Determine whether there is a

Linear Correlation

between two variables x and y

Notation

n represents the number of pairs of data (x,y)

$\sum$ indicates the operation of addition, to sum

$\sum \mathbf{x}$ indicates to add all the x values

$\sum \boldsymbol{y}$ indicates to add all the y values

$\sum \boldsymbol{xy}$ indicates to add all the xy values

$\sum \boldsymbol{x^2}$ indicates to add all the x^2 values

$\sum \boldsymbol{y^2}$ indicates to add all the y^2 values

$(\sum \mathbf{x})^2$ indicates to square the $\sum x$ value

$(\sum \mathbf{y})^2$ indicates to square the $\sum y$ value

r represents the **linear correlation coefficient** for the paired **sample** data values (x,y)

ρ represents the **linear correlation coefficient** for the **population** of paired data values (x,y)

Requirements

The linear coefficient can always be computed for any set of sampled bivariate (x,y) pairs of data. But, the following requirements should be satisfied when using a sample pair of bivariate (x,y) data when making a conclusion about whether there is a linear correlation in the corresponding population of paired data.

1. The sample of paired data (x,y) is a simple **random sample** of quantitative (can be measured) data. Make sure data is not convenient or conducted by voluntary responses.

2. A visual confirmation of a **Scatter Plot** must confirm that the points approximate a **straight-line pattern**.

3. Results are strongly influenced by outliers, so they must be removed when they are known to be errors. At times, we may calculate the linear correlation coefficient with or without other known outliers when considering a linear correlation.

Formally, we require both variables x and y to have Normal Distributions which is known as Bivariate Normal Distribution. This is difficult to check, so we rely on requirements 2 and 3.

Formula

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

Facts

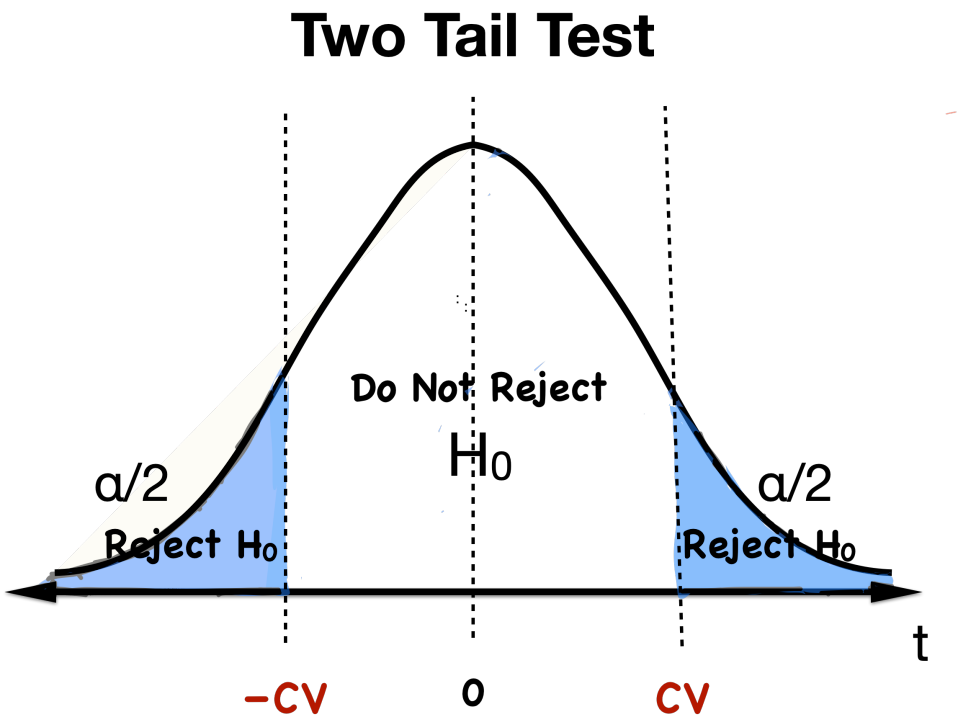
1. $-1 \leq r \leq 1$ and r measures the strength of a linear relationship between two quantities x and y.
2. r is very sensitive to outliers and a single outlier can dramatically change its value.
3. The Test Statistic t is used to test whether there is significant linear correlation between two quantities represented by the variables x and y.

