Testing Linear Correlation

&

Making Prediction

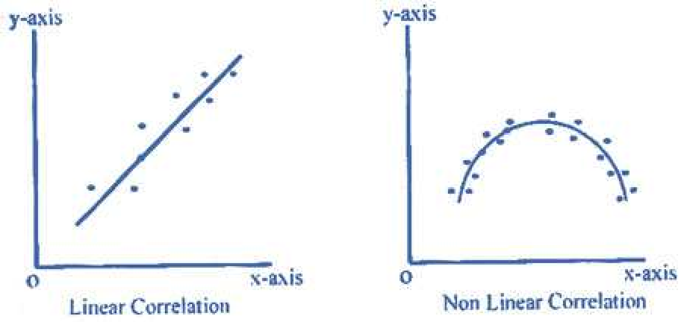
What is **Testing Linear Correlation**?

Testing Linear Correlation is to determine if there is a significant linear correlation between two variables in the given sample of ordered-pairs.

What is a **Correlation**?

A **Correlation** between two variables is when there is an apparent association between the values of one variable with the corresponding values from the other variable.

Linear vs Nonlinear Correlation



What is **Linear Correlation**?

Linear Correlation is defined when the ratio of proportion of two given variables are almost the same for all points.

What is **Linear Correlation Coefficient**?

Linear Correlation Coefficient is a numerical value that measures the strength of the linear correlation between the paired x and y for all values in the sample. We denote this value by r.

What are the properties of r?

- $-1 \leq r \leq 1$
- It is designed to measure the strength of only linear relationship.
- It is very sensitive and changes value if the sample contains any outliers.
- The **Linear Correlation Coefficient** is considered **significant** when $|r|$ is fairly close to 1.

How do we compute r?

- Compute $\sum x$, $\sum y$, and $\sum xy$.
- Compute $\sum x^2$, and $\sum y^2$.
- Now we use the formula

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}}$$

where n is the number of ordered-pairs.

It is worth noting that r is usually calculated with a computer software or a calculator.

Example:

The study time and midterm exam score for a random sample of 10 students in a statistics course are shown in the following table.

Study Time; x (Hours)	3	4	4	5	6	6	8	9	10	9
Midterm Score; y	57	65	72	74	70	80	85	90	97	92

Find the value of linear correlation coefficient r by using the formula.

Solution:

We first identify that $n = 10$, then find and verify that
$\sum x = 64$, $\sum y = 782$, $\sum xy = 5277$, $\sum x^2 = 464$,
$\sum y^2 = 62632$,
and then we apply these values in the the formula

$$\begin{aligned} r &= \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}} \\ &= \frac{10(5277) - (64)(782)}{\sqrt{10(464) - (64)^2}\sqrt{10(62632) - (782)^2}} \\ &= \frac{2722}{\sqrt{544}\sqrt{14796}} \\ &\approx 0.959 \end{aligned}$$

Testing Process for **Linear Correlation**:

1. Set-Up H_0 and H_1.
   - $H_0 : \rho = 0 \Rightarrow$ **Linear Correlation is not significant**
   - $H_1 : \rho \neq 0 \Rightarrow$ **Linear Correlation is significant**
2. Use TI commands or formula to find the P-Value for TTT.
3. By P-Value method:
   - If $P - value \leq \alpha$, then H_1 is valid.
   - If $P - value > \alpha$, then H_0 is valid.
4. Final Conclusion:
   - Draw the final conclusion whether the linear correlation is significant or not.

Example:

The study time and midterm exam score for a random sample of 10 students in a statistics course are shown in the following table.

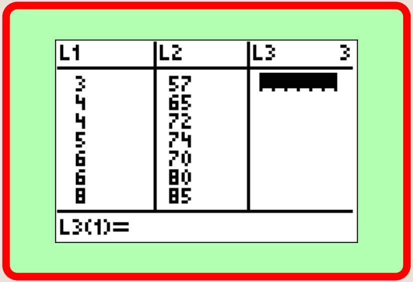
Study Time; x (Hours)	3	4	4	5	6	6	8	9	10	9
Midterm Score; y	57	65	72	74	70	80	85	90	97	92

Use TI command **LinRegTTest** to find

1. CTS t
2. P-Value p
3. Linear Correlation Coefficient r

Solution:

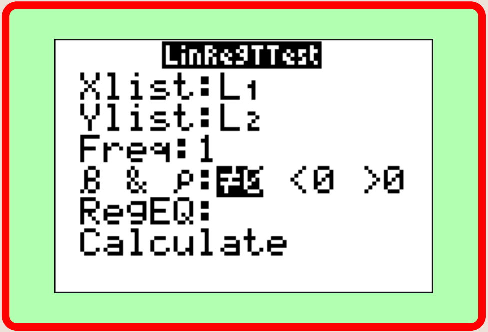
We store all x values in L_1 and corresponding y values in L_2,



then we perform the following steps.

