U4

$$X \sim N(\mu;\sigma)$$

$$P(X \leq 4) = F(4)$$

$$= \text{normdist}(4;\mu;\sigma;1)$$

U5

Muestra

↳ $\overline{X}$

UGA

$\bar{X}$: V.A.

$P(\bar{X} \leq 9) =$ $\bar{X} \sim N(\mu_{\bar{X}}; \sigma_{\bar{X}})$

$= \text{normdist} \ldots$ $\mu_{\bar{X}} = \mu_{X}$

$\sigma_{\bar{X}} = \sigma_{X} / \sqrt{n}$

UGB

$$\bar{X} = \mu \,?$$

UGB

$IC_{p\mu}$ $\bar{x} \pm Algo$

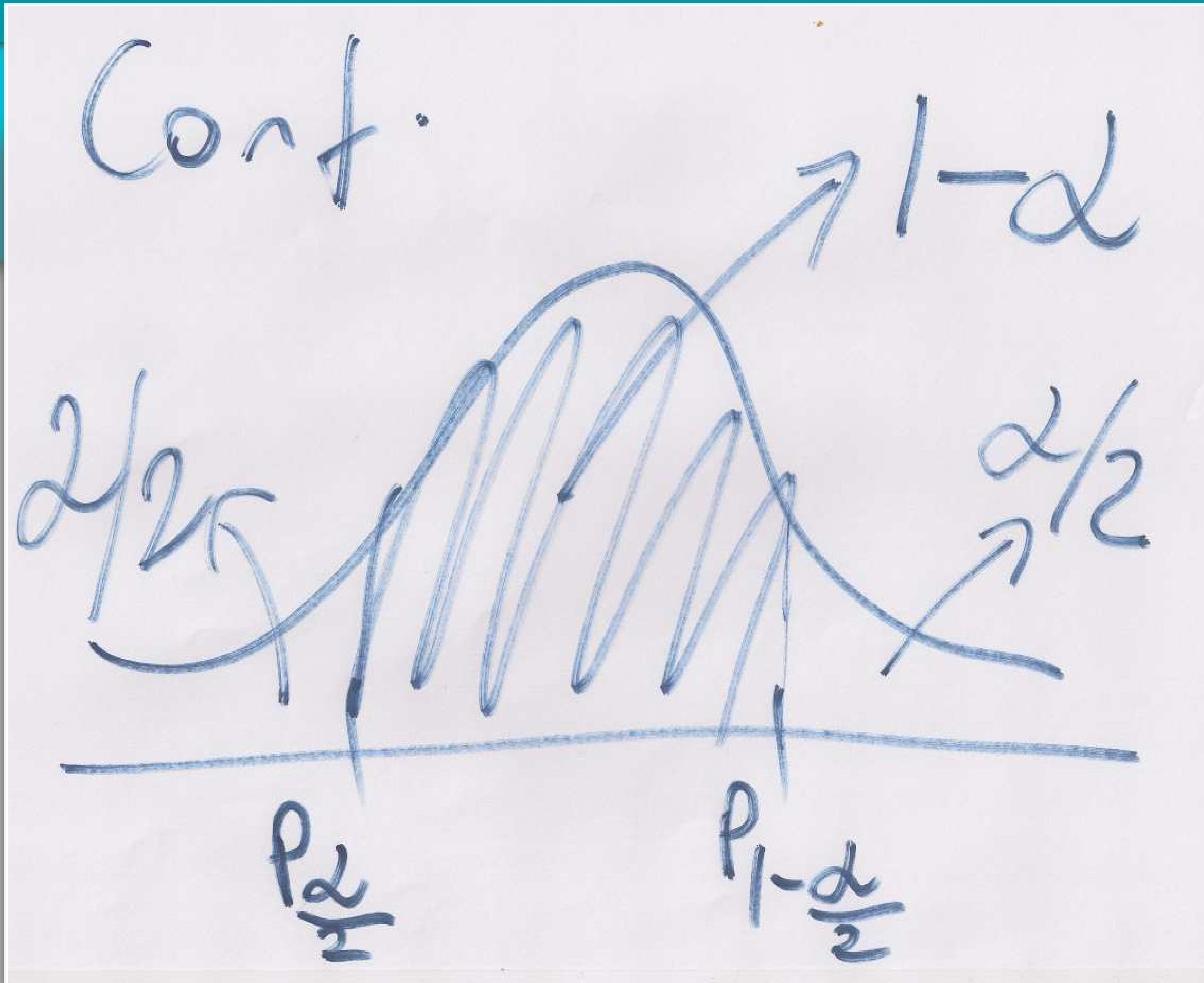
Algo → Conf.

↘ E.S.

consultas: taulamet@unl.edu.ar

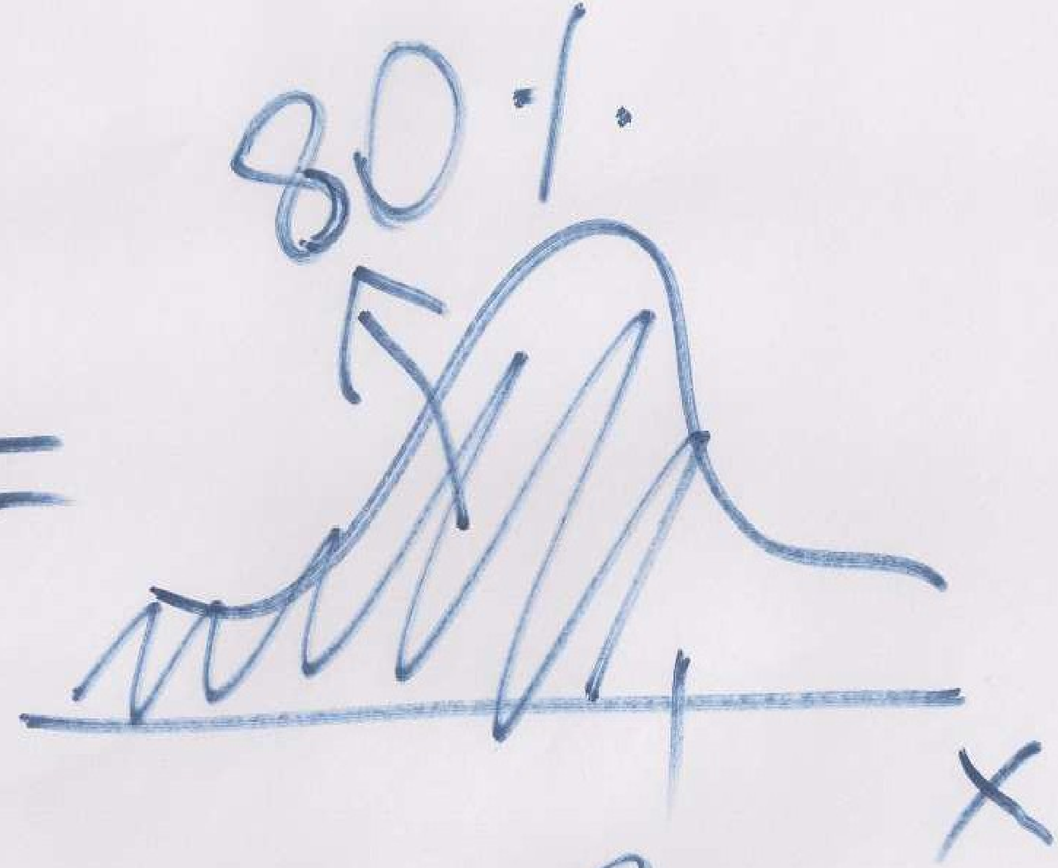
E.S.

$$\bar{x} : \frac{\sigma}{\sqrt{n}}$$


Conf:
1-α
α/2
α/2
P α/2
P 1-α/2

$F(P_{80}) = 80\%$

$P_{80} =$



P_{80}

$P(X < P_{80}) = 80\%$

Percentiles en GNUmeric

$$\mathrm{P}(\mathrm{X} \leq x_1) = 80\%$$

$$\implies x_1 = P_{80}$$

$$[\,\mathrm{X} \sim \mathcal{N}(\mu,\, \sigma)]$$

$$\implies x_1 = norminv(\mu, \sigma, 80\%)$$

$$\mathrm{X} \sim T_{dof}$$

$$\implies x_1 = tinv(40\%, dof)$$

$$\mathrm{X} \sim \chi_{dof}$$

$$\implies x_1 = chiinv(20\%, dof)$$

¡Año Electoral!

Ejercicio 1

Según los dirigentes del partido A, la intención de voto del partido rival B, en la ciudad I, es la misma que la que tiene en la ciudad II. Se realiza una encuesta a 100 personas en la ciudad I de los que 25 mostraron su apoyo al partido B, y a otras 100 personas en la ciudad II de las que 30 se declaran simpatizantes del partido B.

a) ¿Podría decir cuál sería la proporción real de personas que votarían al partido B en la ciudad I? Realícelo y justifique.

b) ¿A cuántas personas habría que encuestar para tener un margen de error de 2%?

¡Año Electoral!

a) “Prop. de votos a B en CII” = ?

$$\pi_{BenII} = ?$$

Datos:

$$\text{p} = 25/100$$

$$\text{n} = 100$$

$$1 - \alpha = ?$$

¡Año Electoral!

a) I.C. para: $\pi_{BenII} = ?$

Elegimos una confianza: $1 - \alpha = 90\%$

$p = 25/100 \quad n = 100$

$$\left(p \pm \left| Z_{\left(1-\frac{\alpha}{2}\right)} \right| \sqrt{\frac{\pi(1-\pi)}{n}} \right)$$

¡Año Electoral!

a) I.C. para: π_{BenII}

1- $\alpha = 90\% \Rightarrow Z_{95\%} \approx 1,64$

p= 25/100 n= 100

Estimamos π

Usando p

Y obtenemos:

I.C. $0{,}25 \pm 1,64\sqrt{\frac{0{,}25 \times 0{,}75}{100}}$

consultas: taulamet@unl.edu.ar

¡Año Electoral!

Ejercicio 1

Según los dirigentes del partido A, la intención de voto del partido rival B, en la ciudad I, es la misma que la que tiene en la ciudad II. Se realiza una encuesta a 100 personas en la ciudad I de los que 25 mostraron su apoyo al partido B, y a otras 100 personas en la ciudad II de las que 30 se declaran simpatizantes del partido B.

a) ¿Podría decir cuál sería la proporción real de personas que votarían al partido B en la ciudad I? Realícelo y justifique.

b) ¿A cuántas personas habría que encuestar para tener un margen de error de 2%?

¡Año Electoral!

b)Para un error del 2%

$$\left(p \pm \underbrace{\left| Z_{\left(1-\frac{\alpha}{2}\right)} \right| \sqrt{\frac{\pi(1-\pi)}{n}}}\right)$$

$$n = ?$$

¡Año Electoral!

b)Para un error del 2%

$$\left(p \pm \overbrace{\left| Z_{\left(1-\frac{\alpha}{2}\right)} \right| \sqrt{\frac{\pi(1-\pi)}{n}}} \right)$$

$$2\% = 1{,}64 \sqrt{\frac{0{,}25 \times 0{,}75}{n}} \Rightarrow n \approx 1268$$

consultas: taulamet@unl.edu.ar

Durante los últimos 10 días de junio un tren llegó tarde a su destino en los siguientes tiempos (en minutos; un número negativo significa que el tren llegó temprano ese número de minutos):

-3	6	4	10	-4	124	2	-1	4	1

Obtener un intervalo de confianza para estimar la esperanza poblacional de los tiempos de de llegada tarde a destino del tren en minutos, con un 90% de confianza.

$$\left(\bar{x} \pm \left| t_{\left(1-\frac{\alpha}{2};n-1\right)} \right| \frac{S'}{\sqrt{n}} \right)$$

consultas: taulamet@unl.edu.ar

$$\left(\bar{x} \pm \left| t_{\left(1-\frac{\alpha}{2};n-1\right)} \right| \frac{S'}{\sqrt{n}}\right)$$

Media	14,3
Error estándar	12,2611

CONF	90,00%
T95	1,83
Error	22,48
LI	-8,2
LS	34,7

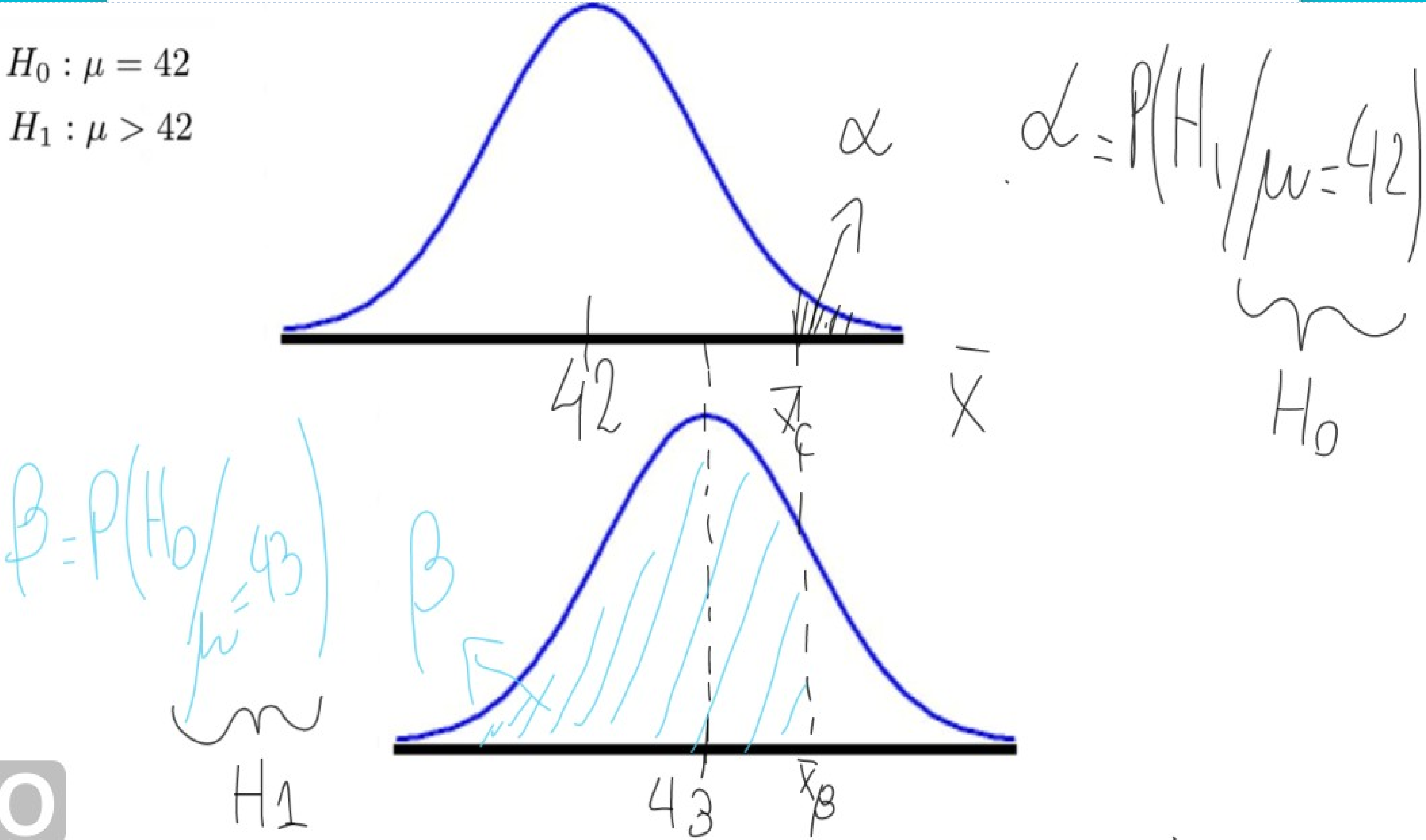
Respuesta: Con un 90% de confianza, el promedio poblacional de los tiempos de demora en llegar a destino del tren se encontrará entre -8,2 y 34,7, en símbolos:

$$\text{IC p/ } \mu, 1-\alpha = 90\% : (-8,2; 34,7)$$

REPASO Y POTENCIA

$H_0 : \mu = 42$

$H_1 : \mu > 42$



<table>
<tr><th>N° de tormentas por estación y por año.</th><th colspan="2">N° de ocurrencias observadas.</th><th>$\frac{e^{-\lambda}\lambda^{x}}{x!}$</th><th colspan="2">Fe=n*P</th><th>$\frac{(fo-fe)^2}{fe}$</th></tr>
<tr><td>0</td><td colspan="2">102</td><td>0.3073</td><td colspan="2">110.62</td><td>0.67</td></tr>
<tr><td>1</td><td colspan="2">144</td><td>0.3626</td><td colspan="2">130.53</td><td>1.39</td></tr>
<tr><td>2</td><td colspan="2">74</td><td>0.2139</td><td colspan="2">77.01</td><td>0.12</td></tr>
<tr><td>3</td><td colspan="2">28</td><td>0.0841</td><td colspan="2">30.29</td><td>0.17</td></tr>
<tr><td>4</td><td>10</td><td rowspan="3">12</td><td>0.025</td><td>8.94</td><td rowspan="3">11.46</td><td rowspan="3">0.025</td></tr>
<tr><td>5</td><td>2</td><td>0.0059</td><td>2.11</td></tr>
<tr><td>6</td><td>0</td><td>0.0011</td><td>0.41</td></tr>
<tr><td></td><td colspan="2"></td><td></td><td colspan="2">359.91</td><td>2.375 $\chi^2_{observado}$</td></tr>
</table>

ESTUDIO DE ASOCIACIÓN ENTRE VARIABLES

REGRESIÓN

ENCONTRAR UN MODELO

CORRELACIÓN

EXACTITUD DEL MODELO

ANÁLISIS DE REGRESIÓN

Permite encontrar un modelo que vincula a dos o más variables, brindando un mecanismo de pronóstico.