Formal Hypothesis Test of ρ
(Linear Correlation Coefficient for the population)

Null Hypothesis $H_0: \rho = 0$ (No Linear Correlation)
Alternate Hypothesis $H_1: \rho \neq 0$ (Linear Correlation)

Use the t-distribution and your test statistic to determine whether there is sufficient evidence to support the claim that there is a linear correlation between two quantities x and y using either $\alpha = 0.01$ or $\alpha = 0.05$ level of significance.

Test Statistic $t = \dfrac{r}{\sqrt{\frac{1-r^2}{n-2}}}$ where $df = n - 2$



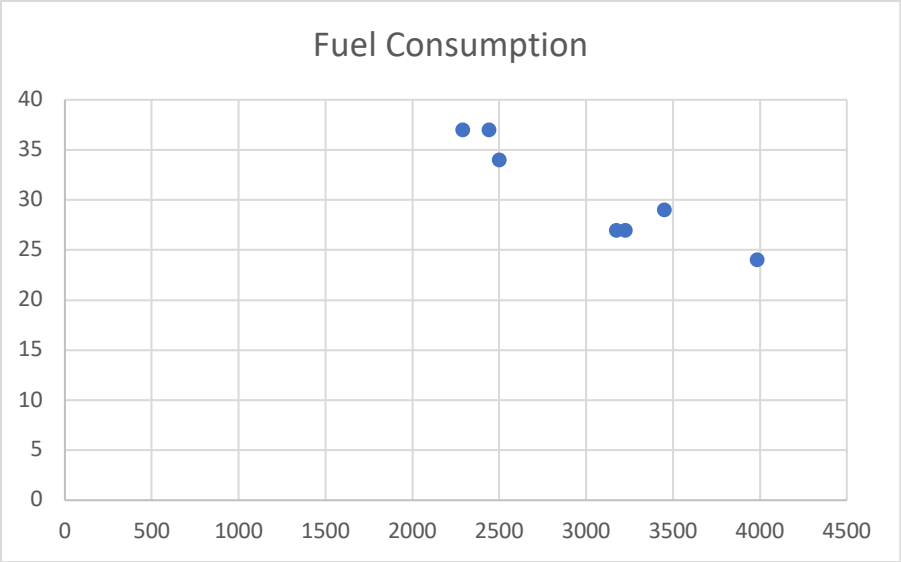
Example

Car Weight and Fuel Consumption The bivariate data listed below are the weights (pounds) and the highway fuel consumption amounts (MPG) of randomly selected cars. Is there a linear correlation between weight and highway fuel consumption using $\boldsymbol{\alpha = 0.05}$.

Weight	3175	3450	3225	3985	2440	2500	2290
Fuel Consumption	27	29	27	24	37	34	37

Step 1

Plot the data on a **Scatter Plot** to see if the data is in the form of a **straight-line pattern** and to determine if there are any **outliers** that may be removed.