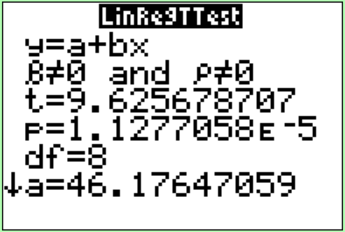
STAT $\rightarrow$ TESTS $\downarrow$ LinRegTTest **Xlist:** L_1 **Ylist:** L_2 Freq: 1 $\rho :\neq 0$ RegEQ: blank

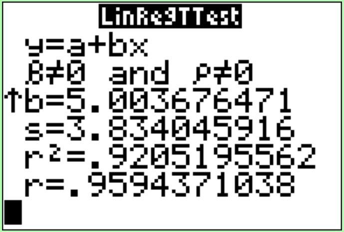
Solution Continued:



Now press **Calculate** key to view the results.

Solution Continued:



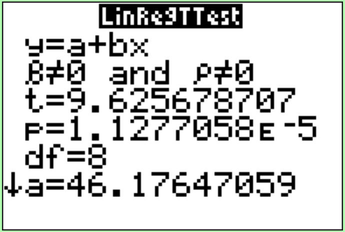


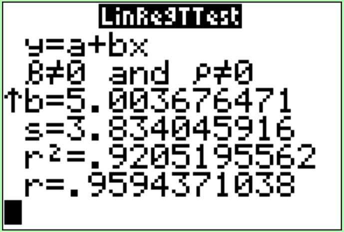
Here are the answers:

1. CTS $t = 9.626$
2. P-Value $p = 1.1 \times 10^{-5}$
3. Linear Correlation Coefficient $r = .959$

Solution Continued:

We can also find the equation of the regression line $\hat{y} = a + bx$.





Here are the answers:

1. $a = 46.176 \approx 46$
2. $b = 5.004 \approx 5$
3. Equation of the regression line $\hat{y} \approx 46 + 5x$

How to Find **CTS** using **Formula**:

The formula to find the CTS is

$$t = r \cdot \sqrt{\frac{n-2}{1-r^2}}.$$

Example:

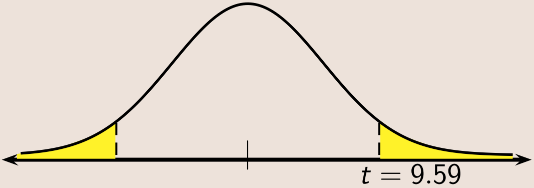
Given $n = 10$ and $r = 0.959$, find the value of the CTS by using the formula and the corresponding P-value.

Solution:

$$\begin{aligned} t &= r \cdot \sqrt{\frac{n-2}{1-r^2}} \\ &= 0.959 \cdot \sqrt{\frac{10-2}{1-0.959^2}} \\ &= 0.959 \cdot \sqrt{\frac{10-2}{1-0.920}} \\ &= 0.959 \cdot \sqrt{\frac{8}{0.08}} \\ &= 0.959 \cdot \sqrt{100} \\ &= 0.959 \cdot 10 \\ &= 9.59 \end{aligned}$$

Solution Continued:

Now, to find the corresponding P-Value, we use the TI command **tcdf(L,U, df)** with $df = n - 2$. Make sure to multiply the area by 2 since it is a TTT.



P-value $p = 2 \cdot \textbf{tcdf}(9.59, E99, 8) = 1.16 \times 10^{-5}$

Example:

The study time and midterm exam score for a random sample of 10 students in a statistics course are shown in the following table.

Study Time; x (Hours)	3	4	4	5	6	6	8	9	10	9
Midterm Score; y	57	65	72	74	70	80	85	90	97	92

Use the P-Value method at 0.05 significance level,

to determine whether the linear correlation is significant or not.

Solution:

We first set up H_0 and H_1,

- $H_0 : \rho = 0 \Rightarrow$ **Linear Correlation is not significant**
- $H_1 : \rho \neq 0 \Rightarrow$ **Linear Correlation is significant**

Since $1.1 \times 10^{-5} \leq 0.05 \Rightarrow P - value \leq \alpha$, then H_1 is valid which implies

Linear Correlation is significant.

The conclusion whether the linear correlation is significant or not is an important factor as we make predictions.

How do we make **prediction**?

- When **linear correlation is significant,** use $\hat{y} = a + bx$.

  Plug in the given x value to find the prediction value y.

- When **linear correlation is not significant,** use $\overline{y}$.

Example:

Eight pairs of data yield the regression equation
$\hat{y} = 55.6 + 2.8x$ with $\bar{y} = 71.5$.
What is the best predicted value for y for $x = 5.5$ if we assume the linear correlation is significant?

Solution:

Since the linear correlation coefficient is significant, we use the equation of the regression line $\hat{y} = 55.6 + 2.8x$. and plug in $x = 5.5$ to find the prediction value.

$$\begin{aligned}\hat{y} &= 55.6 + 2.8x \\ &= 55.6 + 2.8(5.5) \\ &= 55.6 + 15.4 \\ &= 71\end{aligned}$$

So, our prediction value is 71.

Example:

Ten pairs of data yield the regression equation
$\hat{y} = 73.5 - 4.5x$ with $\bar{y} = 58.5$.

What is the best predicted value for y for $x = 4.5$ if we assume the linear correlation is not significant?

Solution:

Since the linear correlation is not significant,

we use $\bar{y}$ as the prediction value regardless of the value of x.

So, our prediction value is 58.5.