ANÁLISIS DE CORRELACIÓN

Determina la medida del grado de exactitud de la relación entre las variables

consultas: taulamet@unl.edu.ar

La siguiente tabla muestra valores de precipitación y escurrimiento en una cuenca urbana.

x (mm)	y (mm)
50	60
60	60
70	80
80	70
90	80
100	100
110	100
120	130
130	110
140	120
150	150
160	140

La siguiente tabla muestra valores de precipitación y escurrimiento en una cuenca urbana.

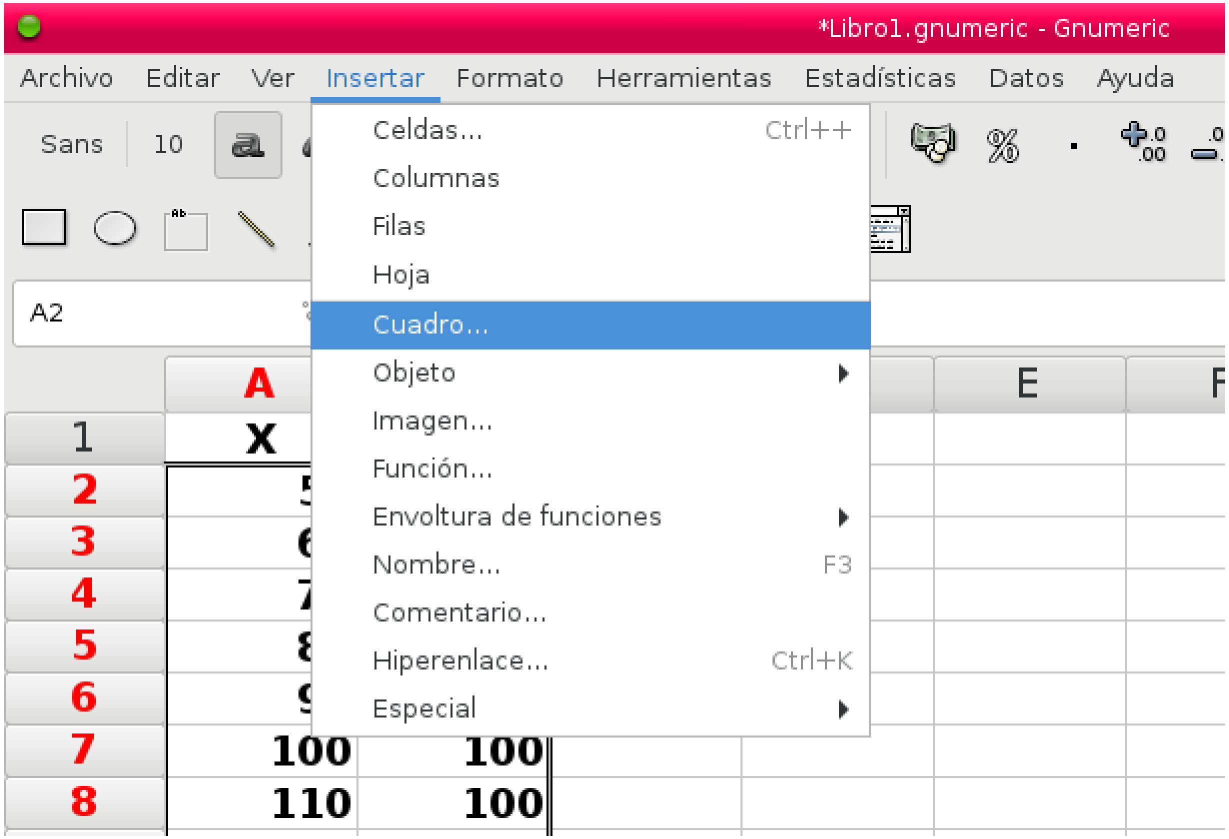
x (mm)	y (mm)
50	60
60	60
70	80
80	70
90	80
100	100
110	100
120	130
130	110
140	120
150	150
160	140

¿Existe relación entre X e Y?

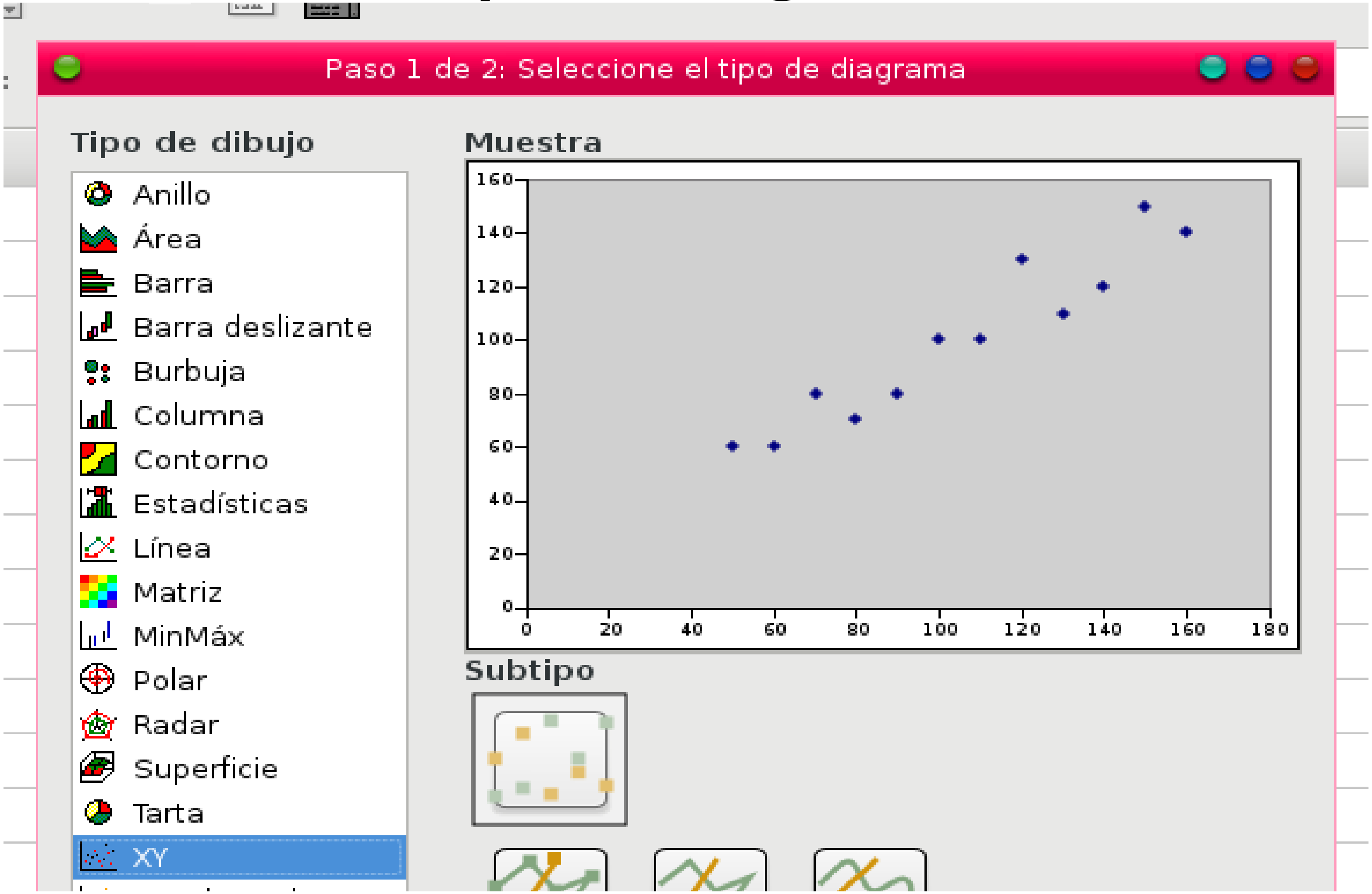
En Gnumeric

	A	B	C	D
1	**X**	**Y**		
2	**50**	**60**		
3	**60**	**60**		
4	**70**	**80**		
5	**80**	**70**		
6	**90**	**80**		
7	**100**	**100**		
8	**110**	**100**		
9	**120**	**130**		
10	**130**	**110**		
11	**140**	**120**		
12	**150**	**150**		
13	**160**	**140**		
14				

Insertamos un Dispersiograma

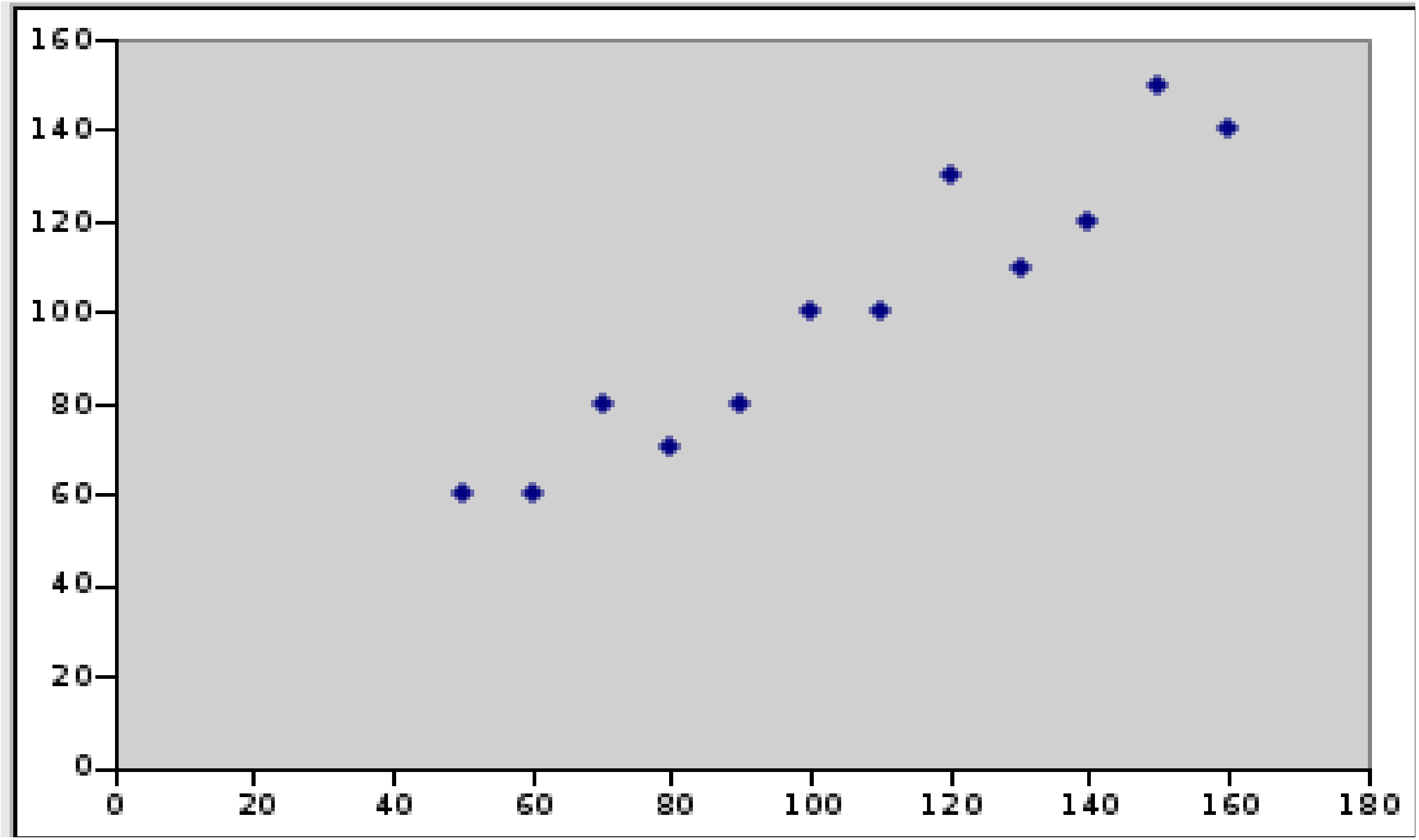


Insertamos un Dispersiograma



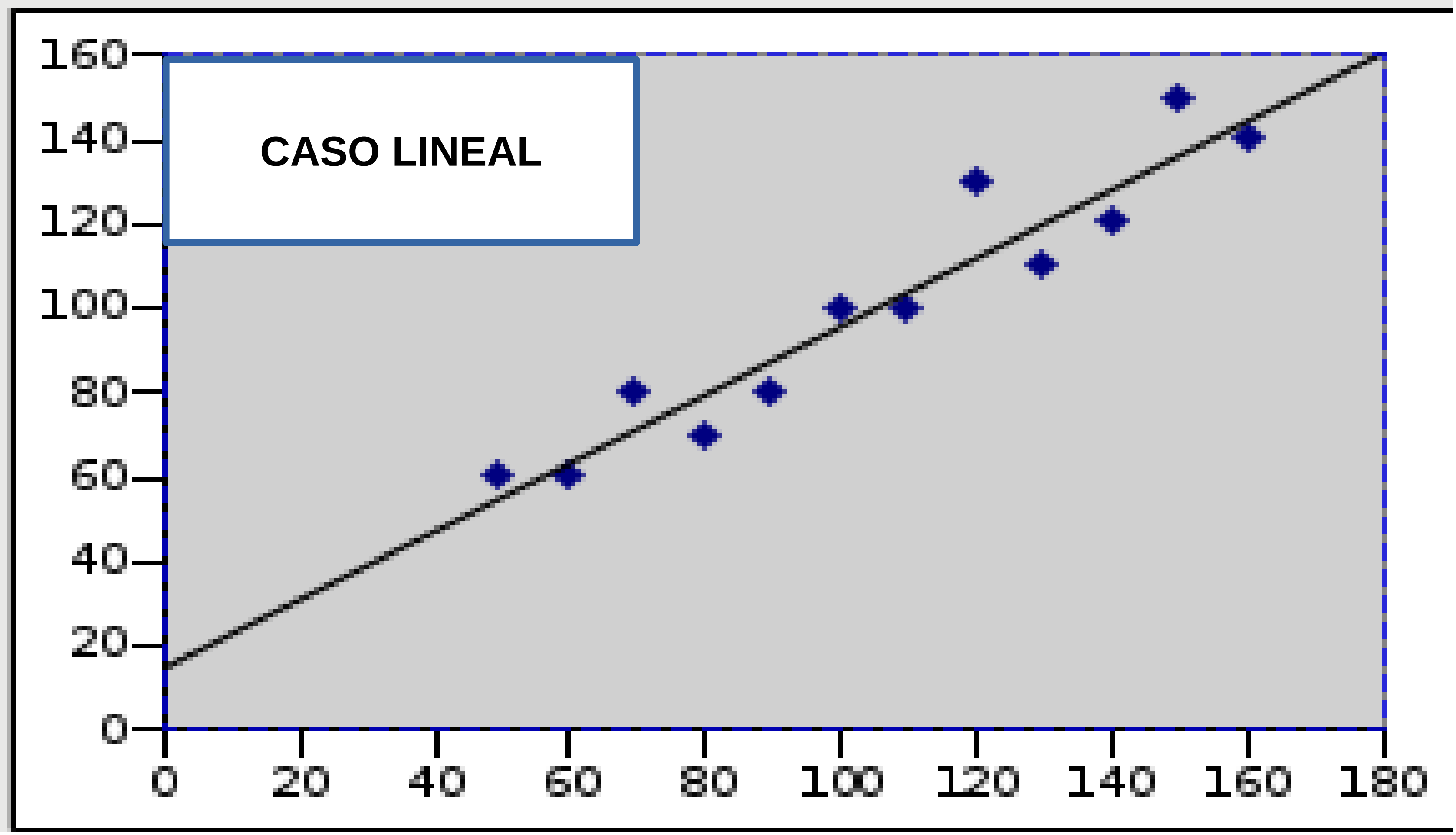
Dispersiograma

¿Existe relación entre X e Y? ¿Cómo es?