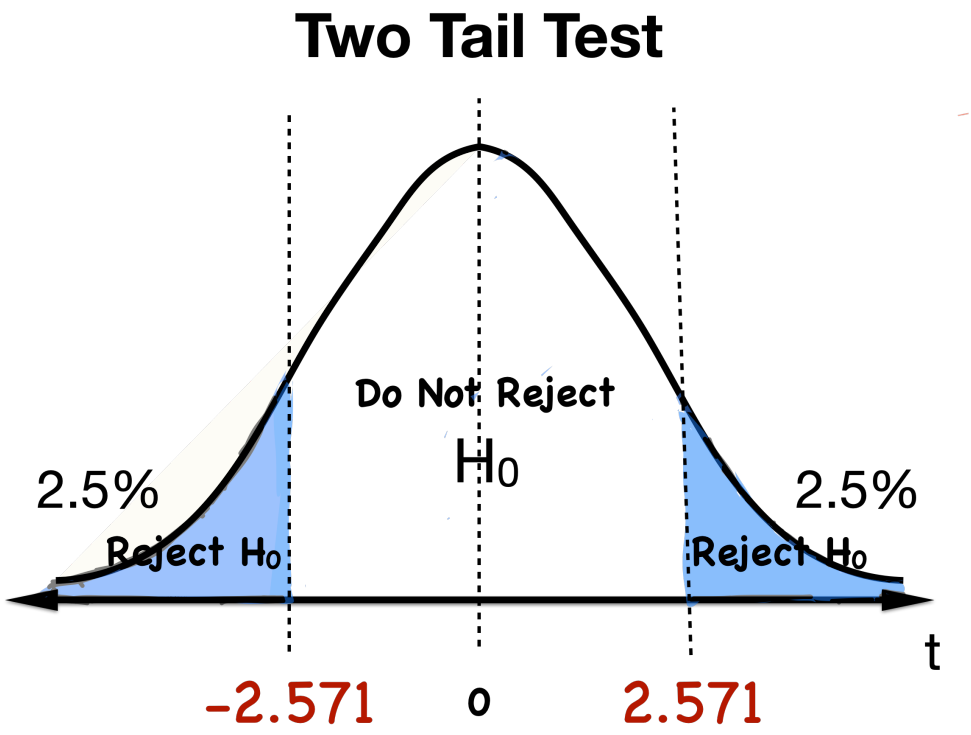
Step 2

Set up your formal **hypothesis test** for the population linear correlation coefficient ρ using either $\alpha = 0.01$ or $\alpha = 0.05$

Using $\alpha = 0.05$ with a sample size of 7 we have a df=5 and critical values of -2.571 and 2.571 according to the t-table

$H_0: \rho = 0$ (No Linear Correlation)

$H_1: \rho \neq 0$ (Linear Correlation)



Step 3

To compute the sample test statistic, we need to compute the sample linear correlation coefficient r which is found by the formula:

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

								sum
$\boldsymbol{x}$ **Weight**	3175	3450	3225	3985	2440	2500	2290	21065
$\boldsymbol{y}$ **Fuel Consumption**	27	29	27	24	37	34	37	215
$\boldsymbol{x^2}$	10080625	11902500	10400625	15880225	5953600	6250000	5244100	65711675
$\boldsymbol{y^2}$	729	841	729	576	1369	1156	1369	6769
$\boldsymbol{xy}$	85725	100050	87075	95640	90280	85000	84730	628500

$$r = \frac{7 \cdot 628500 - 21065 \cdot 215}{\sqrt{7 \cdot 65711675 - 21065^2}\sqrt{7 \cdot 6769 - 215^2}}$$

$$\boldsymbol{r \approx -0.944}$$

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \quad \text{so that} \quad t = \frac{-0.944}{\sqrt{\frac{1-(-0.944)^2}{7-2}}} \text{ or } \boldsymbol{t \approx -6.398}$$

