Aclaración sobre los valores que aparecen en GeoGebra en el Análisis de datos de un Análisis de Regresión de dos variables

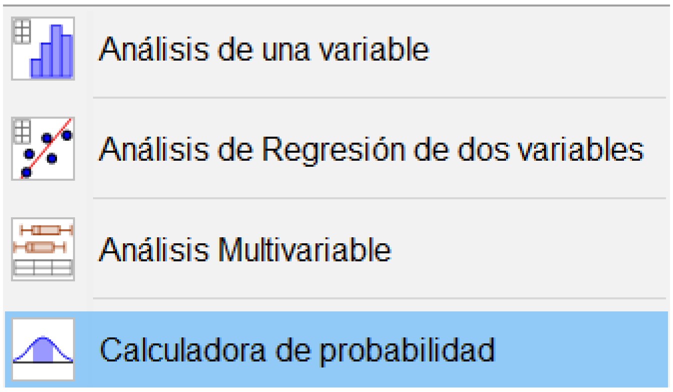
Cuando se efectúa con GeoGebra un Análisis de regresión y se piden datos en el apartado **Estadísticas**, aparecen unos valores que pueden crear confusión comparados con otros valores más conocidos.

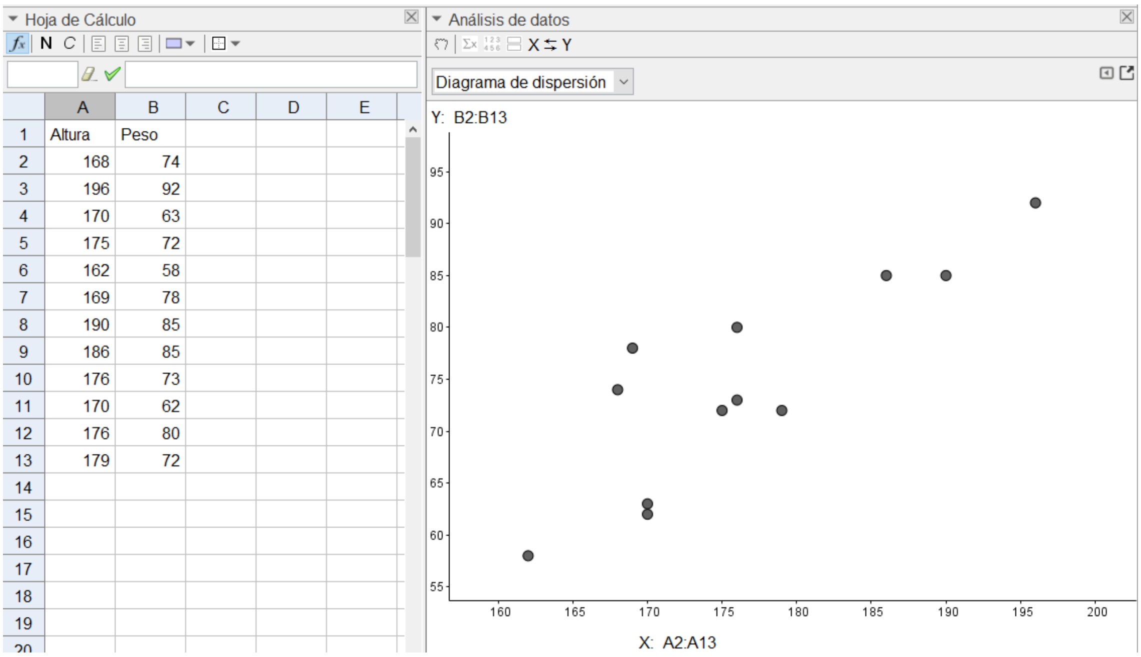
Considérese, para ilustrarlo, una actividad sencilla como la que sigue:

La siguiente tabla incluye información acerca del peso (en kg.) y la altura (en cm.) de 12 individuos:

Altura	168	196	170	175	162	169	190	186	176	170	176	179
Peso	74	92	63	72	58	78	85	85	73	62	80	72

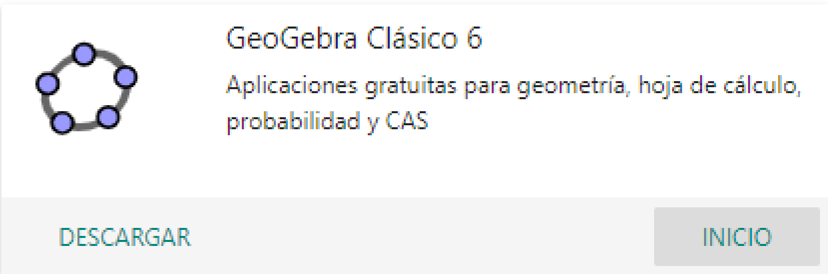
En GeoGebra se introducen los datos en la Hoja de cálculo, se seleccionan las dos columnas y se elige **Análisis de Regresión de dos variables**.