MODELO DE REGRESIÓN BIVARIADO

$$E(Y/x) = \mu_{Y/x} = \alpha + \beta X_i$$



consultas: taulamet@unl.edu.ar

MODELO DE REGRESIÓN LINEAL SIMPLE

PARÁMETROS DEL MODELO

VARIABLE EXPLICATIVA INDEPENDIENTE

VARIABLE DEPENDIENTE

ERROR ALEATORIO

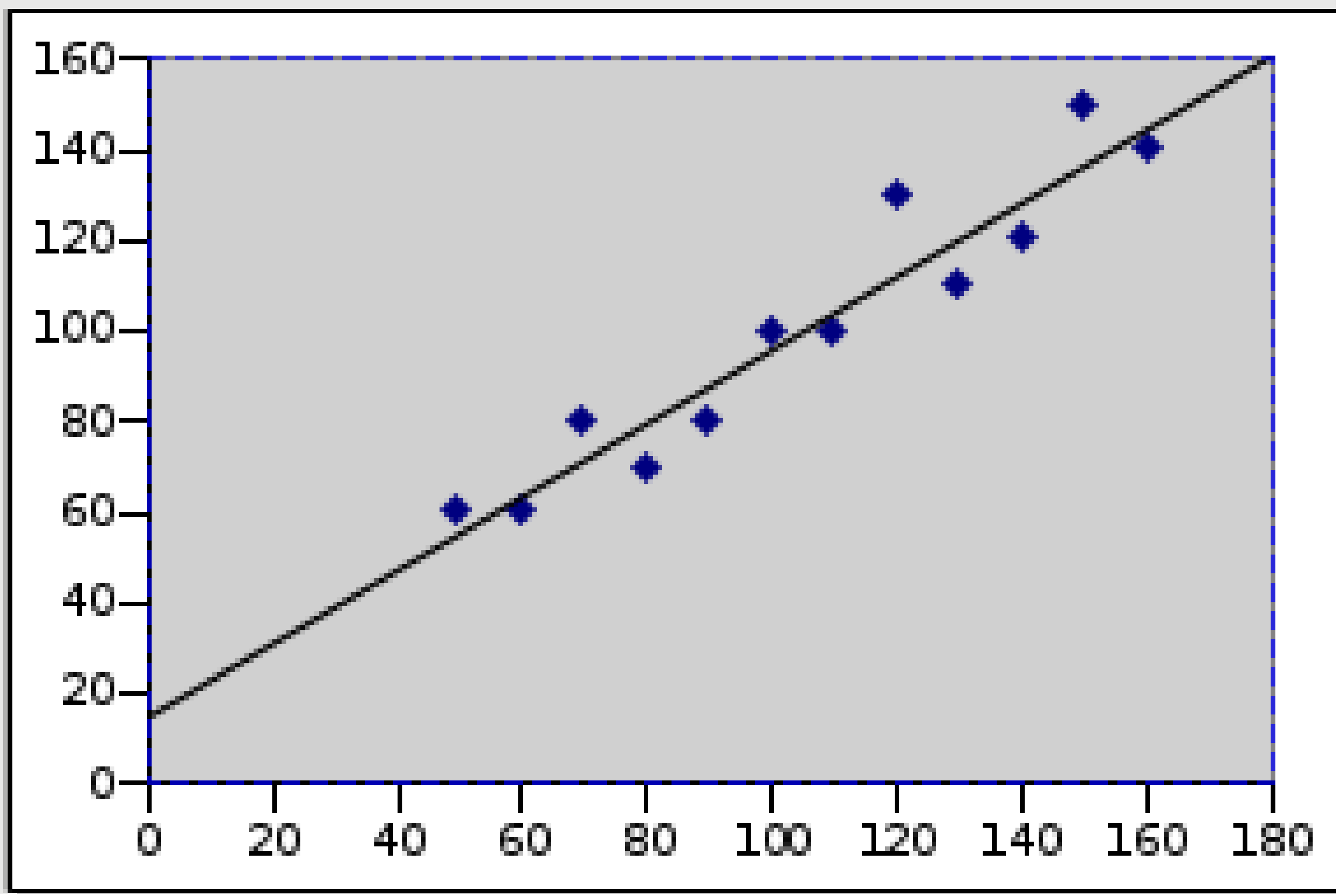
$$Y_i = \underbrace{\alpha + \beta X_i}_{\text{PARTE SISTEMÁTICA}} + \underbrace{\in_i}_{\text{PARTE ESTOCÁSTICA}}$$

PARTE SISTEMÁTICA

PARTE ESTOCÁSTICA

consultas: taulamet@unl.edu.ar

PENSANDO ESTA RELACIÓN ESTADÍSTICA



consultas: taulamet@unl.edu.ar

PENSANDO ESTA RELACIÓN ESTADÍSTICA

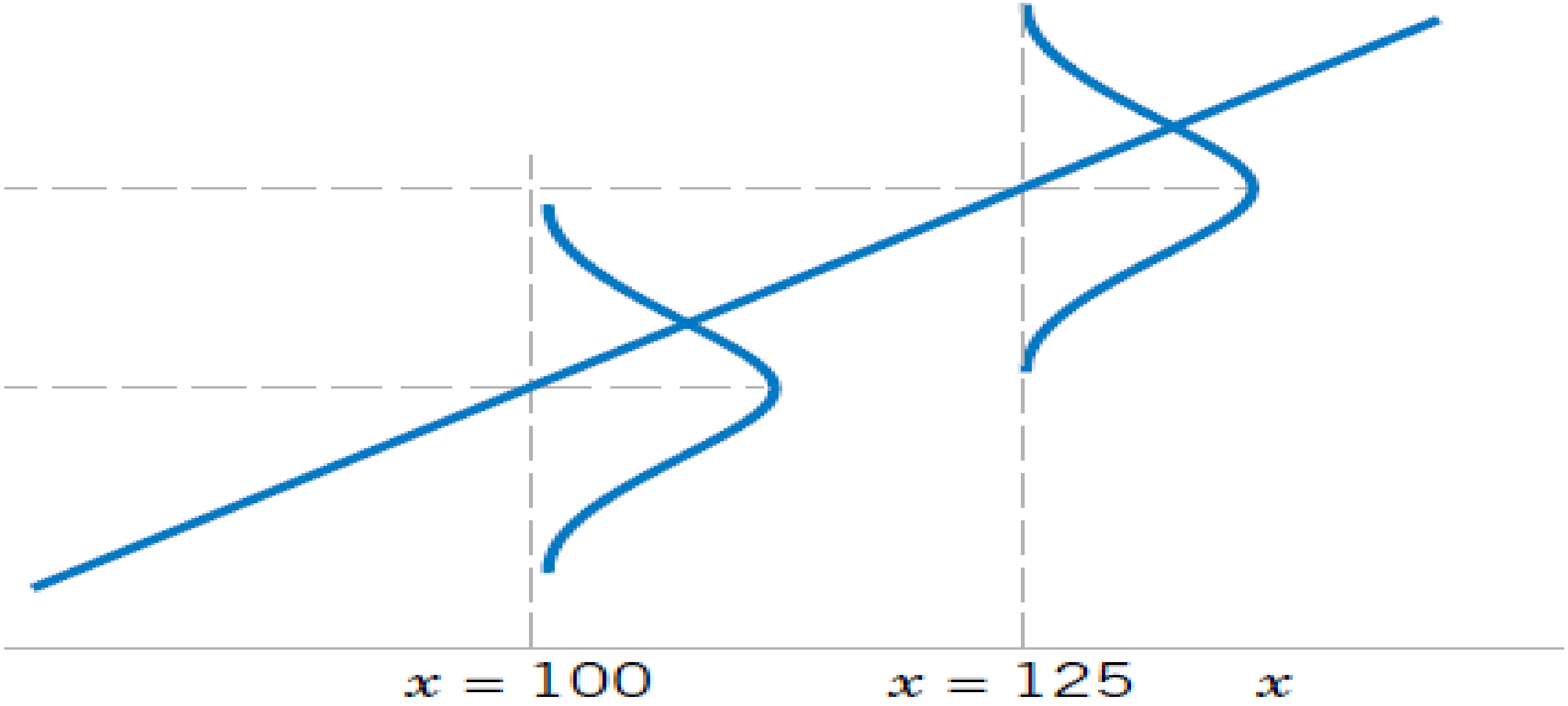
La variable dependiente tiende a variar con la variable independiente de una manera sistemática.

Existe dispersión de los puntos entorno a la curva que relaciona las variables.

consultas: taulamet@unl.edu.ar

PENSANDO ESTA RELACIÓN ESTADÍSTICA

Existe una distribución de probabilidades de la variable dependiente (Y) para cada valor de X. Existen grupos de valores de Y para cada valor de X que llamaremos sub-poblaciones. El modelo de regresión es el que *relaciona los valores medios* de dichas sub-poblaciones:



consultas: taulamet@unl.edu.ar

MODELOS DE REGRESIÓN

LINEAL
NO-LINEAL
(Polinomial, exponencial, etc.)

SIMPLE
(2 Variables)
MÚLTIPLE
(Más de 2 Variables)

MODELO DE REGRESIÓN LINEAL SIMPLE

$$Y_i = \alpha + \beta X_i + \in_i$$

Para conocer más del modelo supondremos que la X se fija en un valor y hallaremos la esperanza y varianza del mismo.

MODELO DE REGRESIÓN LINEAL SIMPLE

Supongamos que la media y la varianza de $\in$ son 0 y σ^2

$$E(Y/x) = E(\alpha + \beta x + \varepsilon) = \alpha + \beta x + E(\varepsilon) = \alpha + \beta x$$

$$Var(Y/x) = Var(\alpha + \beta x + \varepsilon) = Var(\alpha + \beta x) + Var(\varepsilon) = 0 + \sigma^2 = \sigma^2$$

ESTIMACIÓN POR MÍNIMOS CUADRADOS

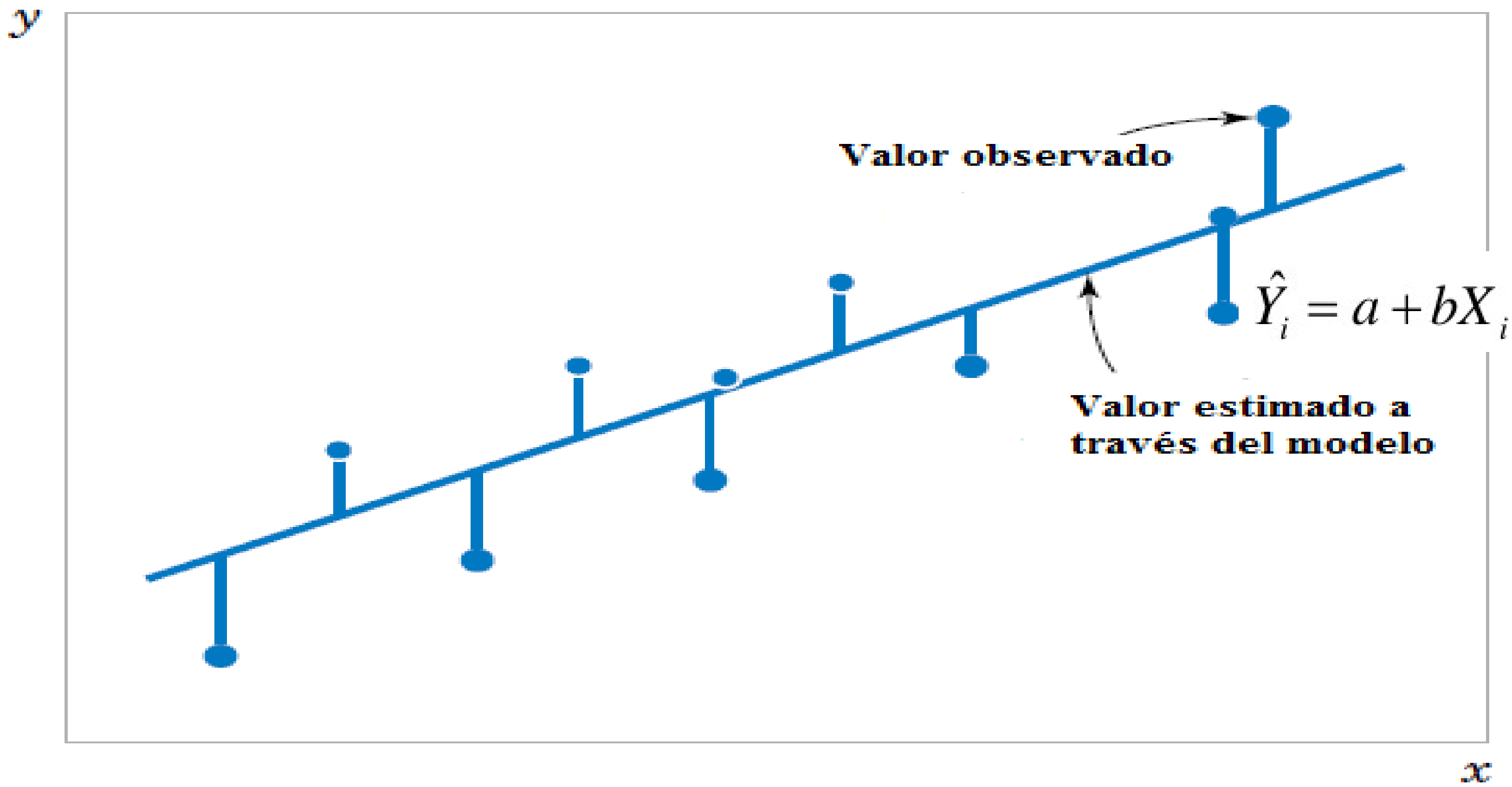
SUPUESTOS:

**La VA del error es independiente.*

**La distribución del error es normal*

**X es un conjunto de números fijos*

ESTIMACIÓN POR MÍNIMOS CUADRADOS



$$S = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

$$\frac{\partial S}{\partial \alpha} = 0 \quad \text{y} \quad \frac{\partial S}{\partial \beta} = 0$$