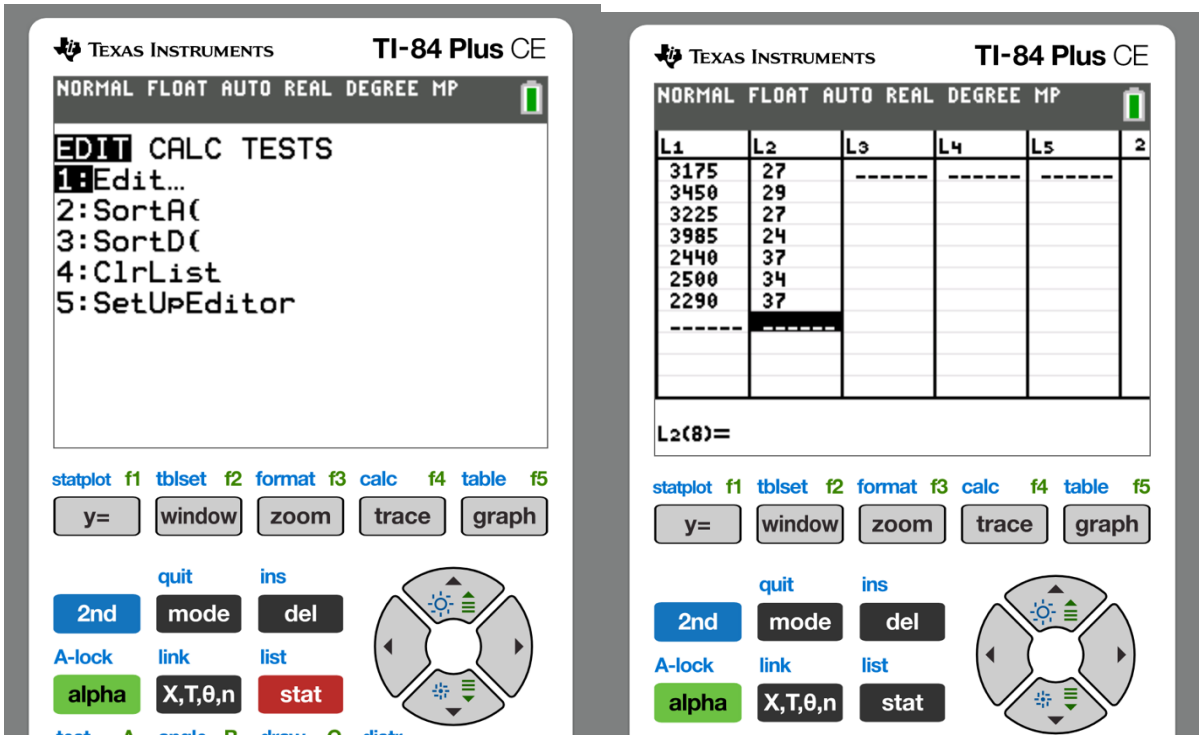
Conclusion

The **test statistic** indicates to reject the null hypothesis ($\rho = 0$) or we can say the data suggest there is a linear correlation between the weight of a car and its fuel consumption.

What about the TI-83 or TI-84 Plus CE Calculator?

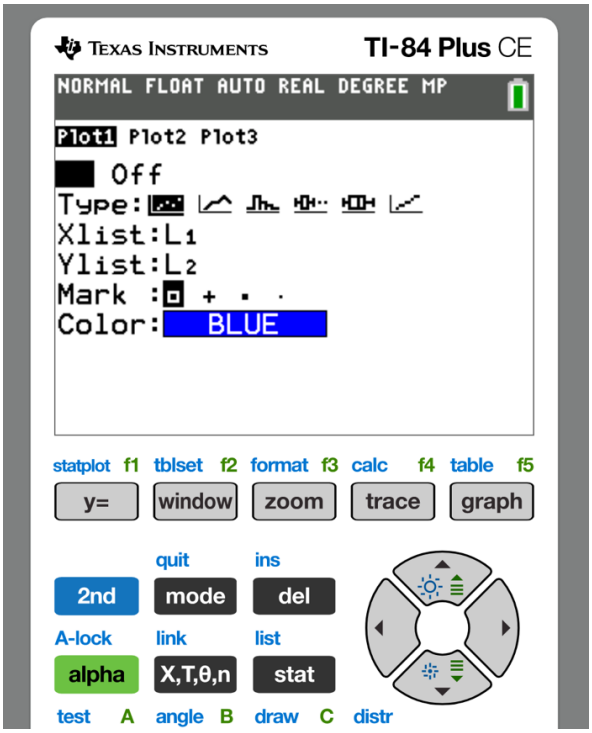
You will need to enter a list of data values for the x and y variables by following the procedure below to enter a list of data.