A continuación, se elige  (**Mostrar estadísticas**) en el grupo de herramientas:

En GeoGebra 5 y GeoGebra Clásico 6 (versión "DESCARGAR" de escritorio)



se obtiene:

Estadísticas	
MediaX	176.4167
MediaY	74.5
Sx	9.9312
Sy	10.2025
r	0.8537
ρ	0.6972
Sxx	1084.9167
Syy	1145
Sxy	951.5

En la versión "INICIO" de GeoGebra Clásico 6 se obtiene:

MediaX	176.4167
MediaY	74.5
Sx	9.9312
Sy	10.2025
r	0.8537
ρ	0.6972
nVarX	1084.9167
nVarY	1145
nCov	951.5

Se observa que solo cambia el nombre de los tres últimos valores. Se aclarará más adelante.

Para hallar los distintos valores estadísticos y compararlos con los datos obtenidos se van a creer tres listas:

- altura = {168, 196, 170, 175, 162, 169, 190, 186, 176, 170, 176, 179}
- conjunta = {(168, 74), (196, 92), (170, 63), (175, 72), (162, 58), (169, 78), (190, 85), (186, 85), (176, 73), (170, 62), (176, 80), (179, 72)}
- peso = {74, 92, 63, 72, 58, 78, 85, 85, 73, 62, 80, 72}

Al calcular, en la Vista Algebraica, las medias de las alturas (MediaX) y de los pesos (MediaY) se obtiene:

MediaX = 176.4167
MediaY = 74.5

que corresponde, como era de esperar, con los valores que aparecen en **Mostrar estadísticas**

MediaX = 176.4167
MediaY = 74.5

MediaX	176.4167
MediaY	74.5

Al calcular, en la Vista Algebraica, las varianzas de las alturas (VarianzaX) y de los pesos (VarianzaY) se obtiene:

VarianzaX = 90.4097

VarianzaY = 95.4167

Esos valores no coinciden con ningún valor de los que aparecen en **Mostrar estadísticas**.

Pero recuérdese que:

$$var(x) = \frac{\sum_1^n (x_i - \bar{x})^2}{n} \qquad var(y) = \frac{\sum_1^n (y_i - \bar{y})^2}{n}$$

Al calcular los numeradores de las expresiones anteriores se obtiene:

Hoja de Cálculo

C20

	A	B	C	D
1	Altura	Peso	$(x_i - \text{MediaX})^2$	$(y_i - \text{MediaY})^2$
2	168	74	70.8403	0.25
3	196	92	383.5069	306.25
4	170	63	41.1736	132.25
5	175	72	2.0069	6.25
6	162	58	207.8403	272.25
7	169	78	55.0069	12.25
8	190	85	184.5069	110.25
9	186	85	91.8403	110.25
10	176	73	0.1736	2.25
11	170	62	41.1736	156.25
12	176	80	0.1736	30.25
13	179	72	6.6736	6.25
14				
15			1084.9167	1145
16				

Obsérvese que precisamente son los valores que aparecen en **Mostrar estadísticas:**

Sxx=nVarX=1084.9167 Syy=nVarY=1145

Por tanto, se puede decir que Sxx es n veces la varianza de la variable x; Syy es n veces la varianza de la variable y.

Con lo cual:

$$var(x) = \frac{Sxx}{n} = \frac{nVarX}{n} = \frac{1084.9167}{12} = 90.4097$$

$$var(y) = \frac{Syy}{n} = \frac{nVarY}{n} = \frac{1145}{12} = 95.4167$$

Por otra parte, para calcular la desviación típica:

$$\sigma_x = \sqrt{var(x)} = \sqrt{90.4097} = 9.5084$$

$$\sigma_y = \sqrt{var(y)} = \sqrt{95.4167} = 9.7681$$

Al calcular, en la Vista Algebraica, las desviaciones de las alturas (DesviacionX) y de los pesos (DesviacionY) se obtienen los mismos valores anteriores:

DesviacionX = 9.5084
DesviacionY = 9.7681

Esos valores no coinciden con ningún valor de los que aparecen en **Mostrar estadísticas**.