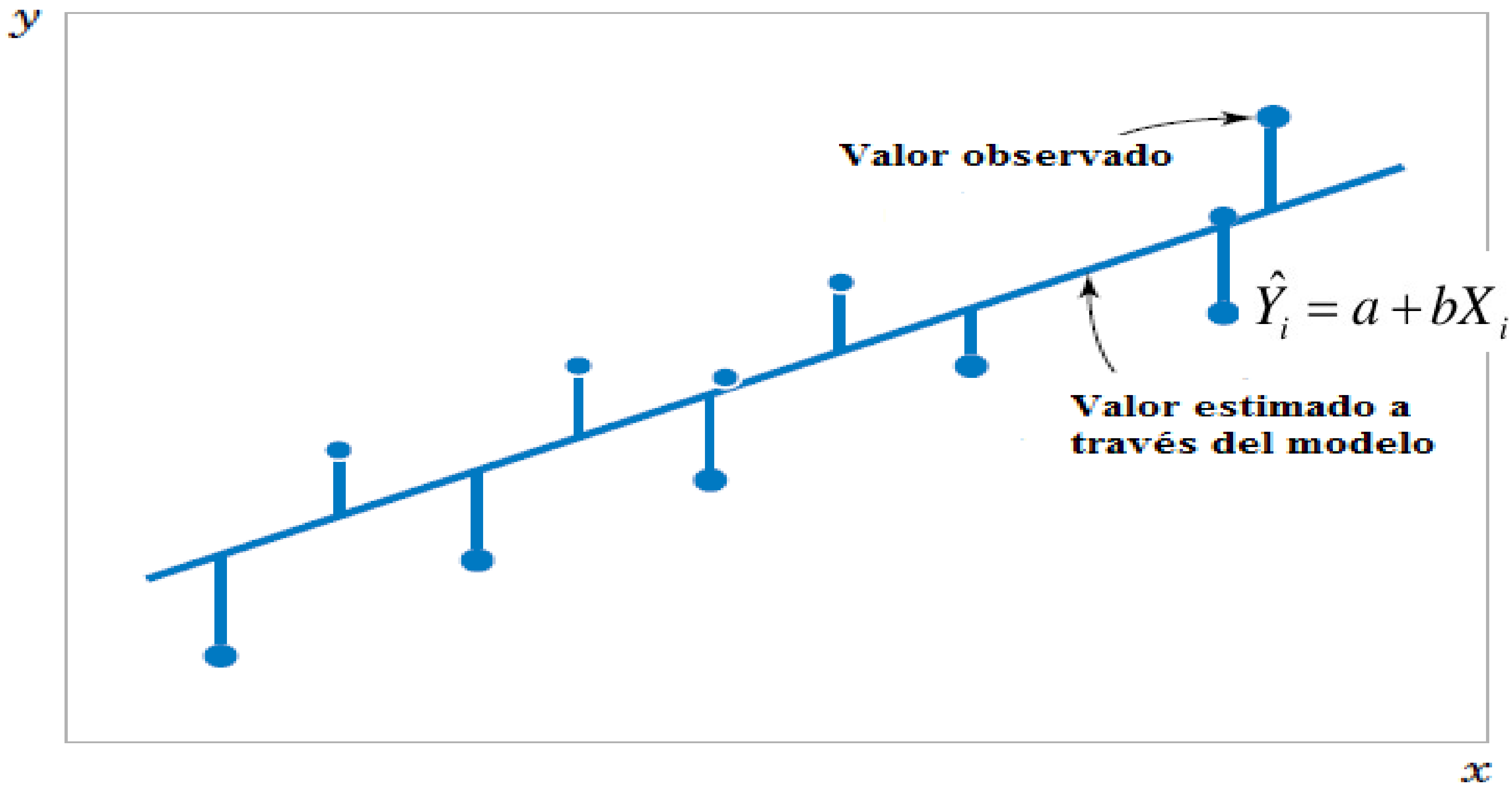
ESTIMACIÓN POR MÍNIMOS CUADRADOS

$$\hat{\beta} = b = \frac{m_{1,1}}{S_x^2} = \frac{\text{cov}}{S_x^2}$$

$$\hat{\alpha} = a = \overline{Y} - \hat{\beta}\overline{X}$$

Este método minimiza las distancias verticales

VARIANZA DE LA REGRESIÓN



$$S^2_{y/x} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-2}$$

consultas: taulamet@unl.edu.ar

ANÁLISIS DE CORRELACIÓN

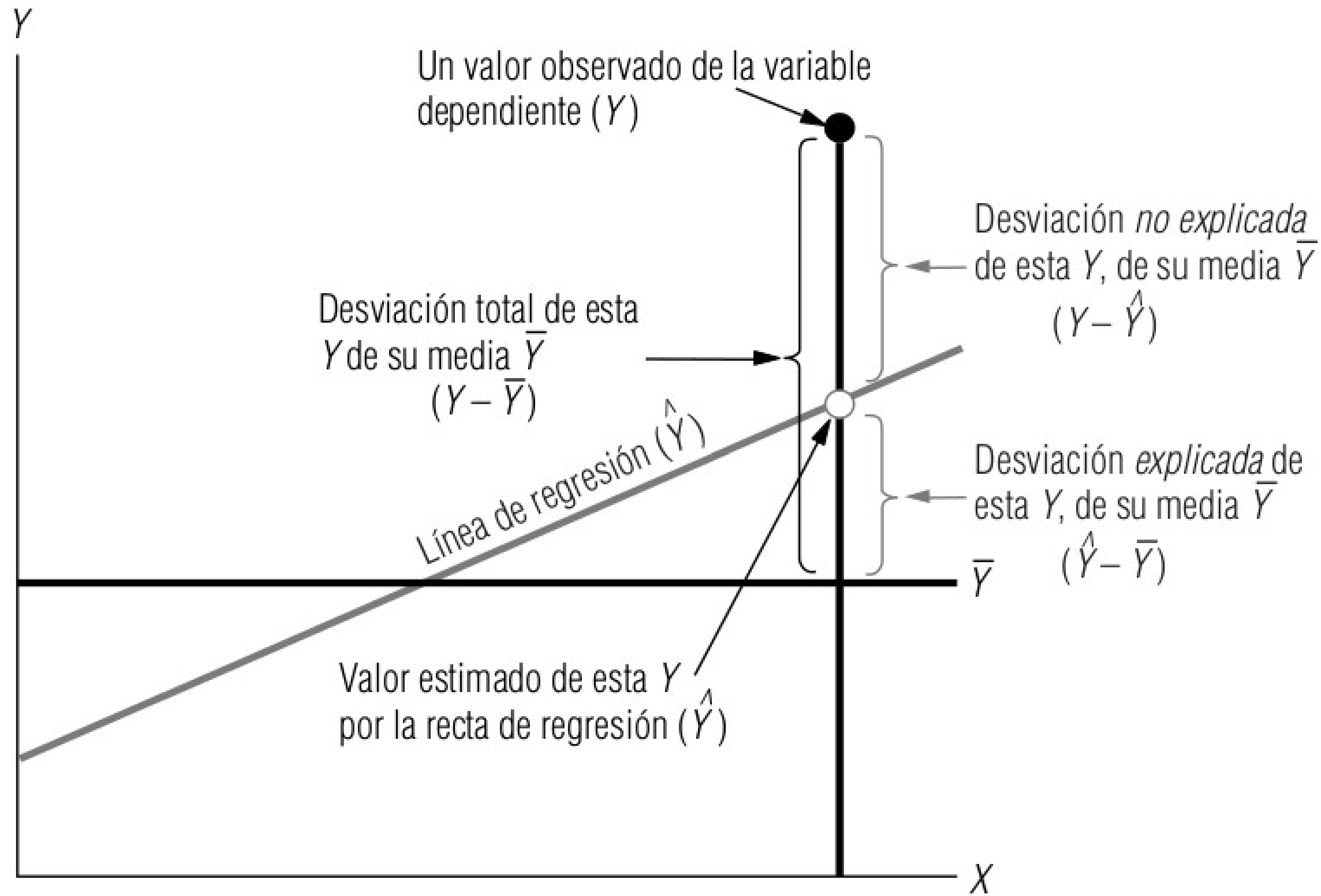
CASO LINEAL
Se utiliza el coeficiente de correlación r:

$$r = \hat{\rho} \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}} = \frac{m_{1,1}}{S_x S_y}$$

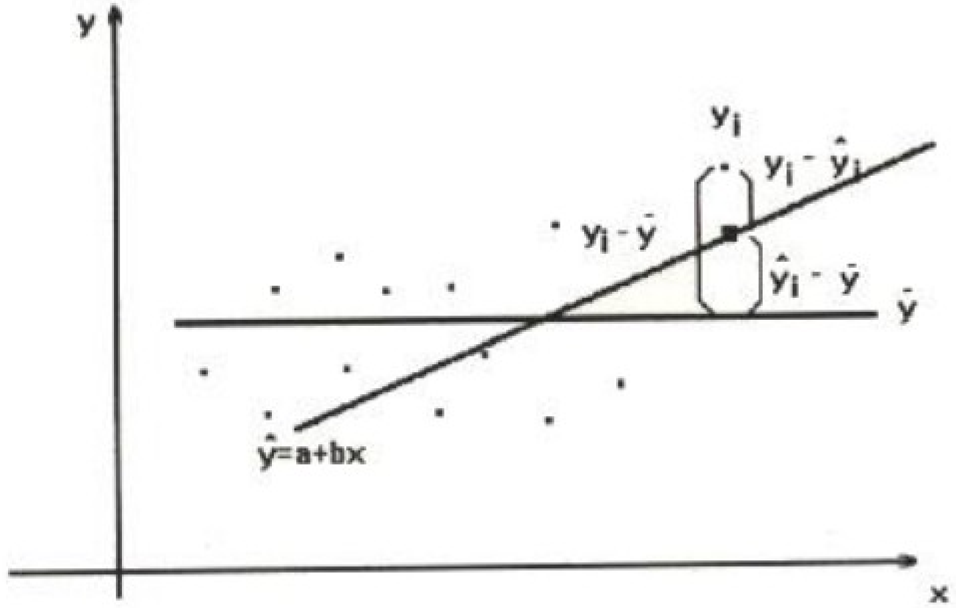
$$-1 \leq r \leq 1$$

Cuanto más cerca se encuentre del 0 menor relación lineal habrá entre la X y la Y. En la medida que se acerque a 1 o a -1, expresará un mayor grado de asociación lineal ya sea relación directa o inversamente proporcional.

PARTICIÓN DE LOS CUADRADOS



PARTICIÓN DE LOS CUADRADOS



$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2$$

$$\sum_i (y_i - \bar{y})^2 = \textit{suma de cuadrados totales o error total (SCT)}$$

$$\sum_i (\hat{y}_i - \bar{y})^2 = \textit{suma de cuadrados explicados por la regresión (SCR)}$$

$$\sum_i (y_i - \hat{y}_i)^2 = \textit{suma de cuadrados no explicados (SCE)}$$

ANÁLISIS DE CORRELACIÓN

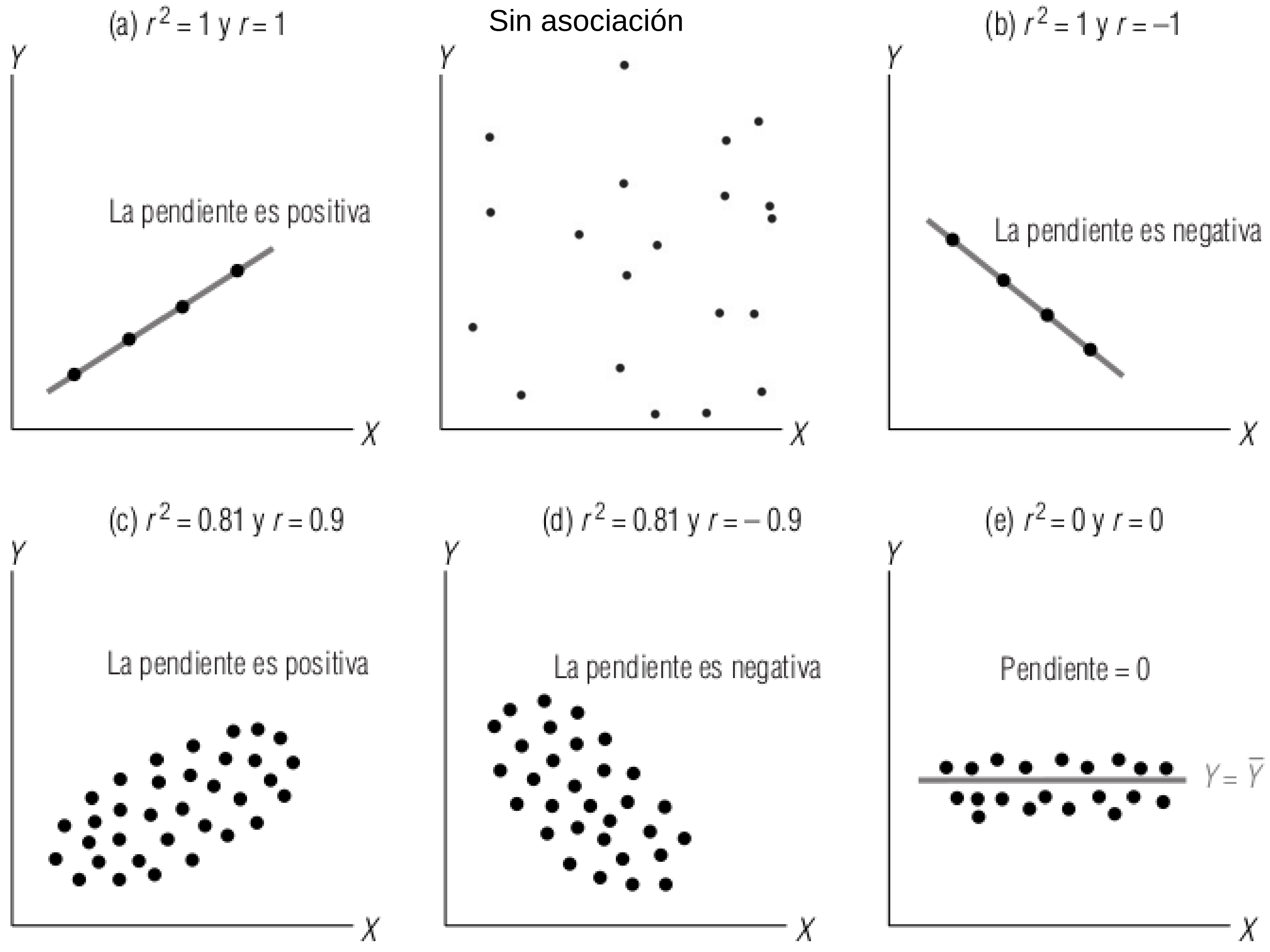
CASO NO LINEAL
Se utiliza el coeficiente de determinación r^2:

$$r^2 = 1 - \frac{\Sigma(Y - \hat{Y})^2}{\Sigma(Y - \bar{Y})^2}$$

$$-0 \leq r^2 \leq 1$$

Cuanto más se acerque a 1, mayor será la proporción de suma de cuadrados explicados por la regresión en relación a la suma de cuadrados totales.