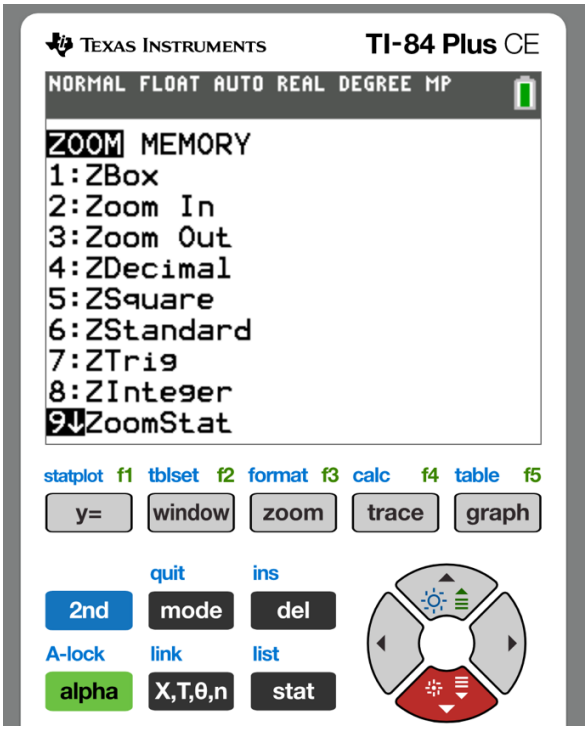
1. Press **STAT** make sure you are on **EDIT** in the top menu.
2. Scroll to **EDIT** then press **ENTER**. This will allow you to access the L_1 and L_2 lists for your x and y variables. Use the arrow keys and enter to input the bivariate data values.

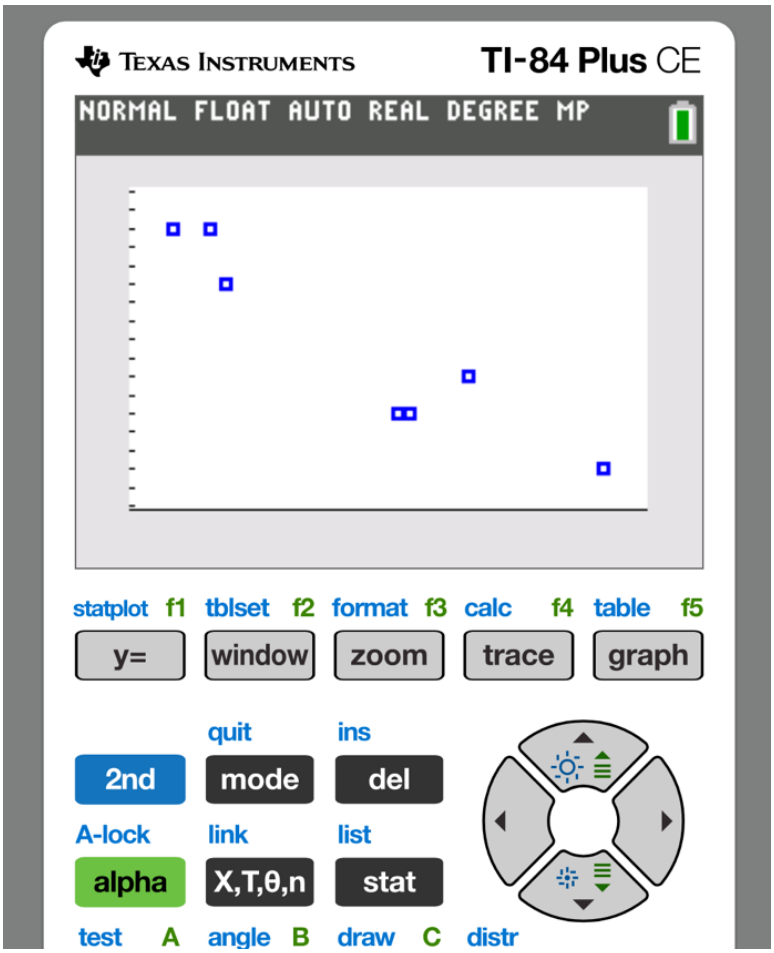


Using the TI-83/TI-84 Plus Calculator to create a Scatter Plot

Scatter Plot Open the **STATS PLOTS** menu
1. By pressing **2nd y=**
2. Press **ENTER** to access the **Plot 1** settings.
a) Select **ON** and press **ENTER**.
b) Select the first chart option (Scatter Plot), then press **ENTER.**
c) Enter names of lists containing data for the x and y variables.
3. Press **zoom** then **9** (Zoom STAT) to generate the scatter plot.





