ESTADÍSTICA DE LA OBSERVACIÓN

OBJETIVOS DE APRENDIZAJE

Declarar lo que es la muestra de una población
Reconocer una distribución normal
Calcular la mediana, la media y la moda
Calcular la varianza
Utilizar el Método de Mínimos Cuadrados para un ajuste lineal

La Estadística está formada por el conjunto de métodos y técnicas que permiten la obtención, organización, síntesis, descripción e interpretación de los datos para la toma de decisiones en ambiente de incertidumbre.

Diferencia entre muestras y población

Al conjunto que contiene a todos los elementos de iguales características y propiedades se denomina población, por ejemplo; la población de estudiantes registrados en el laboratorio de física A es 320, mientras que la muestra es una fracción de la población, por ejemplo; para determinar la edad promedio de los estudiantes que toman el laboratorio de física A se escoge una muestra de 60. En estadística se conoce como muestreo a la técnica para la selección de una muestra a partir de una población. Recuerde los elementos de la muestra no guardan ninguna característica especial que los diferencie de los demás elementos de una población. Al contrario, con una muestra lo que se pretende es representar a toda la población. Podríamos decir que la muestra es una población de tamaño reducido. Esto es una desventaja, aunque la muestra pretenda representar lo más fielmente posible a la población, nunca dejará de ser eso, una muestra. Con los datos de la muestra solo podremos conocer las características de esos valores muestrales, para ello se obtendrán los valores estimados de los parámetros de la población, como a estos últimos casi nunca se los podrá conocer con certeza, entonces a esta diferencia se conoce como error muestral.

Variables y Atributos

Observar una población es equivalente a observar sus elementos. Ahora bien, esos elementos poseen una serie de características que son las que realmente se observan todas estas características de los elementos de una población se les conoce de forma genérica como caracteres. Estos últimos, según su naturaleza, pueden ser de tipo cuantitativo o cualitativo.

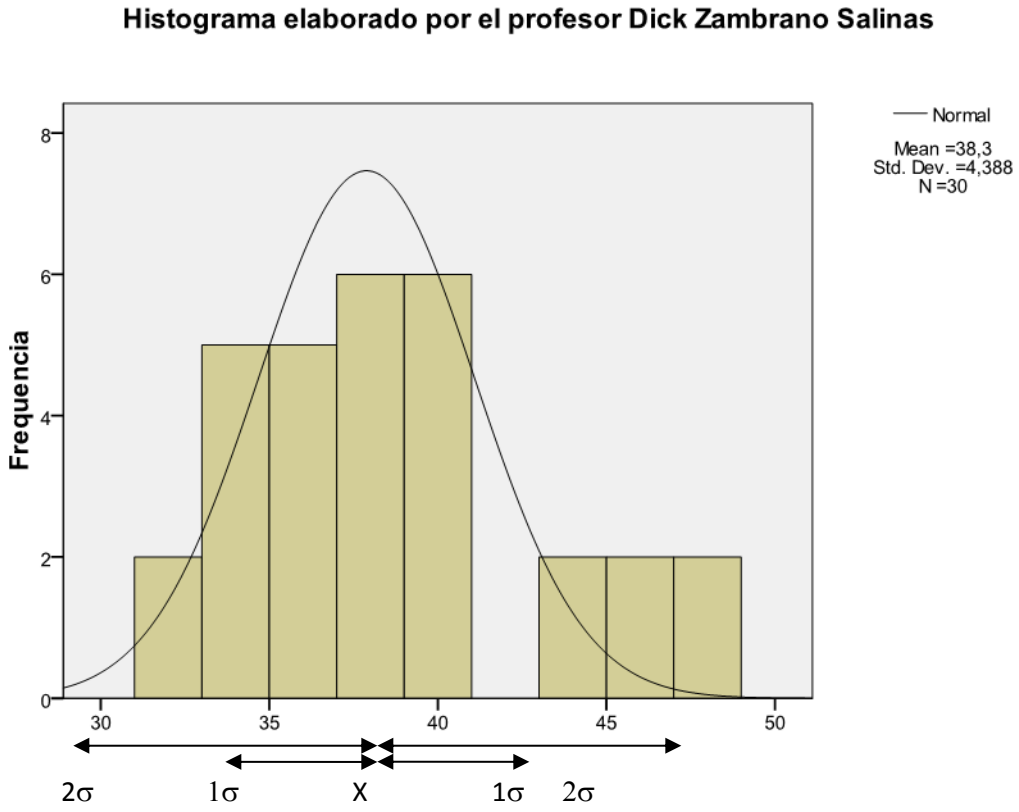
En estadística es más habitual hablar de variables que de caracteres cuantitativos y de atributos en lugar de caracteres cualitativos. Las variables pueden medirse en términos cuantitativos y a cada una de esas posibles mediciones o realizaciones se les conoce como valores, datos u observaciones. A su vez, en función del número posible de valores que tome una variable, a las mismas se las puede clasificar en discretas y continuas. Serán discretas cuando el número de valores sea finito o infinito numerable, ejemplo, el número de alumnos registrados en física A en la ESPOL, mientras que una variable será continua cuando el número de sus valores sea infinito no numerable, por ejemplo, el tiempo en caída libre de un objeto desde una altura determinada.