EJEMPLOS DE DISPERSIOGRAMAS



consultas: taulamet@unl.edu.ar

2020

Año del General
Manuel Belgrano

Facultad de Ingeniería y
Ciencias Hídricas
~ 50 aniversario ~

CLASE VIRTUAL DEL 24 DE JUNIO

Unidad 7 – *Regresión y Correlación*

Estadística - Ingeniería en Informática

Año 2020

Prof. Juan Pablo Taulamet

consultas: taulamet@unl.edu.ar

EJEMPLO – COVID 19

Unidad 7 – *Regresión y Correlación*

Estadística - Ingeniería en Informática

Año 2020

Prof. Juan Pablo Taulamet

consultas: taulamet@unl.edu.ar

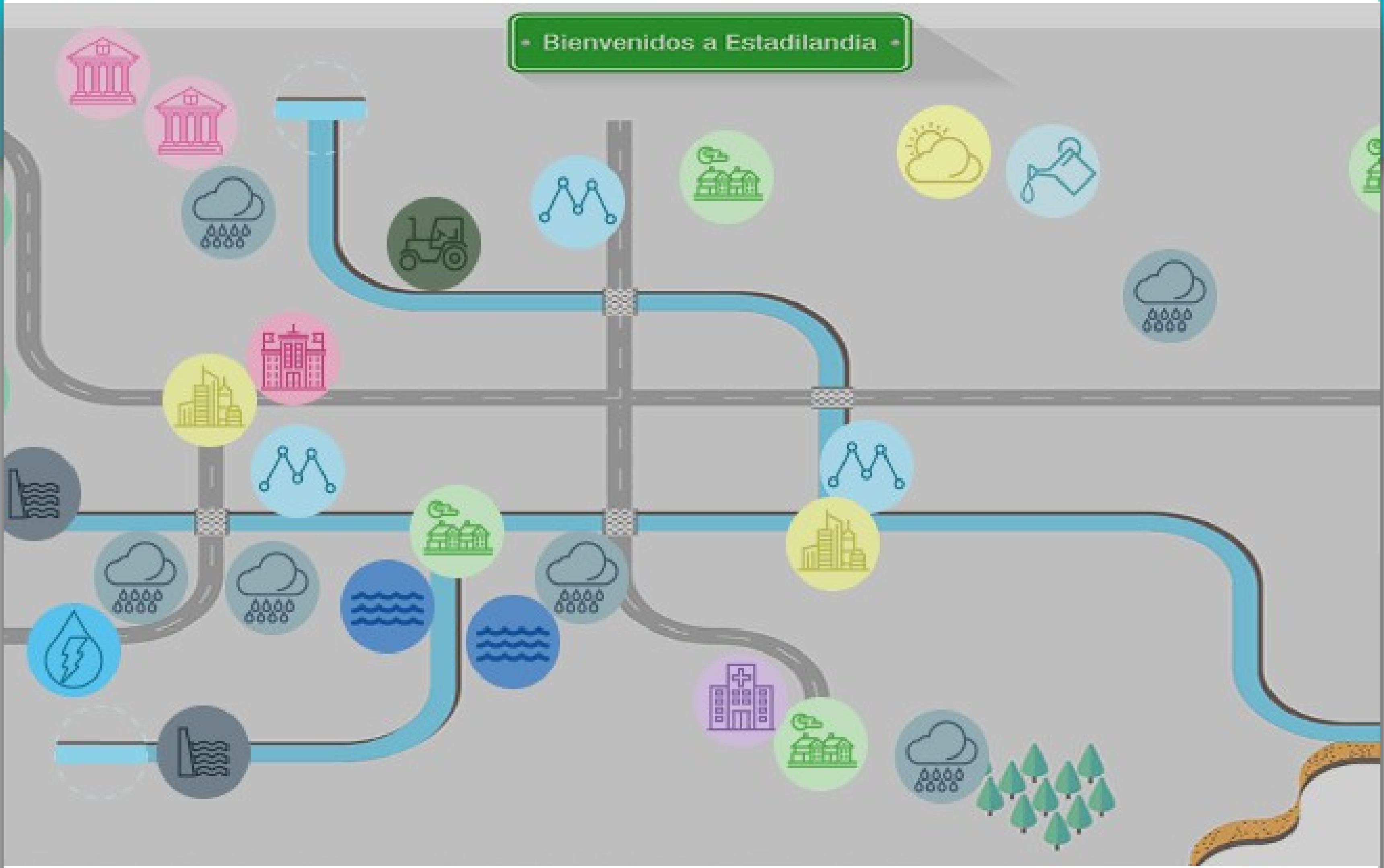
Múltiples variables

- Casos positivos
- Casos recuperados
- Casos acumulados
- Camas disponibles
- Proporción de enfermos graves
- Cantidad de muertes

Múltiples variables

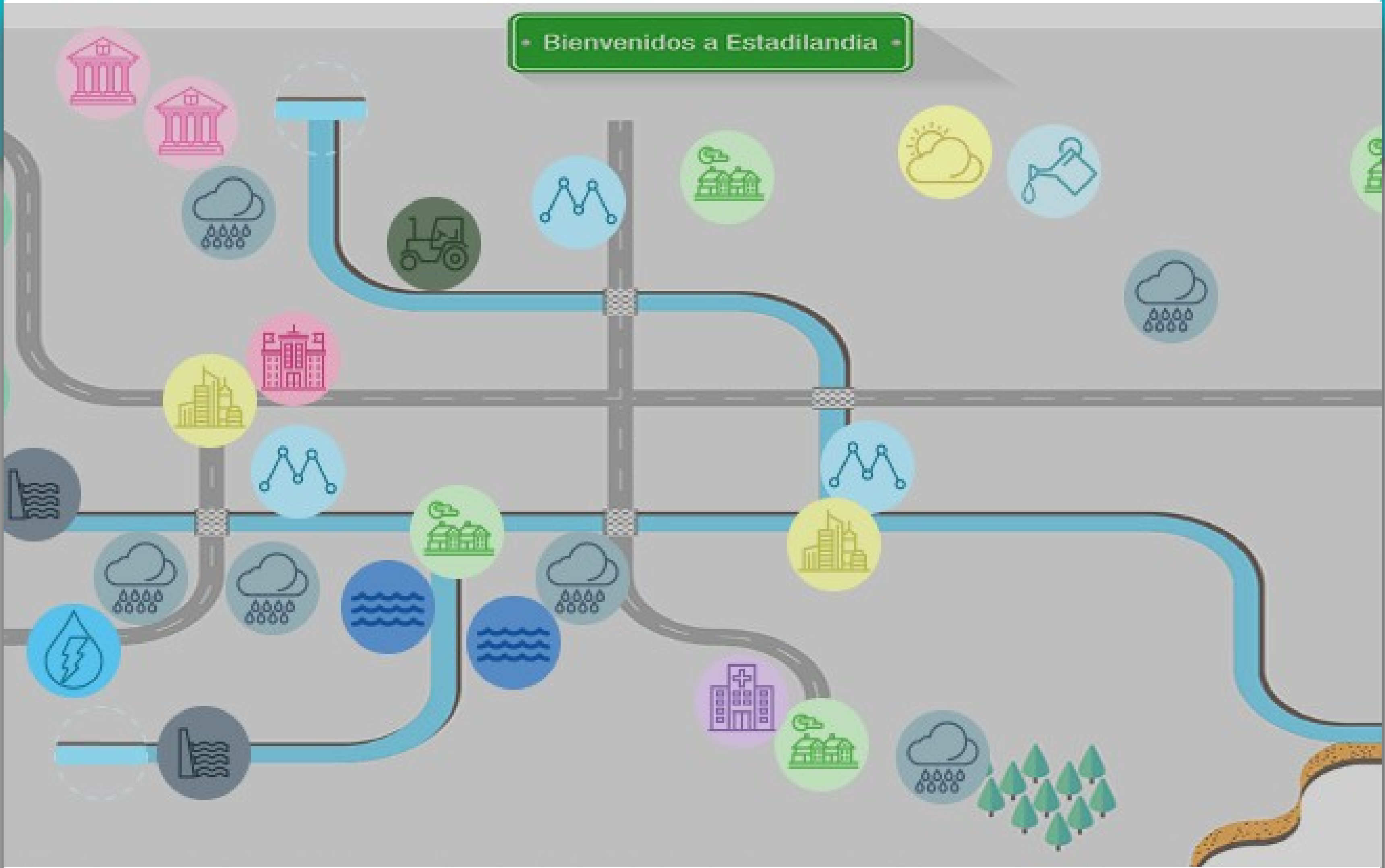
- **Casos positivos**
- Casos recuperados
- Casos acumulados
- Camas disponibles
- Proporción de enfermos graves
- Cantidad de muertes

DATOS



consultas: taulamet@unl.edu.ar

DATOS



DATOS

Hospital

Modificar Eliminar Visible

Dirección de Control Sanitario

Centro de salud del pueblo MMM

Descargas [+]

[GNUMERIC] Casos COVID-19 Argentina al 23/06
Modificar Eliminar

CASOS POSITIVOS AL 23/06

E2 | = CASOS POSITIVOS DIAF

	A	B	C	D	E	F
1	D	X	Y		CASOS DE COVID-19 E	
2	3/3/2020	1	1		CASOS POSITIVOS DIA	
3	4/3/2020	2	0			
4	5/3/2020	3	1			
5	6/3/2020	4	6			
6	7/3/2020	5	0			
7	8/3/2020	6	3			
8	9/3/2020	7	5			
9	10/3/2020	8	2			

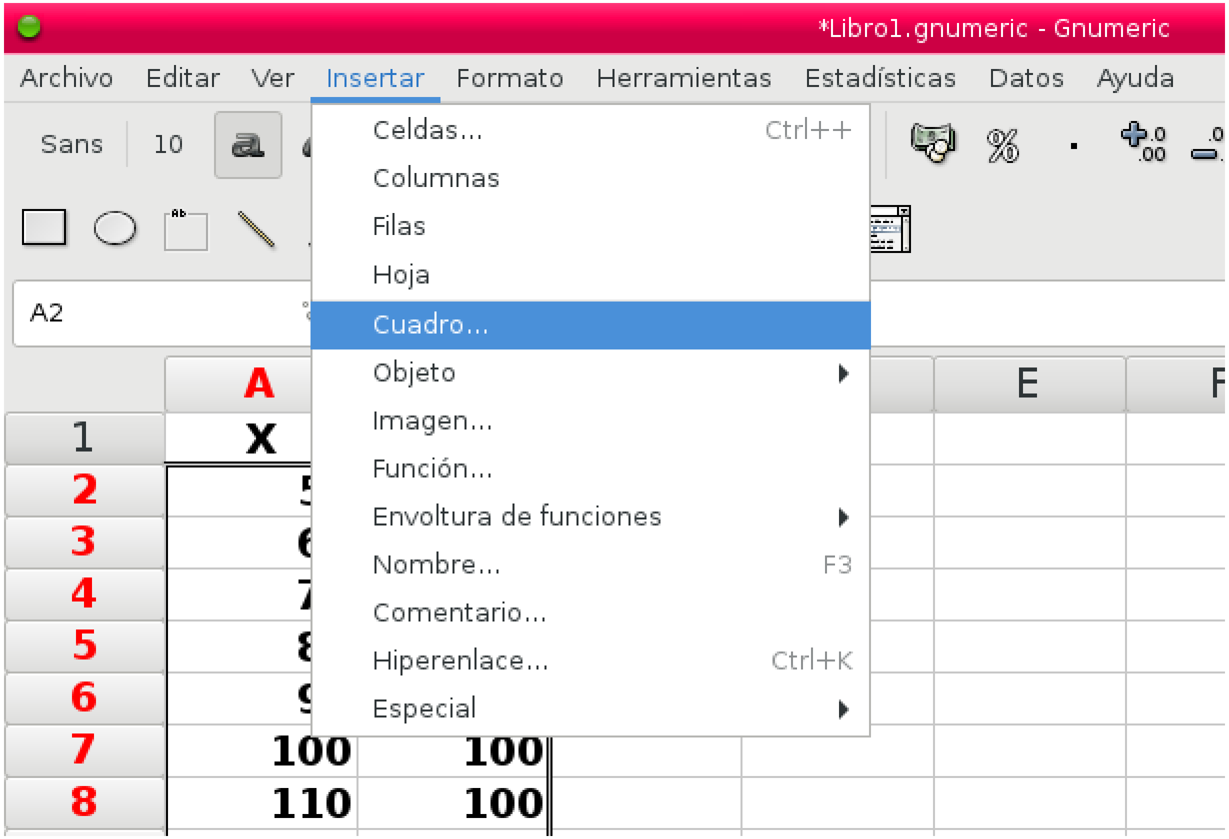
Consigna

a) Encontrar una herramienta para estimar la variable dependiente a partir de la variable independiente.

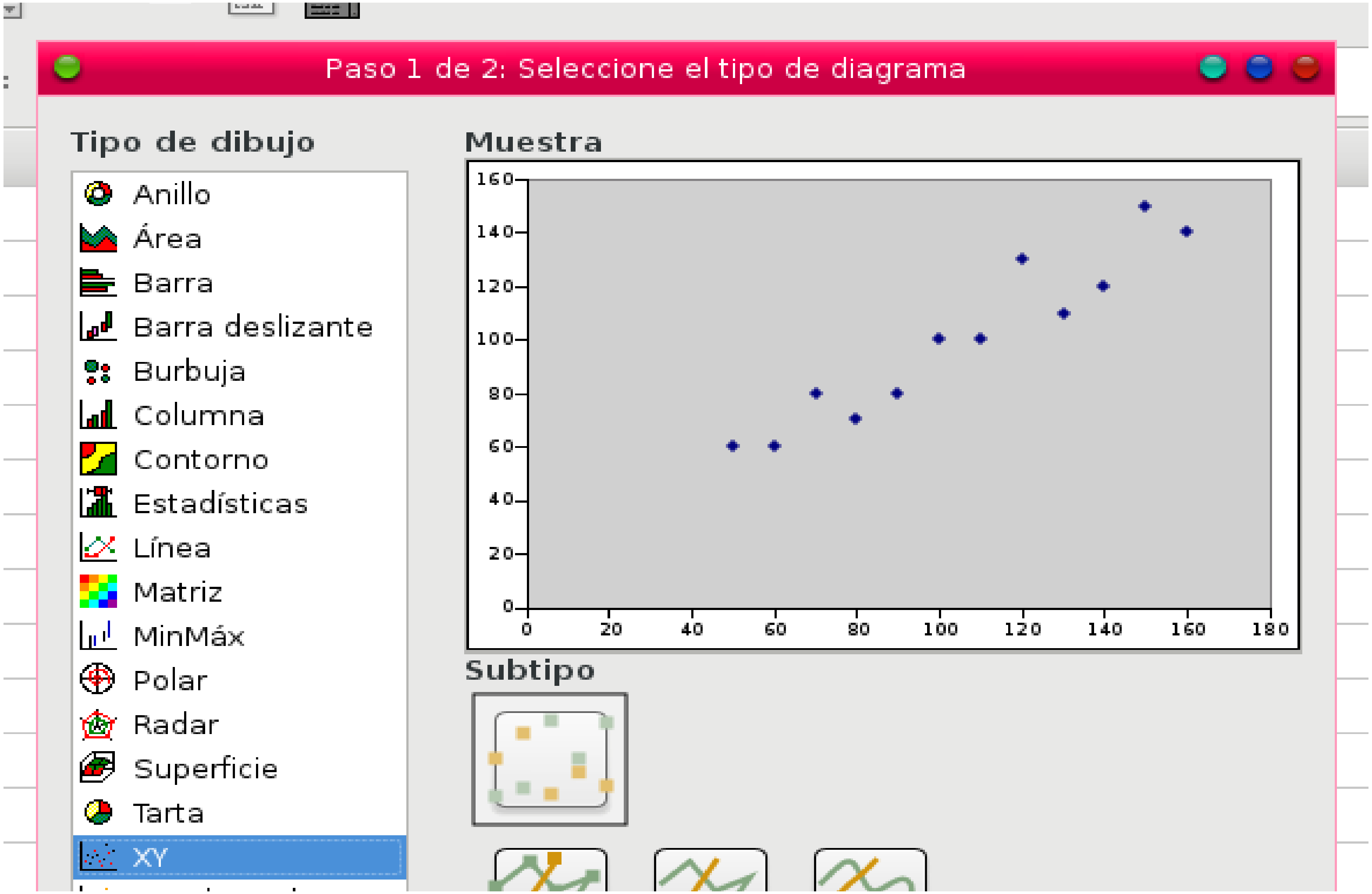
b) ¿Qué opina del ajuste?

c) Estimar los casos dentro de 10, 15 y 20 días.