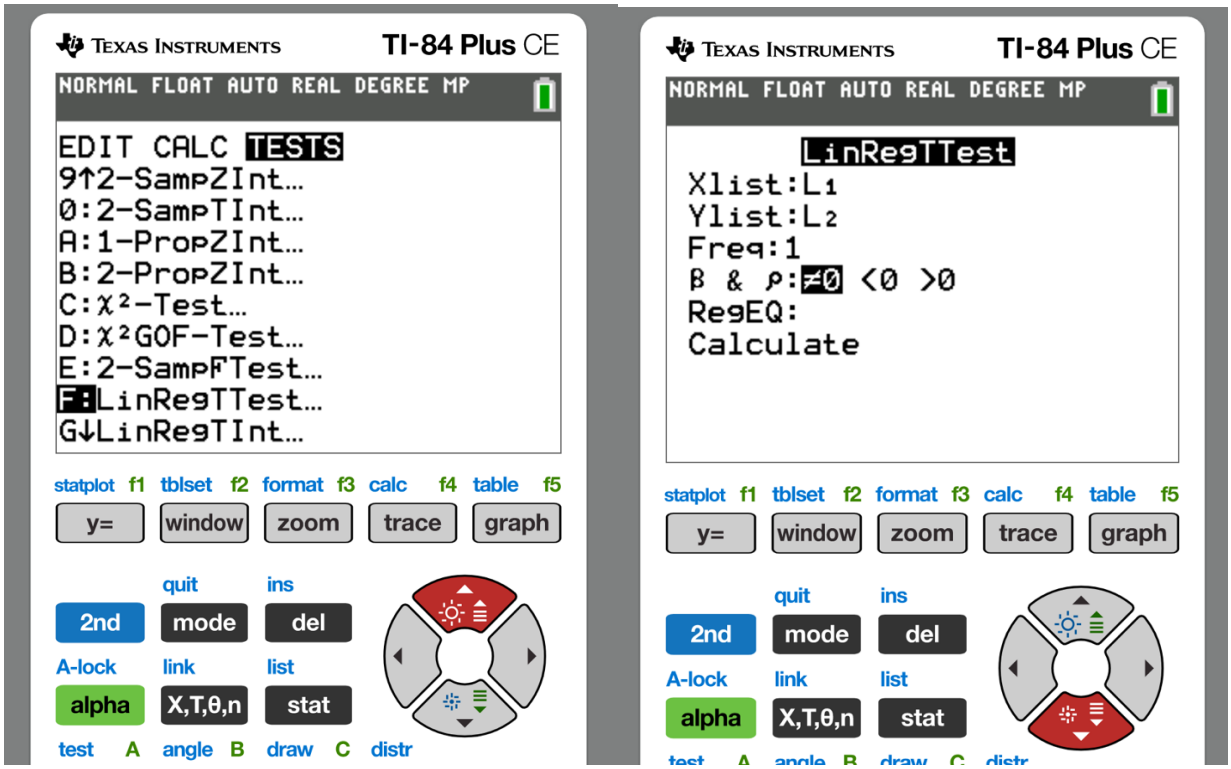


This is what you should see, a **Scatter Plot**.