Distribución Normal o de Gauss

La distribución normal frecuentemente es identificada en la mayoría de las mediciones físicas y por tal motivo se le confiere especial importancia, ésta distribución puede deducirse a partir de la hipótesis de que la desviación total de una cantidad medida x, respecto de un valor central X, es la resultante de una gran cantidad de pequeñas fluctuaciones que ocurren al azar, la curva que representa esta distribución tiene la forma de una campana como se muestra en la gráfica en donde se relaciona la distribución de Gauss y las mediciones reales [5]. La función matemática que representa a la campana de Gauss viene dada por $y = Ce^{-h^2(x-X)^2}$

Donde C es una medida de la altura de la campana, observe que la curva es simétrica alrededor de la media (X = 38.3) y tiende a cero asintóticamente. Mientras que h determina la amplitud de la curva, si h es grande, la campana es angosta y alta, si h es pequeña, la campana es ancha y baja, la relación entre h y la desviación estándar viene dada por $\sigma = \frac{1}{\sqrt{2}h}$

Ahora daremos una interpretación a la desviación estándar en términos de probabilidad, así tenemos que el área bajo la curva que se extiende desde menos infinito a más infinito es igual a 1. Para una desviación alrededor de X $(X \pm \sigma)$ el área bajo la curva corresponde al 68% del total, esto quiere decir que, 68 de cada 100 muestras tomadas de una población, dan la confianza de que cada una de ellas contenga la mejor estimación de la media, esto se conoce como un intervalo de confianza. Mientras que dentro del intervalo de confianza $(X \pm 2\sigma)$ la probabilidad es del 95%



La **media** o promedio de la distribución se define, como: $X = \bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$

X es la media aritmética de los valores observados.

La **moda** corresponde al valor de la variable donde está la máxima frecuencia, o sea, que en un histograma la moda corresponde al valor de la variable donde hay un pico o máximo. Si una distribución tiene dos máximos la denominamos distribución bimodal, y si tiene tres máximos trimodal y así sucesivamente.

La **mediana** es el valor de la variable que separa los datos entre aquellos que definen el primero 50% de los valores de los de la segunda mitad. O sea que la mitad de los datos de la población o muestra están arriba de la mediana y la otra mitad están abajo de la misma.

Para estimar la mediana tenemos que observar la lista de datos ordenados de menor a mayor, y ubicar el valor central de la lista. Si el número de datos es impar, la mediana corresponde precisamente al valor central. Si el numero N de datos es par, la mediana se estima como $Mediana = \frac{X_{\frac{N}{2}} + X_{\frac{N}{2}+1}}{2}$

En una distribución dada, una línea vertical trazada desde la mediana divide a la distribución en dos partes de área equivalentes.

Varianza

La varianza es la media aritmética del cuadrado de las desviaciones respecto a la media de una distribución estadística. La varianza de la población se representa por σ^2 y de la muestra por S^2.

$$S^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots\ldots + (x_N - \bar{x})^2}{N} \quad \Rightarrow \quad S^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}$$

Desviación estándar

La desviación estándar (σ) mide cuánto se separan los datos. La fórmula es la raíz cuadrada de la varianza.