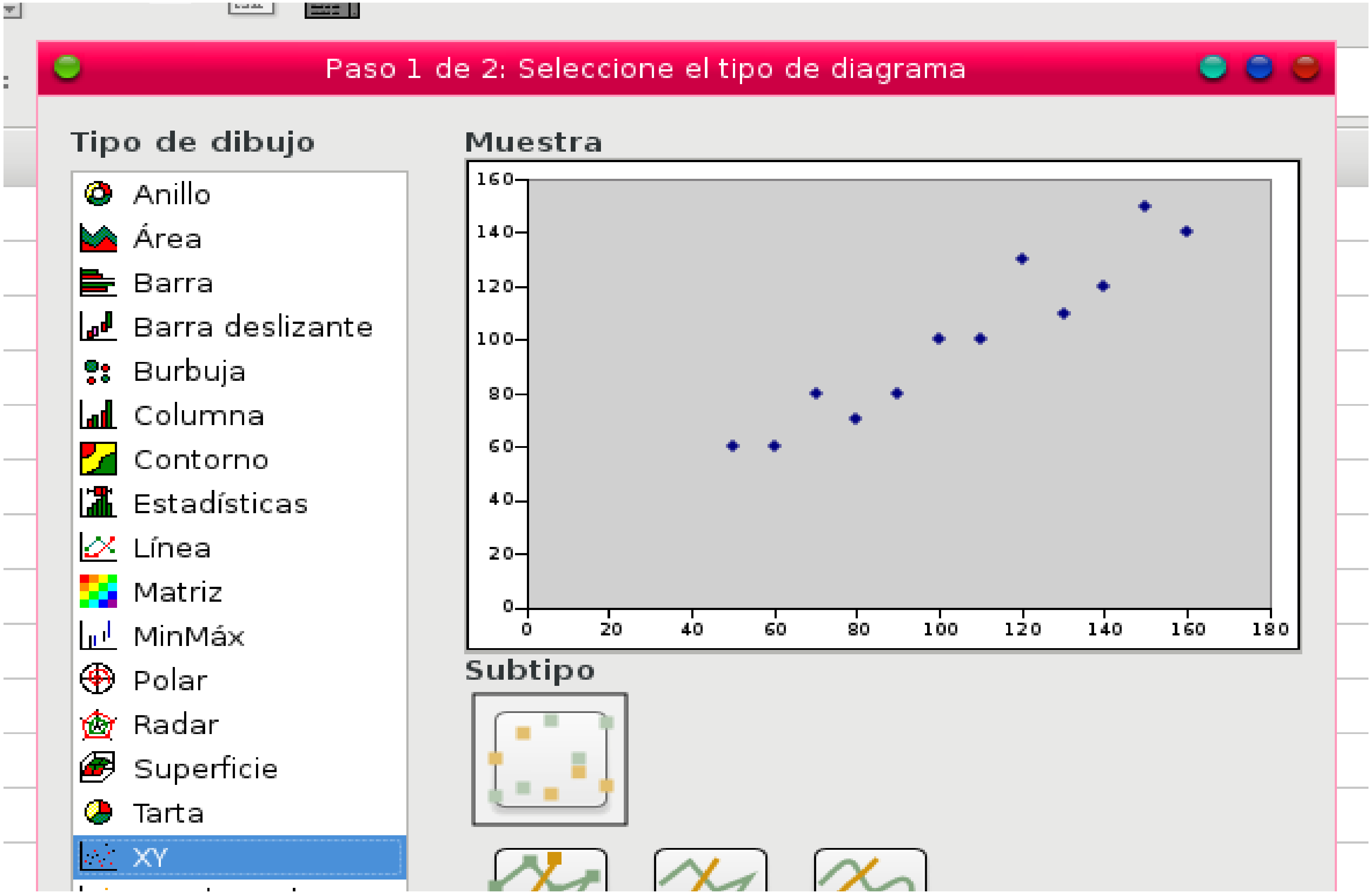
Dispersiograma



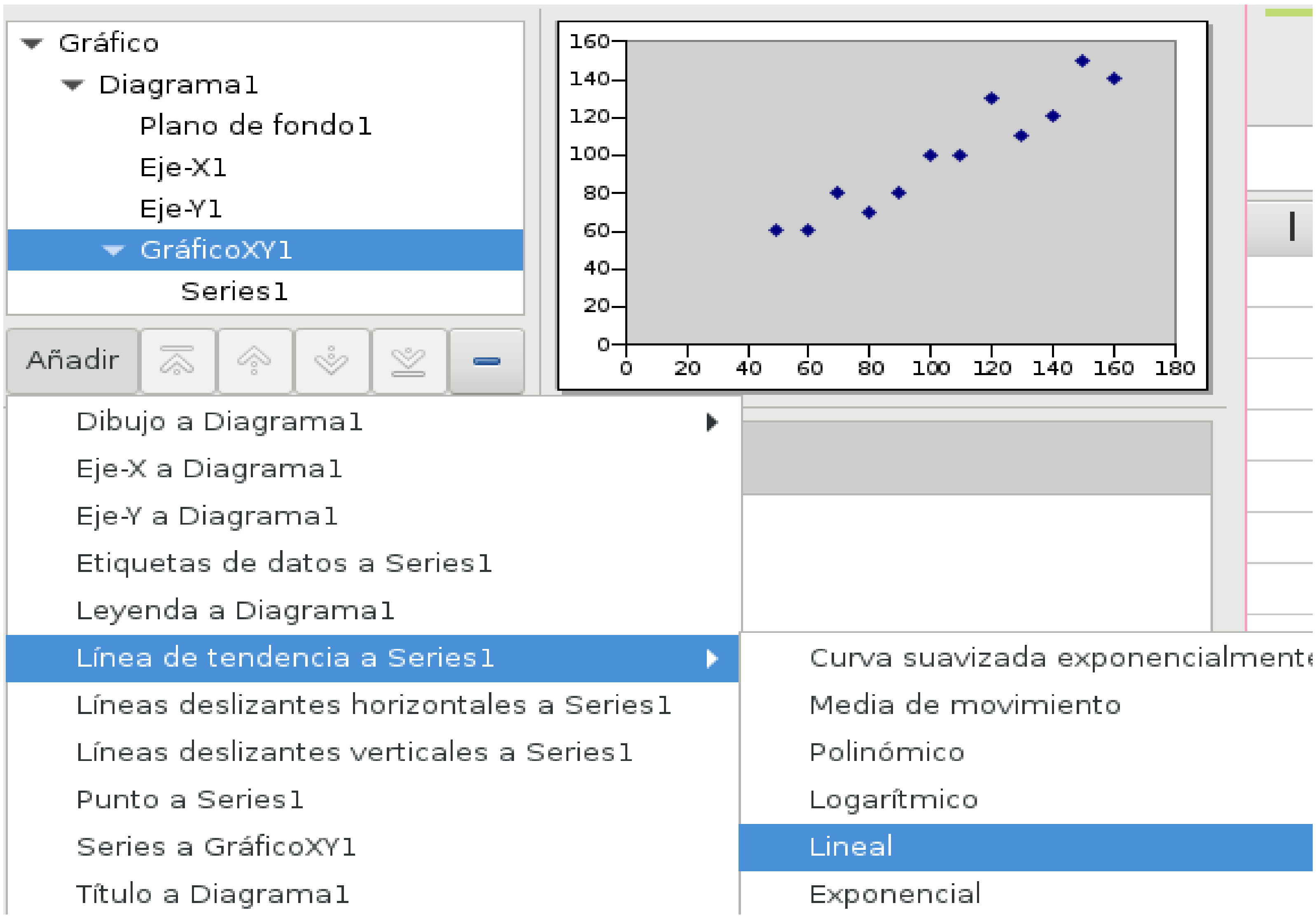
Dispersiograma



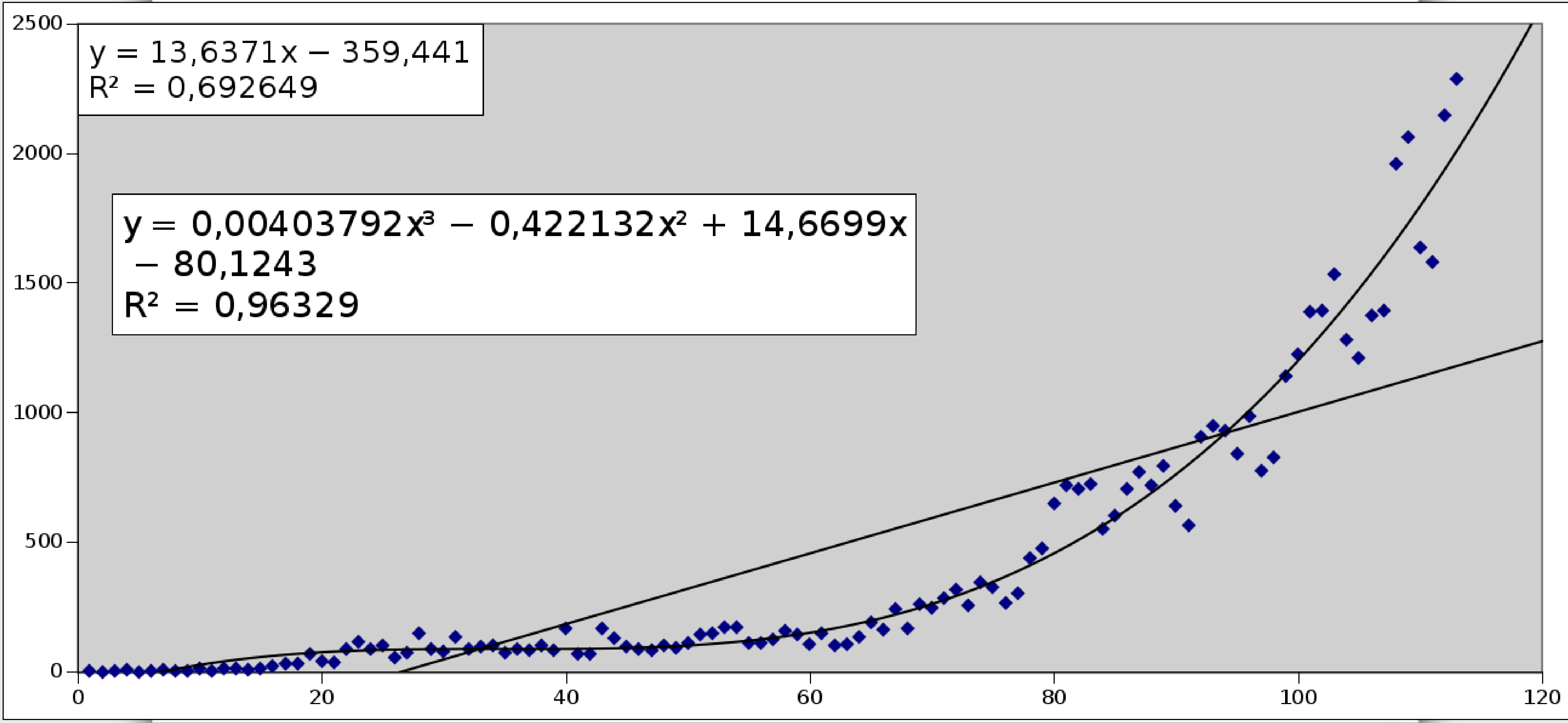
Dispersiograma



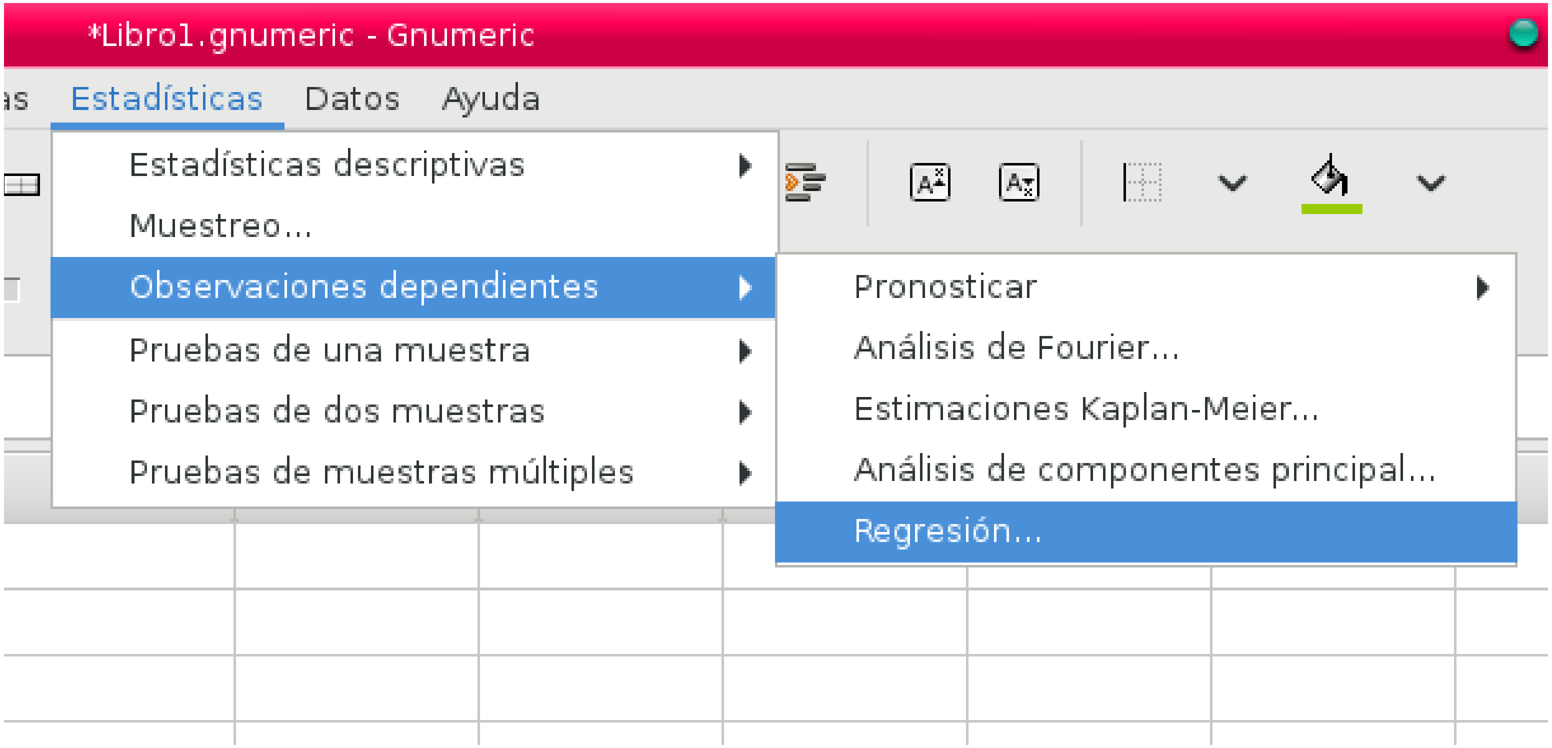
Línea de tendencia



Líneas de tendencia
Lineal y Polinomial Orden 3



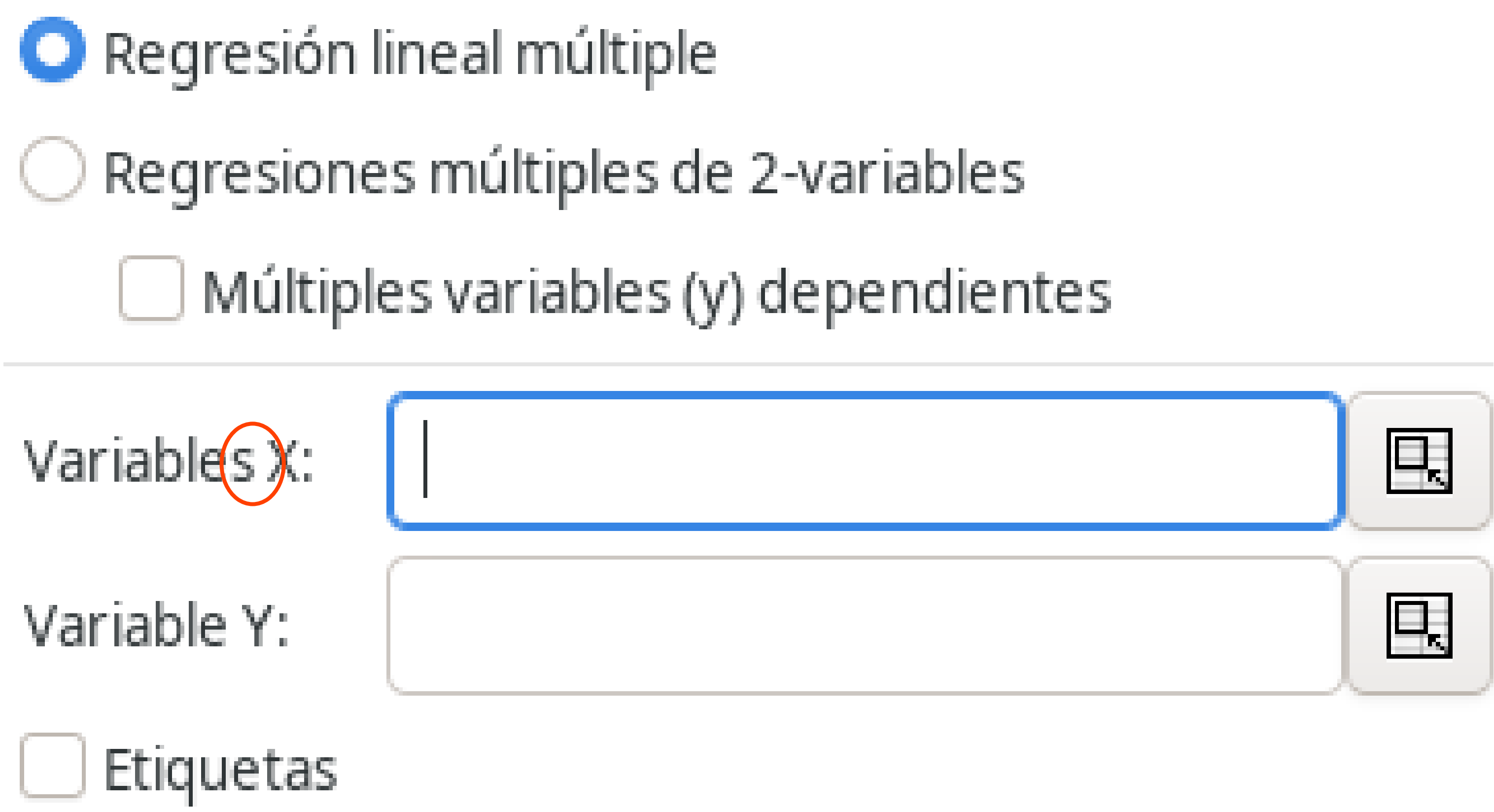
Salida de Regresión



Salida de Regresión

X	X2	X3	Y
1	1	1	1
2	4	8	0
3	9	27	1
4	16	64	6
5	25	125	0
6	36	216	3
7	49	343	5
8	64	512	2

Salida de Regresión



Salida de Regresión

Estadísticas de regresión	
R múltiple	0,981473430091944
R^2	0,9632900939764462
Error estándar	104,26894356955502
R^2 ajustado	0,9622797295904768
Observaciones	113

Análisis de varianza

	df	*SS*		
Regresión	3	31096[illegible]		
Residual	109	1185049,5720491046	10872,0125991105	
Total	112	32281460,265486725		

	Coeficientes	*Error estándar*	*Estadísticas-t*	*Valor-P*	*0,95*	*0,95*
Interceptar	−80,12425878355081	40,574113698941076	−1,9747630072235425	0,05082393549987373	−160,5408330597237	0,29231549262208034
X	14,669876051808165	3,0687461638793385	4,780413650526025	5,504903406814156E-06	8,58772092074577	20,75203118287056
X2	−0,42213162237494506	0,0624024095142403	−6,764668634768248	6,936991070951056E-10	−0,5458111693133919	−0,2984520754364982
X3	0,004037923492669804	0,00035991304140760137	11,219164153868093	6,982911376239668E-20	0,0033245875454677573	0,004751259439871851

	Coeficientes	Error e
Interceptar	−80,1242587835081	4
X	14,669876051808165	3
X2	−0,42213162237494506	0
X3	0,004037923492669804	0,0003

consultas: taulamet@unl.edu.ar

Predicción vs Pronóstico

Predicción: es la estimación del **valor medio de Y** dado un valor particular de X.

$$\hat{Y}_h = a + bX_h$$

Pronóstico: es la estimación **de un sólo valor de Y** dado un valor particular de X.

$$\tilde{Y}_h = a + bX_h$$

Predicción vs Pronóstico

PREDICCIÓN
$$\left(\hat{Y}_h \pm t_{n-2;1-\frac{\alpha}{2}} \hat{\sigma}_{pred}\right)$$ $$\hat{\sigma}_{pred} = \hat{S}_{y/x}\sqrt{\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}$$

$$S^2_{y/x} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-2}$$

Series Cronológicas

- Análisis
  - Representación gráfica
  - Suavizado
  - Aislar y eliminar componentes
  - Realizar predicciones

Suavizado

- Representación gráfica
- Promedio móvil
- Suavizado exponencial

Series Cronológicas

- Componentes
  - Tendencia
  - Estacionalidad
  - Variaciones Cíclicas
  - Variaciones Irregulares

Modelos

$$Y = T * E * CI$$

$$Y = T + E + CI$$

Modelo Multiplicativo

$$Y = T * E * CI$$

Hotel Los Silos

Se dispone de los datos tabulados del número de clientes (en miles) que este hotel ha recibido durante cada estación entre 2010 y 2014.

Los datos

	Verano	Otoño	Invierno	Primavera
2010	300	125	325	200
2011	250	150	375	175
2012	300	200	450	225
2013	350	225	460	249
2014	362	240	500	282

Situaciones a resolver

	Verano	Otoño	Invierno	Primavera
2010	300	125	325	200
2011	250	150	375	175
2012	300	200	450	225
2013	350	225	460	249
2014	362	240	500	282