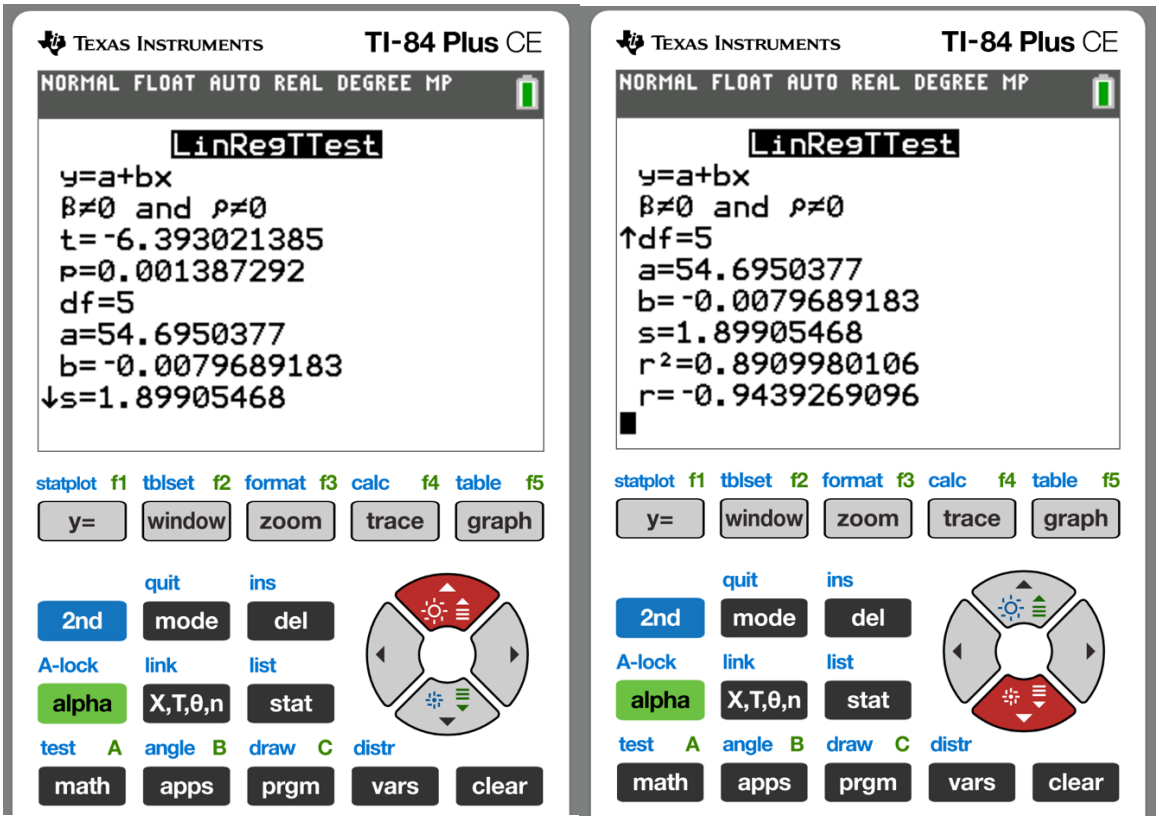
Computing the Linear Correlation Coefficient r, Test Statistic t, and p value p

TI-83 or TI-84 Plus Using the **Linear Regression T-Test** function and entering **Bi-variate Data**

1. Press **STAT** then **TESTS** in the top menu.
2. Select **LinRegTTest** in the menu and press **ENTER**.
3. Enter the list names for the bivariate data (x and y variables). Enter 1 for the **FREQ** and for β and ρ select $\neq 0$ to test the null hypostesis of no correlation.
the menu and press **enter**.
4. Select **Calculate** and press **ENTER.**



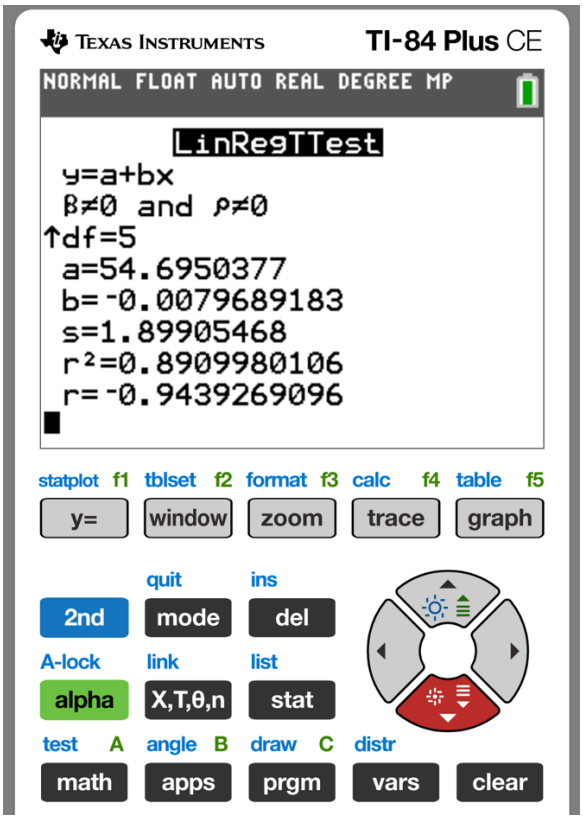
This is the results you obtain when following the procedure.