Expresión de la desviación estándar de la muestra:

$$\boldsymbol{S = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N-1}}}$$

El error estándar de la muestra o también conocido como error cuadrático medio (SEM) se calcula así:

$$\boldsymbol{S_m = \frac{S}{\sqrt{N}}}$$

Tratamiento de datos para distribuciones bidimensionales

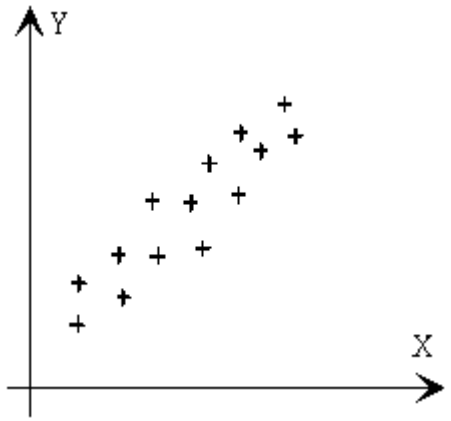
Supongamos que hemos medido un conjunto de pares de datos (x_i, y_i) en una experiencia, por ejemplo, la posición de un móvil en ciertos instantes de tiempo. Queremos obtener una función y = f(x) que se ajuste lo mejor posible a los valores experimentales. Se pueden ensayar muchas funciones, rectas, polinomios, funciones potenciales o logarítmicas. Una vez establecido la función a ajustar se determina sus parámetros. La función más sencilla es la función lineal y = ax + b. El procedimiento de ajustar los datos experimentales a una línea recta se denomina Regresión Lineal.

Regresión Lineal

La importancia de las distribuciones bidimensionales radica en investigar cómo influye una variable sobre la otra. Esta puede ser una dependencia causa efecto, por ejemplo, a mayor altura de caída (causa), mayor es la rapidez de impacto con el suelo (efecto). O bien, el aumento de la masa de un sistema sometido a una fuerza constante, da lugar a una disminución de la aceleración del mismo.

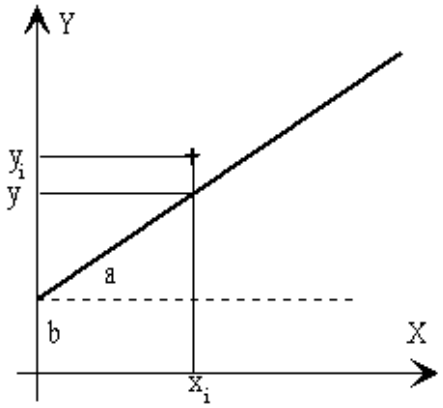
Si utilizamos un sistema de coordenadas cartesianas para representar la distribución bidimensional, obtendremos un conjunto de puntos conocido como el diagrama de dispersión, cuyo análisis permite estudiar cualitativamente, la relación entre ambas variables tal como se ve en la figura. El siguiente paso, es la determinación de la dependencia funcional entre las dos variables x e y que mejor ajusta a la distribución bidimensional. Se denomina regresión lineal cuando la función es lineal, es decir, requiere la determinación de dos parámetros: la pendiente y la ordenada en el origen de la recta de regresión, y = b + ax

La regresión nos permite además, determinar el grado de dependencia de las series de valores X e Y, prediciendo el valor Y estimado que se obtendría para un valor X que no esté en la distribución.



Vamos a determinar la ecuación de la recta que mejor ajusta a los datos representados en la figura. Se denomina error $(e_i=y_i-y)$ a la diferencia, entre el valor medido y_i, y el valor ajustado $y= ax_i+b$, tal como se ve en la figura inferior. El criterio de ajuste se toma como aquél en el que la desviación cuadrática media sea mínima, es decir, la suma debe de ser mínima.

$$Error\ de\ Ajuste = \sum_1^N e_i^2 = \sum_1^N [y_i - (ax_i + b)]^2$$



Los extremos de una función: máximo o mínimo se obtienen cuando las derivadas de s respecto de a y de b sean nulas. Lo que da lugar a un sistema de dos ecuaciones con dos incógnitas del que se despeja a y b.

$$a = \frac{N\sum x_i y_i + \sum x_i \sum y_i}{N\sum x_i^2 - (\sum x_i)^2} \quad (1) \qquad b = \frac{\sum y_i - a\sum x_i}{N} \quad (2)$$

El coeficiente de correlación es otra técnica de estudiar la distribución bidimensional [6], que nos indica la intensidad o grado de dependencia entre las variables X e Y. El coeficiente de correlación r es un número que se obtiene mediante la fórmula.