- **Use las herramientas disponibles para mejorar la apariencia de la gráfica de la serie y haga un pronóstico para el Verano de 2015.**
- Si el hotel contrata 15 personas en el verano, ¿Cuántos empleados necesitará en el invierno, suponiendo iguales requerimientos de servicio?
- Aislar la tendencia y graficar todas las componentes en forma separada.

Suavizado Exponencial

10	W	0,5		
11	Estación	Año	Y	S
12	Verano	2010	300	300
13	Otoño	2010	125	212,5
14	Invierno	2010	325	268,8
15	Primavera	2010	200	234,4
16	Verano	2011	250	242,2
			150	196,1
			375	285,5
			175	230,3
			300	265,1
21	Otoño	2012	200	232,6
22	Invierno	2012	450	341,3
23	Primavera	2012	225	283,1
24	Verano	2013	350	316,6
25	Otoño	2013	225	270,8

$$S_i = W * Y_i + (1 - W) * S_{i-1}$$

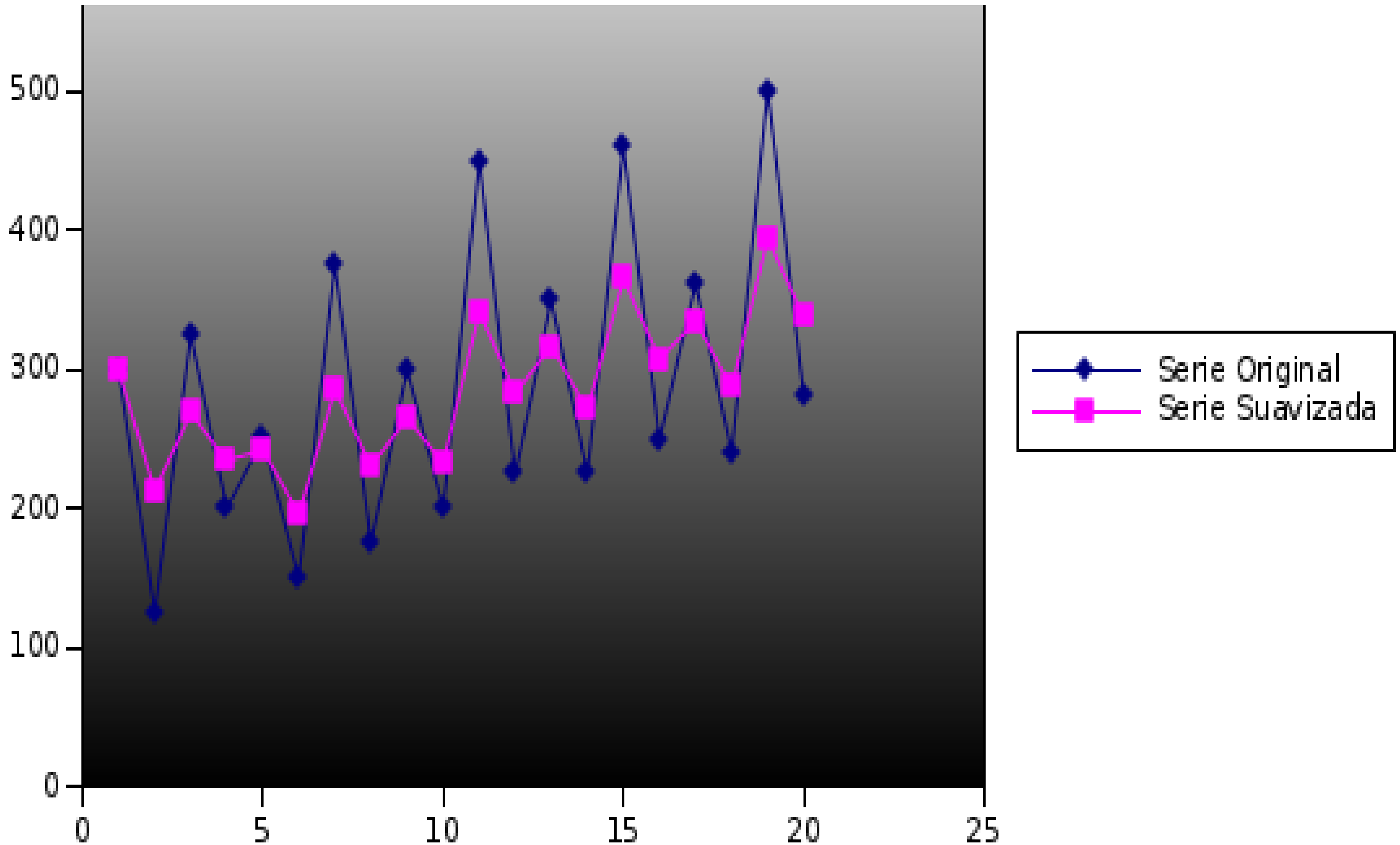
$$S_1 = Y_1$$

Suavizado Exponencial

W	0,5		
Estación	Año	Y	S
Verano	2010	300	300
Otoño	2010	125	212,5
Invierno	2010	325	268,8
Primavera	2010	200	234,4
Verano	2011	250	242,2
Otoño	2011	150	196,1
Invierno	2011	375	285,5
Primavera	2011	175	230,3
Verano	2012	300	265,1
Otoño	2012	200	232,6
Invierno	2012	450	341,3
Primavera	2012	225	283,1
Verano	2013	350	316,6
Otoño	2013	225	270,8
Invierno	2013	460	365,4
Primavera	2013	249	307,2
Verano	2014	362	334,6
Otoño	2014	240	287,3
Invierno	2014	500	393,6
Primavera	2014	282	337,8

Pronóstico para el 1er trimestre de 2015: 337,8

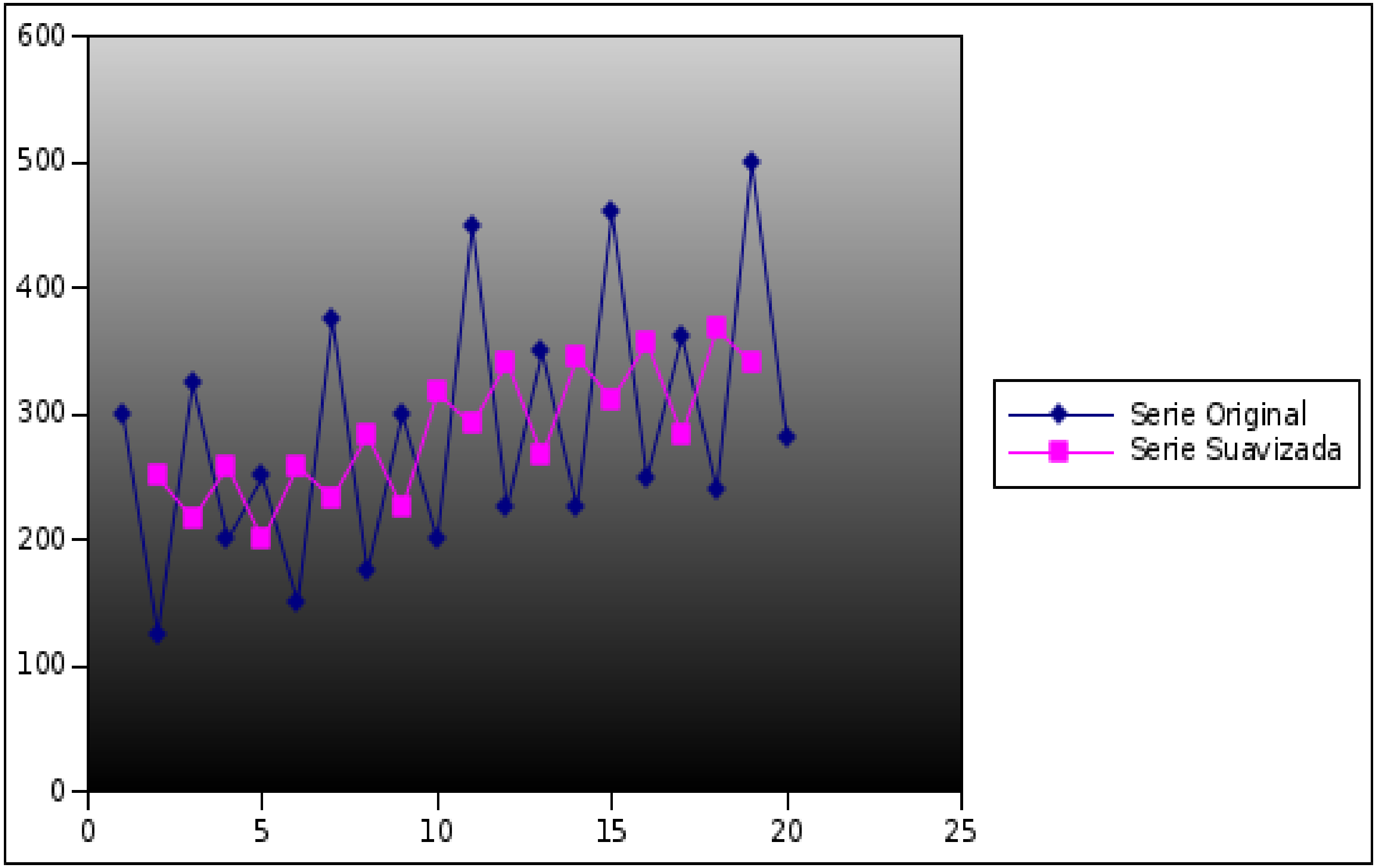
Suavizado Exponencial



Promedio Móvil de Orden 3

	A	B	C	D	E	F
10						
11	Estación	Año	Y	PM3		
12	Verano	2010	300			
13	Otoño	2010	125	=average(C12:C14)		
14	Invierno	2010	325	216,7		
15	Primavera	2010	200	258,3		
16	Verano	2011	250	200		
17	Otoño	2011	150	258,3		
18	Invierno	2011	375	233,3		
19	Primavera	2011	175	283,3		
20	Verano	2012	300	225		
21	Otoño	2012	200	316,7		
22	Invierno	2012	450	291,7		
23	Primavera	2012	225	341,7		
24	Verano	2013	350	266,7		
25	Otoño	2013	225	345		
26	Invierno	2013	460	311,3		

Promedio Móvil de Orden 3



Promedio Móvil de Orden 3

11	Estación	Año	Y	PM3
12	Verano	2010	300	
13	Otoño	2010	125	250
14	Invierno	2010	325	216,7
15	Primavera	2010	200	258,3
16	Verano	2011	250	200
17	Otoño	2011	150	258,3
18	Invierno	2011	375	233,3
19	Primavera	2011	175	283,3
20	Verano	2012	300	225
21	Otoño	2012	200	316,7
22	Invierno	2012	450	291,7
23	Primavera	2012	225	341,7
24	Verano	2013	350	266,7
25	Otoño	2013	225	345
26	Invierno	2013	460	311,3
27	Primavera	2013	249	357
28	Verano	2014	362	283,7
29	Otoño	2014	240	367,3
30	Invierno	2014	500	340,7
31	Primavera	2014	282	
32				

Se perdieron 2 datos (N-1)

Situaciones a resolver

	Verano	Otoño	Invierno	Primavera
2010	300	125	325	200
2011	250	150	375	175
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
2013	350	225	460	249
2014	362	240	500	282

- Use las herramientas disponibles para mejorar la apariencia de la gráfica de la serie y haga un pronóstico para el Verano de 2015.
- **Si el hotel contrata 15 personas en el verano, ¿Cuántos empleados necesitará en el invierno, suponiendo iguales requerimientos de servicio?**
- Aislar la tendencia y graficar todas las componentes en forma separada.

