Requerimientos del Programa 6

Utilizando el **PSP 2.1** escribe un programa que:

- Lea del teclado el nombre de un archivo
- Lea de este archivo lo siguiente:
  - El primer renglón contiene un número real mayor o igual a cero, al cual llamaremos x_k.
  - A partir del segundo renglón habrá en cada renglón una pareja (x, y) de dos números reales mayores o iguales a cero, separados por una coma
  - El fin del archivo marca el final de las parejas de datos
- Calcule los siguientes datos
  - La cantidad de parejas de datos leídas (N)
  - Los coeficientes de correlación $r_{x,y}$ y r^2
  - ***La significancia de tal correlación***
  - Los parámetros de regresión lineal β_0 y β_1
  - Una predicción mejorada y_k, donde $y_k = \beta_0 + \beta_1 x_k$
  - ***El intervalo de predicción al 70% para tal estimado***
- Escriba en pantalla estos valores calculados de acuerdo con el siguiente formato:

```
N   = x
xk  = x.xxxxx
r   = x.xxxxx
r2  = x.xxxxx
b0  = x.xxxxx
b1  = x.xxxxx
yk  = x.xxxxx
sig= x.xxxxxxxxxx
ran= x.xxxxx
LS  = x.xxxxx
LI  = x.xxxxx
```

NOTAS:

- ✓ Explicación de siglas: "sig" = significancia, "ran" = rango (intervalo de predicción 70%), "LS" = límite superior (UPI en inglés), "LI" = Límite inferior (LPI en inglés)
- ✓ Los valores de xk, r, r2, b0, b1, yk, ran, LS y LI se desplegarán con 5 decimales (redondeados hacia arriba en su último dígito, por ejemplo: 0.123455 se desplegará como 0.12346, mientras que 0.123454 se desplegará como 0.12345) mientras que sig se desplegará con 10 decimales (redondeado hacia arriba en su último dígito)
- ✓ "LI" no puede ser negativo (su valor mínimo es cero)

Otras características que ***debe*** cumplir el programa:

- No utilizará ningún GUI para operar (funcionará desde la consola)
- Debe estar construido con programación orientada a objetos
- Debe contar con al menos 3 clases "relevantes"
- El ***único*** código que puede ser reutilizado es el de tus programas 1 a 5
- Debe manejar apropiadamente ***todas*** las condiciones normales y anormales
- Debe pasar exitosamente ***todos*** los casos de prueba (*error máximo 0.0001, excepto para significancia cuyo error máximo es 0.00000001*):
  - Los diseñados por ti en la fase de diseño, y
  - Los siguientes 2 casos de prueba (es obligatorio incluirlos en el Diseño de las Pruebas):

Objetivo de la prueba	Instrucciones y datos de entrada	Resultados Esperados
Probar con una lista de datos	Teclear en pantalla: Arch1.txt	N = 10 xk = 386.00000 r = 0.95450 r2 = 0.91106 b0 = -22.55253 b1 = 1.72793 yk = 644.42938 sig= 0.0000177517 ran= 230.00172 LS = 874.43110 LI = 414.42766
Probar con una lista de datos	Teclear en pantalla: Arch2.txt	N = 14 xk = 149.00000 r = 0.91381 r2 = 0.83505 b0 = -23.43891 b1 = 1.42554 yk = 188.96720 sig= 0.0000049053 ran= 204.66397 LS = 393.63116 LI = 0.00000

Fin de los requerimientos

Explicación y ejemplo de cómo se realizan los cálculos (no son requerimientos)

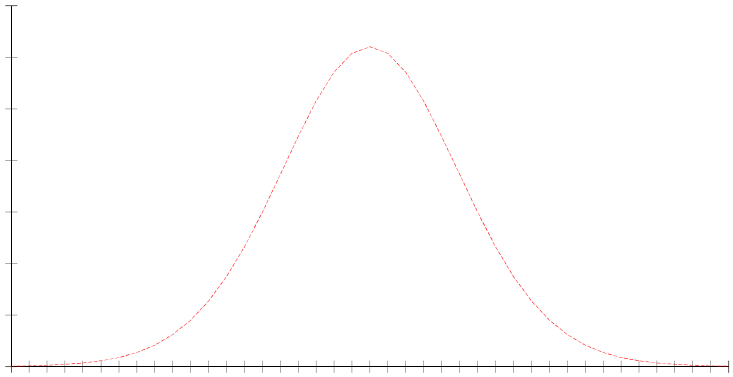
(Tomado del curso original del PSP del Software Engineering Institute)

Significance

The significance test

The significance test determines the likelihood that a strong correlation is random, and is therefore of no practical significance.
For example, a data set with only two points will always have an $r^2 = 1$, but this correlation is not significant.

Student *t*-Distribution



Calculating significance

The procedure for calculating the correlation significance is as follows.

1. Compute the value of x, such that

$$x = \frac{\left|r_{x,y}\right|\sqrt{n-2}}{\sqrt{1-r_{x,y}{}^2}}$$

where

- $r_{x,y}$ is the correlation
- n is the number of data points

2. Find the probability p by numerically integrating the t distribution for $n-2$ degrees of freedom, from 0 to x.
3. Calculate the tail area as $1-2*p$. (The area under the curve from –x to +x is twice the area from 0 to x, or $2*p$; the remaining area in the upper and lower tails is $1-2*p$).

A tail area ≤ 0.05 is considered as strong evidence that there is a relationship.
A tail area ≥ 0.2 indicates a relationship that is due to chance.

Prediction interval

Prediction interval

The prediction interval provides a likely range around the estimate.

- A 70% prediction interval gives the range within which 70% of the estimates will fall.
- It is not a forecast, only an expectation.
- It only applies if the estimate behaves like the historical data.

It is calculated from the same data used to calculate the regression parameters.

Prediction interval procedure

To calculate the prediction interval, use the following steps.

1. Calculate the *Range* for a 70% interval.
2. Calculate the UPI as $y_k + Range(70\%)$.
3. Calculate the LPI as $y_k - Range(70\%)$ (NOTE: it can't be negative).

The formula for calculating the prediction range is

$$Range = t(0.35, dof)\sigma \sqrt{1 + \frac{1}{n} + \frac{(x_k - x_{avg})^2}{\sum_{i=1}^{n} (x_i - x_{avg})^2}}$$

where

- x is your historical data
- n is the number of historical data points
- $t(0.35, dof)$ is the value of x for a t distribution for n - 2 degrees of freedom where $p = 0.35$

The formula for calculating the standard deviation term is

$$\sigma = \sqrt{\left(\frac{1}{n-2}\right)\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2}$$

where

- x, y are your historical data
- n is the number of historical data points