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{N\sigma_x\sigma_y}$$

El numerador es el producto de las desviaciones de los valores X e Y respecto de sus valores medios. En el denominador tenemos las desviaciones cuadráticas medias de X y Y

El coeficiente de correlación puede valer cualquier número comprendido entre -1 y +1.

· Cuando r = 1, la correlación lineal es perfecta, directa.

· Cuando r = -1, la correlación lineal es perfecta, inversa.

· Cuando r = 0, no existe correlación alguna, independencia total de los valores X e Y

Variantes de la regresión lineal

La función potencial

Supongamos que en un experimento de caída libre, la distancia h recorrida por el cuerpo y el tiempo t de caída se registraron como se indica en la tabla 1. Por otra parte, dado que es conocido el modelo matemático que relaciona las variables, estas pueden ser representadas por una función potencial, $h = ct^a$ la cual puede transformarse en $logh = alogt + logc$ Si usamos las nuevas variables Y=log h y X=log t, obtenemos la relación lineal. $Y=aX+b$ donde b=log c. Obtener la función matemática que relacione las variables h y t

h(cm)	92	84	73	66	57	45	36	24
t(s)	0.438	0.418	0.390	0.371	0.345	0.306	0.274	0.224

Tabla 1

Para lograr esto debemos primero completar la siguiente tabla.

Y= logh (cm)	1.96	1.92	1.86	1.82	1.76	1.65	1.56	1.38
X= logt (s)	- 0.356	- 0.379	- 0.409	- 0.431	- 0.462	- 0.514	- 0.562	- 0.650

Luego calculamos la pendiente (a) y el intercepto (b) usando las formulas (1) y (2)
En el internet podrán encontrar diferentes sitios [7], [8] que les proporcionen programas ejecutables que realizan ajustes por mínimos cuadrados. A continuación se muestran los resultados de dos de esos programas.

1) Y = 2.670 + 1.980X; a= 1.980; δa = 0.014; b = 2.670; δb = 0.007; r = 0.9998

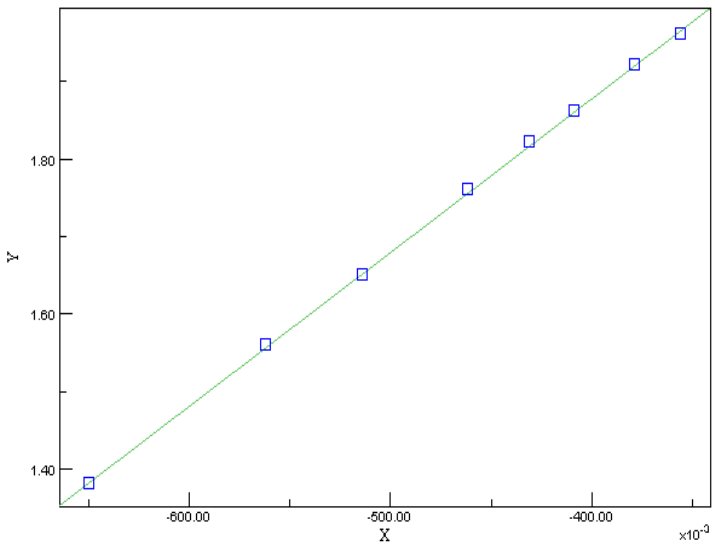
Cuadro de resultados:

Σx_i	Σy_i	Σx_i^2	Σy_i^2	$\overline{x}$
-3.763e+000	-1.391e+001	+1.839e+000	+2.446e+001	-4.704e-001

$\Sigma(x_i - \overline{x})^2$	$\Sigma x_i y_i$	$\Sigma(y_i - ax_i - b)^2$
+6.93819e-002	-6.40554e+000	+8.22527e-005

r = +9.99849e-001

a = +1.98000e+000 **b = +2.67009e+000**
E(a) = +1.406e-002 **E(b) = +6.740e-003**