Cálculo del Índice Estacional

Promedio Móvil de Orden 4

Estación	Año	X	Xc	Y	PM4	PM4C	Y
Verano	2010	1	−19	300			
Otoño	2010	2	−17	125			
Invierno	2010	3	−15	325	=average(E8:E1		
Primavera	2010	4	−13	200	225	228,125	
Verano	2011	5	−11	250	231,3	237,5	
Otoño	2011	6	−9	150	243,8	240,625	
Invierno	2011	7	−7	375	237,5	243,75	
Primavera	2011	8	−5	175	250	256,25	
Verano	2012	9	−3	300	262,5	271,875	
Otoño	2012	10	−1	200	281,3	287,5	
Invierno	2012	11	1	450	293,8	300	
Primavera	2012	12	3	225	306,3	309,375	
Verano	2013	13	5	350	312,5	313,75	
Otoño	2013	14	7	225	315	318	
Invierno	2013	15	9	460	321	322,5	

consultas: taulamet@unl.edu.ar

¿Cómo codificar X?

Para calcular la variable X codificada, cargamos los datos y utilizamos la siguiente fórmula:

x_i = Numeramos los períodos

$$x_c = (x_i - \bar{x})*2$$

Obteniendo X_c

CODIF:

2*(C8-average(C$8:C$27))

	A	B	C	D	E
7	**Estación**	**Año**	**X**	**Xc**	**Y**
8	Verano	2010	1	−19	300
9	Otoño	2010	2	−17	125
10	Invierno	2010	3	−15	325
11	Primavera	2010	4	−13	200
12	Verano	2011	5	−11	250
13	Otoño	2011	6	−9	150
14	Invierno	2011	7	−7	375
15	Primavera	2011	8	−5	175
16	Verano	2012	9	−3	300
17	Otoño	2012	10	−1	200
18	Invierno	2012	11	1	450
19	Primavera	2012	12	3	225
20	Verano	2013	13	5	350
21	Otoño	2013	14	7	225

Cálculo del Índice Estacional

Promedio Móvil de Orden 4

CENTRADO

A	B	C	D	E	F	G	H
Estación	Año	X	Xc	Y	PM4	PM4C	Y/PM4C = I
Verano	2010	1	−19	300			
Otoño	2010	2	−17	125			
Invierno	2010	3	−15	325	237,5	=average(F10:F11)	
Primavera	2010	4	−13	200	225	228,125	0,8

consultas: taulamet@unl.edu.ar

Cálculo del Índice Estacional

Y / PMO4C = E

Estación	Año	X	Xc	Y	PM4	PM4C	**Y/PM4C = E**
Verano	2010	1	−19	300			
Otoño	2010	2	−17	125			
Invierno	2010	3	−15	325	237,5	231,25	1,41
Primavera	2010	4	−13	200	225	228,125	0,88
Verano	2011	5	−11	250	231,3	237,5	1,05
Otoño	2011	6	−9	150	243,8	240,625	0,62
Invierno	2011	7	−7	375	237,5	243,75	1,54
Primavera	2011	8	−5	175	250	256,25	0,68
Verano	2012	9	−3	300	262,5	271,875	1,10
Otoño	2012	10	−1	200	281,3	287,5	0,70
Invierno	2012	11	1	450	293,8	300	1,50
Primavera	2012	12	3	225	306,3	309,375	0,73
Verano	2013	13	5	350	312,5	313,75	1,12
Otoño	2013	14	7	225	315	318	0,71
Invierno	2013	15	9	460	321	322,5	1,43
Primavera	2013	16	11	249	324	325,875	0,76
Verano	2014	17	13	362	327,8	332,75	1,09
Otoño	2014	18	15	240	337,8	341,875	0,70
Invierno	2014	19	17	500	346		
Primavera	2014	20	19	282			

Cálculo del Índice Estacional

	2010	2011	2012	2013	2014	Mediana	IE
Verano		1,05	1,10	1,12	1,09	1,10	1,09
Otoño		0,62	0,70	0,71	0,70	0,70	0,70
Invierno	1,41	1,54	1,50	1,43		1,46	1,46
Primavera	0,88	0,68	0,73	0,76		0,75	0,75
					Sum	4,00	4,00
					FC	1,00	

Cálculo del Índice Estacional

	2010	2011	2012	2013	2014	Mediana	IE
Verano		1,05	1,10	1,12	1,09	1,10	1,09
Otoño		0,62	0,70	0,71	0,70	0,70	0,70
Invierno	1,41	1,54	1,50	1,43		1,46	1,46
Primavera	0,88	0,68	0,73	0,76		0,75	0,75
					Sum	4,00	4,00
					FC	1,00	

Verano (1,09)……...15 empleados

Invierno (1,46)……….X

Empleados =15*1,46/1,09
20,09

Situaciones a resolver

	Verano	Otoño	Invierno	Primavera
2010	300	125	325	200
2011	250	150	375	175
2012	300	200	450	225
2013	350	225	460	249
2014	362	240	500	282

- Use las herramientas disponibles para mejorar la apariencia de la gráfica de la serie y haga un pronóstico para el Verano de 2015.
- Si el hotel contrata 15 personas en el verano, ¿Cuántos empleados necesitará en el invierno, suponiendo iguales requerimientos de servicio?
- **Aislar la tendencia y graficar por separado con las variaciones Cíclicas e Irregulares.**

Ubicamos los ÍE

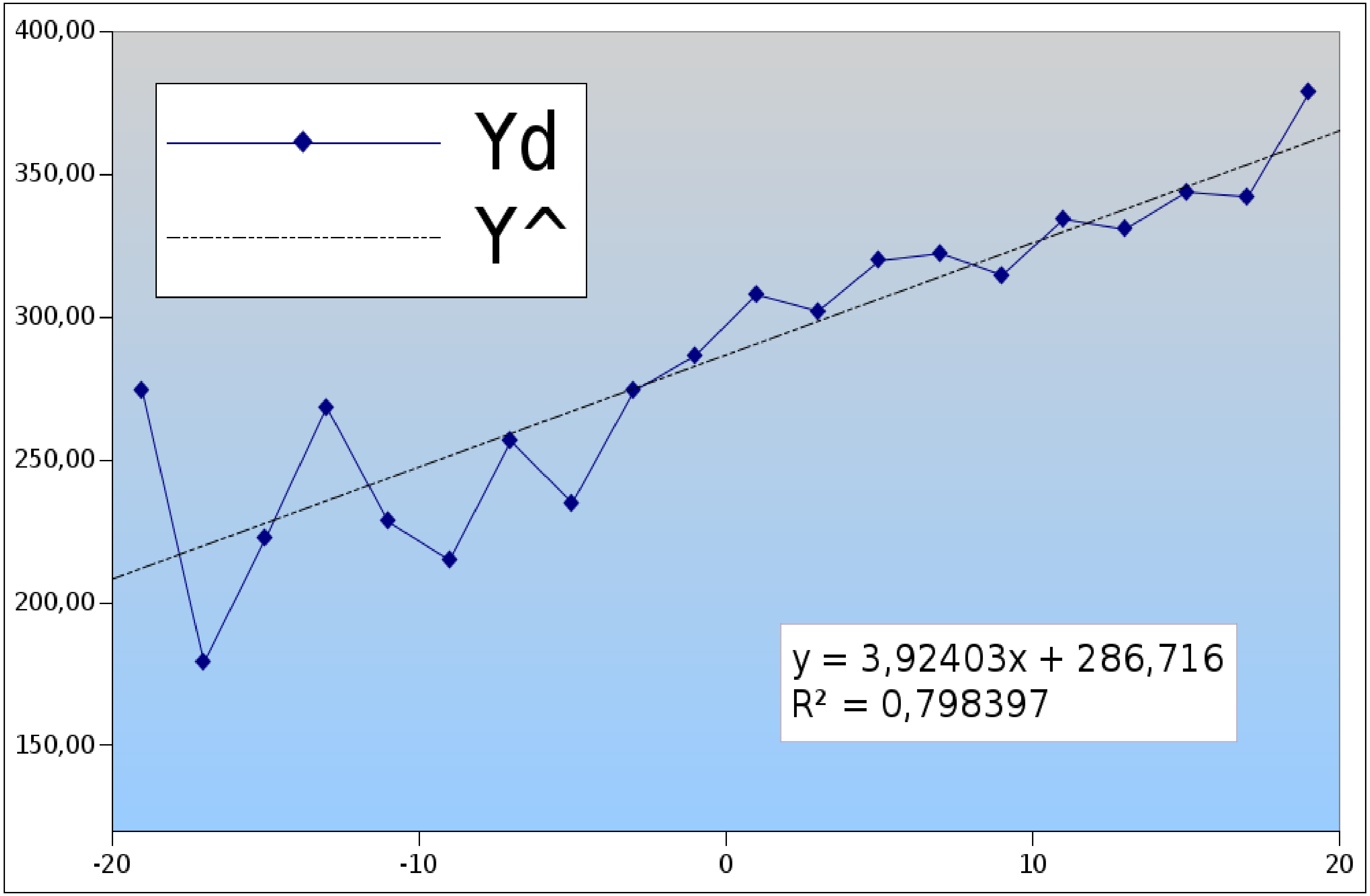
Estación	Año	X	Xc	Y	PM4	PM4C	Y/PM4C = E	I.E.
Verano	2010	1	−19	300				1,09
Otoño	2010	2	−17	125				0,70
Invierno	2010	3	−15	325	237,5	231,25	1,41	1,46
Primavera	2010	4	−13	200	225	228,125	0,88	0,75
Verano	2011	5	−11	250	231,3	237,5	1,05	1,09
Otoño	2011	6	−9	150	243,8	240,625	0,62	0,70
Invierno	2011	7	−7	375	237,5	243,75	1,54	1,46
Primavera	2011	8	−5	175	250	256,25	0,68	0,75
Verano	2012	9	−3	300	262,5	271,875	1,10	1,09
Otoño	2012	10	−1	200	281,3	287,5	0,70	0,70
Invierno	2012	11	1	450	293,8	300	1,50	1,46
Primavera	2012	12	3	225	306,3	309,375	0,73	0,75
Verano	2013	13	5	350	312,5	313,75	1,12	1,09
Otoño	2013	14	7	225	315	318	0,71	0,70
Invierno	2013	15	9	460	321	322,5	1,43	1,46
Primavera	2013	16	11	249	324	325,875	0,76	0,75
Verano	2014	17	13	362	327,8	332,75	1,09	1,09
Otoño	2014	18	15	240	337,8	341,875	0,70	0,70
Invierno	2014	19	17	500	346			1,46
Primavera	2014	20	19	282				0,75

Eliminamos ÍE

$Y_{desestacionalizada}$

Xc	Y	PM4	PM4C	Y/PM4C = E	I.E.	Yd = Y/I.E.
−19	300				1,09	274,03
−17	125				0,70	179,02
−15	325	237,5	231,25	1,41	1,46	222,31
−13	200	225	228,125	0,88	0,75	268,44
−11	250	231,3	237,5	1,05	1,09	228,36
−9	150	243,8	240,625	0,62	0,70	214,82
−7	375	237,5	243,75	1,54	1,46	256,51
−5	175	250	256,25	0,68	0,75	234,88
−3	300	262,5	271,875	1,10	1,09	274,03
−1	200	281,3	287,5	0,70	0,70	286,43
1	450	293,8	300	1,50	1,46	307,81
3	225	306,3	309,375	0,73	0,75	301,99
5	350	312,5	313,75	1,12	1,09	319,71
7	225	315	318	0,71	0,70	322,24
9	460	321	322,5	1,43	1,46	314,65
11	249	324	325,875	0,76	0,75	334,20
13	362	327,8	332,75	1,09	1,09	330,67
15	240	337,8	341,875	0,70	0,70	343,72
17	500	346			1,46	342,01
19	282				0,75	378,49

Estimamos la tendencia a partir de la Y_d



Aislamos la tendencia a partir de la Y_d

Xc	Y	PM4	PM4C	Y/PM4C = E	I.E.	Yd = Y/I.E	Yd^ = T
−19	300				1,09	274,03	=3,92403*D12+286,716
−17	125				0,70	179,02	220,01
−15	325	237,5	231,25	1,41	1,46	222,31	227,86
−13	200	225	228,125	0,88	0,75	268,44	235,70
−11	250	231,3	237,5	1,05	1,09	228,36	243,55
−9	150	243,8	240,625	[illegible]	[illegible]	[illegible]	[illegible]
−7	375	237,5	243,75	[illegible]	[illegible]	[illegible]	[illegible]
−5	175	250	256,25	[illegible]	[illegible]	[illegible]	[illegible]
−3	300	262,5	271,875	[illegible]	[illegible]	[illegible]	[illegible]
−1	200	281,3	287,5	[illegible]	[illegible]	[illegible]	[illegible]
1	450	293,8	300	[illegible]	[illegible]	[illegible]	[illegible]
3	225	306,3	309,375	0,73	0,75	301,99	298,49
5	350	312,5	313,75	1,12	1,09	319,71	306,34
7	225	315	318	0,71	0,70	322,24	314,18
9	460	321	322,5	1,43	1,46	314,65	322,03
11	249	324	325,875	0,76	0,75	334,20	329,88
13	362	327,8	332,75	1,09	1,09	330,67	337,73
15	240	337,8	341,875	0,70	0,70	343,72	345,58
17	500	346			1,46	342,01	353,42

$y = 3{,}92403x + 286{,}716$
$R^2 = 0{,}798397$

Eliminamos la tendencia obteniendo CI

I.E.	Yd = Y/I.E.	Yd^ = T	Yd/T=C*I
1,09	274,03	212,16	1,29
0,70	179,02	220,01	0,81
1,46	222,31	227,86	0,98
0,75	268,44	235,70	1,14
1,09	228,36	243,55	0,94
0,70	214,82	251,40	0,85
1,46	256,51	259,25	0,99
0,75	234,88	267,10	0,88
1,09	274,03	274,94	1,00
0,70	286,43	282,79	1,01
1,46	307,81	290,64	1,06
0,75	301,99	298,49	1,01
1,09	319,71	306,34	1,04
0,70	322,24	314,18	1,03
1,46	314,65	322,03	0,98
0,75	334,20	329,88	1,01
1,09	330,67	337,73	0,98
0,70	343,72	345,58	0,99
1,46	342,01	353,42	0,97
0,75	378,49	361,27	1,05

Graficamos CI