Entonces, para relacionar las desviaciones halladas anteriormente con algunos valores de **Mostrar estadísticas** recuérdese que cuando los datos se consideran extraídos de una muestra en vez de tomados como todos los datos de la población es conveniente usar la cuasivarianza y la cuasidesviación típica que se obtienen como:

$$S^2x = \sigma_{n-1}^2\, x = \frac{\sum_1^n (x_i - \overline{x})^2}{n-1} = \frac{n}{n-1}\sigma_x^2$$

$$S^2y = \sigma_{n-1}^2\, y = \frac{\sum_1^n (y_i - \overline{y})^2}{n-1} = \frac{n}{n-1}\sigma_y^2$$

$$Sx = \sigma_{n-1}\, x = \sqrt{\frac{n}{n-1}}\, \sigma x$$

$$Sy = \sigma_{n-1}\, x = \sqrt{\frac{n}{n-1}}\, \sigma y$$

Así: $\sqrt{\frac{12}{11}} * 9.5084 = 9.9312 = Sx$ $\qquad$ $\sqrt{\frac{12}{11}} * 9.7681 = 10.2025 = Sy$

que corresponden con los valores que aparecen en **Mostrar estadísticas.**

Con lo cual:

$\sigma x = \sqrt{\frac{n-1}{n}}\, Sx = \sqrt{\frac{11}{12}} * \ 9.9312 = 9.5084$ | **DesviacionX = 9.5084**

$\sigma y = \sqrt{\frac{n-1}{n}}\, Sy = \sqrt{\frac{11}{12}} * 10.2025 = 9.7681$ **DesviacionY = 9.7681**

Al calcular, en la Vista Algebraica, la covarianza de las alturas y de los pesos mediante **Covarianza(conjunta)** se obtiene:

CovarianzaXY = 79.2917

Este valor no coincide con ningún valor de los que aparecen en **Mostrar estadísticas**.

Pero recuérdese que:

$$cov(x,y) = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{n}$$

Al calcular el numerador de la expresión anterior se obtiene:

E
$(x_i - \text{MediaX})(y_i - \text{MediaY})$
4.2083
342.7083
73.7917
3.5417
237.875
-25.9583
142.625
100.625
0.625
80.2083
-2.2917
-6.4583
951.5

Obsérvese que precisamente es el valor que aparece en **Mostrar estadísticas:**

Sxy=nCov= 951.5

Por tanto, se puede decir que Sxy es n veces la covarianza de las variables x e y.

Con lo cual:

$$cov(x,y) = \frac{Sxy}{n} = \frac{nCov}{n} = \frac{951.5}{12} = 79.2917$$

Al calcular, en la Vista Algebraica, el coeficiente de correlación de Pearson entre las variables altura y peso mediante **CoeficienteCorrelación(conjunta)** se obtiene:

r = 0.8537

que corresponde con el valor que aparece en **Mostrar estadísticas**.

También se podría haber calculado con los otros valores que aparecen en **Mostrar estadísticas:**

$$r = \frac{cov(x,y)}{\sigma x\ \sigma y} = \frac{\frac{Sxy}{n}}{\sqrt{\frac{n-1}{n}}\ Sx\ \sqrt{\frac{n-1}{n}}\ Sy} = \frac{Sxy}{n\frac{n-1}{n}Sx\ Sy} = \frac{Sxy}{(n-1)Sx\ Sy} = \frac{951.5}{11 * 9.9312 * 10.2025} = 0.8537$$

Por último, al calcular en la Vista Algebraica, el coeficiente de correlación de Spearman (Rho o ro, ρ, que es recomendable utilizarlo cuando los datos presentan valores extremos, ya que dichos valores afectan mucho al coeficiente de correlación de Pearson, o ante distribuciones no normales) mediante **Spearman(conjunta)** se obtiene:

ρ = 0.6972

que corresponde con el valor que aparece en **Mostrar estadísticas**.

Como resumen, la siguiente tabla muestra cómo usar los valores que aparecen en **Mostrar estadísticas** para hallar los parámetros estadísticos más usuales en una distribución bidimensional.

MediaX	MediaX	$\bar{x}$
MediaY	MediaY	$\bar{y}$
Sx	Sx	$\sigma x = \sqrt{\frac{n-1}{n}}\ Sx$
Sy	Sy	$\sigma y = \sqrt{\frac{n-1}{n}}\ Sy$
r	r	Pearson
ρ	ρ	Spearman
Sxx	nVarX	$var(x) = \frac{Sxx}{n} = \frac{nVarX}{n}$
Syy	nVarY	$var(y) = \frac{Syy}{n} = \frac{nVarY}{n}$
Sxy	nCov	$cov(x,y) = \frac{Sxy}{n} = \frac{nCov}{n}$