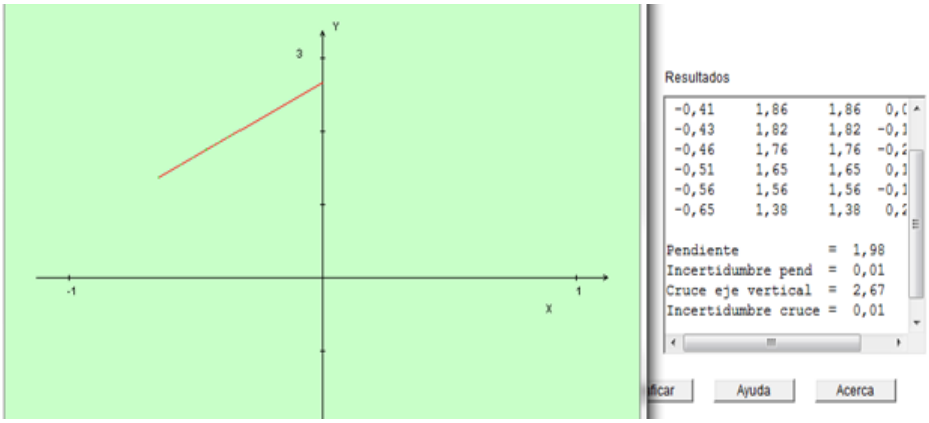


n = 8.0

X_1 = -0.356	;	Y_1 = 1.96	
X_2 = -0.379	;	Y_2 = 1.92	
X_3 = -0.409	;	Y_3 = 1.86	
X_4 = -0.431	;	Y_4 = 1.82	
X_5 = -0.462	;	Y_5 = 1.76	
X_6 = -0.514	;	Y_6 = 1.65	
X_7 = -0.562	;	Y_7 = 1.56	
X_8 = -0.650	;	Y_8 = 1.38	
X_9 = 0.0	;	Y_9 = 0.0	
X_{10} = 0.0	;	Y_{10} = 0.0	

Update

2) Y = 2.67 + 1.98X; a= 1.98; δa = 0.01; b = 2.67; δb = 0.01



La función exponencial

En la tabla 2 se muestra el monto (capital más intereses) de un capital C_0 que un banco ha prestado y que al final del año recuperará. Supongamos que el modelo matemático que relaciona las variables está representado por la siguiente función exponencial, $M = C_0 10^{at}$ la cual puede transformarse en $logM = at + logC_0$. Si usamos las nuevas variables Y=log M y X=t, obtenemos la relación lineal. $Y = aX+b$ donde b = logC_0. Obtener la función matemática que relacione las variables M y t

M($)	120.7	132.6	160.1	212.4	309.6
t(s)	2.0	3.0	5.0	8.0	12.0

Tabla 2

Para lograr esto debemos primero completar la siguiente tabla.

Y = logM($)	2.08	2.12	2.20	2.33	2.49
X = t(s)	2.0	3.0	5.0	8.0	12.0

Luego calculamos la pendiente (a) y el intercepto (b) usando las formulas (1) y (2)
Y = 2.00 + 0.04t; a = 0.04; b = 2.00; b = logC_0; $C_0 = 10^b$; $C_0 = 10^2 = 100$

Siendo la función: M = 100x10 $^{0.04t}$

Utilizando el applet mostrado en la siguiente dirección URL
http://cafpe10.ugr.es/test/teaching/labo_fisica_general/texto/applets/regresion.html
Obtuvimos el siguiente resultado
Y = 1.997 + 0.0412t; a = 0.0412; δa = 0.0003; b = 1.997; r = 0.999; b = logC_0; $C_0 = 10^b$; $C_0 = 10^{1.997} = 99.3$; así la función es: M = 100x10 $^{0.04t}$

Realización de la práctica

El objetivo en esta práctica es que el estudiante pueda obtener la función matemática que relaciona dos variables mediante el método de ajuste lineal por mínimos cuadrados, para

esto el estudiante debe venir revisando los fundamentos teóricos en los que se basará la práctica, tener claridad sobre el procedimiento a seguir y resolver las preguntas planteadas al final de la unidad

Problema a resolver

Saber calcular el error cuadrático medio (SEM) y aplicar el ajuste lineal a un conjunto de datos bidimensionales mediante el método de mínimos cuadrados.

Base Teórica

Para esta práctica es necesario revisar los conceptos de: Diferencia entre muestra y población. Distribución normal. Definición de media, varianza, desviación estándar y error cuadrático medio. Ajuste lineal por mínimos cuadrados. Se recomienda visitar los enlaces en internet y utilizar el applet para realizar el ajuste lineal, esto le servirá para comprobar sus resultados obtenidos manualmente.

Actividades a desarrollar

1) Calcular la media, mediana, moda, desviación estándar y SEM, de la siguiente serie de números: 5, 3, 6, 5, 4, 5, 2, 8, 6, 5, 4, 8, 3, 4, 5, 4, 8, 2, 5, 4.

2) Dada la siguiente información:

X	10	20	30	40	50	60	70	80	$\Sigma X=$	
Y	1.06	1.33	1.52	1.68	1.81	1.91	2.01	2.11	$\Sigma Y=$	
XY									$\Sigma(XY)=$	
X^2									$\Sigma(X^2)=$	

a) Completar la tabla mostrada.
b) Calcular el valor de la pendiente y el intercepto de la recta Y = aX + b, mediante la fórmula (1) y (2)
c) Completar la tabla, realizando el cambio de variables indicado

W =logX									$\Sigma W=$	
Z = logY									$\Sigma Z=$	
WZ									$\Sigma(WZ)=$	
W^2									$\Sigma(W^2)=$	

d) Calcular el valor de la pendiente y el intercepto de la recta Z = aW + b, mediante la fórmula (1) y (2)
e) Comparar los resultados encontrados en la parte b y d, con los obtenidos del enlace http://cafpe10.ugr.es/test/teaching/labo_fisica_general/texto/applets/regresion.html
f) ¿Cuál de los dos ajustes presenta una mejor correlación de los datos medidos?
g) ¿En qué se basó para dar su respuesta?

Preguntas para Prueba de Entrada

1. Se ha tomado una muestra de 16 anillos de un lote de producción, para medir su diámetro interno d(mm).

18.43	18.42	18.47	18.35	18.25	18.22	18.53	18.54
18.65	18.32	18.45	18.36	18.45	18.43	18.48	18.45

Usted debe completar la información requerida en el cuadro adjunto

N	Moda	Mediana	Media	Desviación estándar	Error Cuadrático Medio (SEM)

2. Dos estudiantes A y B independientemente, realizan las observaciones de la velocidad media (V) en diferentes intervalos (t), de un móvil que se mueve por un plano inclinado sin fricción.

t(s)	0.321	0.722	1.222	2.118
V(cm/s)	62.3	55.4	49.2	37.8

Datos tomados por el estudiante A

t(s)	0.151	0.292	0.427	0.568
V(cm/s)	66.3	68.6	70.3	71.5

Datos tomados por el estudiante B

Aplicando el método mínimos cuadrados, obtenga la relación lineal de las variables medidas, a partir de los datos tomados por cada estudiante.

3. Un disco de masa M y radio R, rota alrededor de un eje fijo que pasa por el centro del disco, si se registra los valores de aceleración angular (α) que adquiere para diferentes momentos de torsión (τ) en la siguiente tabla.

$X=\tau(Nm)x10^{-4}$	3.7	1.5	1.9	2.3	3.0	3.4
$Y=\alpha(rad/s^2)$	0.19	0.59	0.72	0.86	1.11	1.25

Aplicando el método de mínimos cuadrados obtener la relación: $\alpha = a\tau + b$

4. Un estudiante registra las mediciones en un experimento, en la siguiente tabla.

$X(m^2)x10^{-3}$	1.6	6.4	14.4	25.6
$Y(kgm^2)\ x10^{-3}$	14	18	23	31

Mediante un ajuste lineal obtener la ecuación de la recta.

Completar la siguiente tabla.

$X(m^2)x10^{-3}$	1.6	6.4	14.4	25.6	$\Sigma X=$	
$Y(kgm^2)\ x10^{-3}$	14	18	23	31	$\Sigma Y=$	
XY					$\Sigma(XY)=$	
X^2					$S(X^2)=$	

5. Los datos observados en un experimento se anotan en la siguiente tabla.

X	0	5	10	15	20
Y	68	72	75	77	80

Mediante un ajuste lineal obtener la ecuación de la recta.

Completar la siguiente tabla.

X	0	5	10	15	20	ΣX=	
Y	68	72	75	77	80	ΣY=	
XY						Σ(XY)=	
X^2						S(X^2)=	

Usando el programa ejecutable obtenemos el siguiente resultado

Referencias

[7] http://cafpe10.ugr.es/test/teaching/labo_fisica_general/texto/applets/regresion.html

[8] http://didactica.fisica.uson.mx/applets/laboratorio/laboratorio/recta.html