Notice the Test Statistic value of $\boldsymbol{t = -6.393}$ so Reject H_0
p value $\boldsymbol{p \approx 0.00}$ and $0.00 < 0.05$ so Accept H_1
The Sample Suggests **There is a Linear Correlation**
The Sample Linear Correlation Coefficient $\boldsymbol{r \approx -0.944}$
Weak Negative Linear Correlation

Best Fit Line aka Regression Line

Using the values of a and b in the calculator we can determine the regression line for the bivariate data. $y = a + bx$



$$\hat{y} = -.008x + 54.695$$

This linear equation can be used to be predictive for any data value x which will deliver a corresponding predictive value for y.

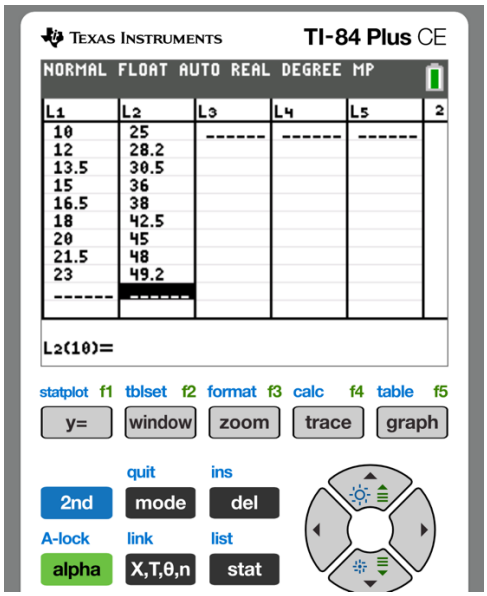
Example

The following data (x, y) is bivariate. Use the **TI-84 calculator** and the **5% level of significance** to test the claim that there is a linear correlation between the variables x and y. Determine the test statistic and the linear coefficient r.

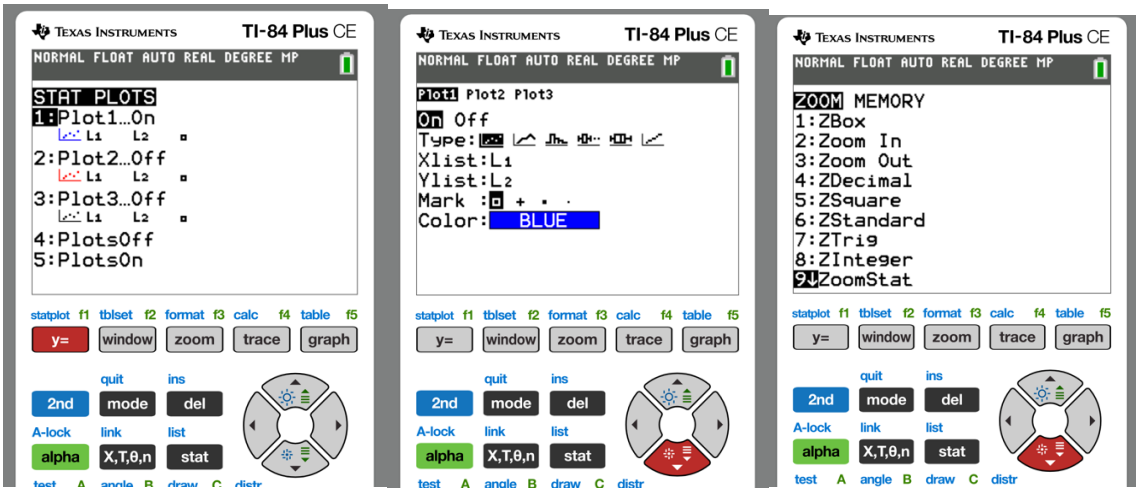
x	10	12	13.5	15	16.5	18	20	21.5	23
y	25	28.2	30.5	36	38	42.5	45	48	49.2

If there is a linear correlation between the variables x and y , what is the regression line?
If there is a linear correlation and $x = 11$, then what is the value of y?

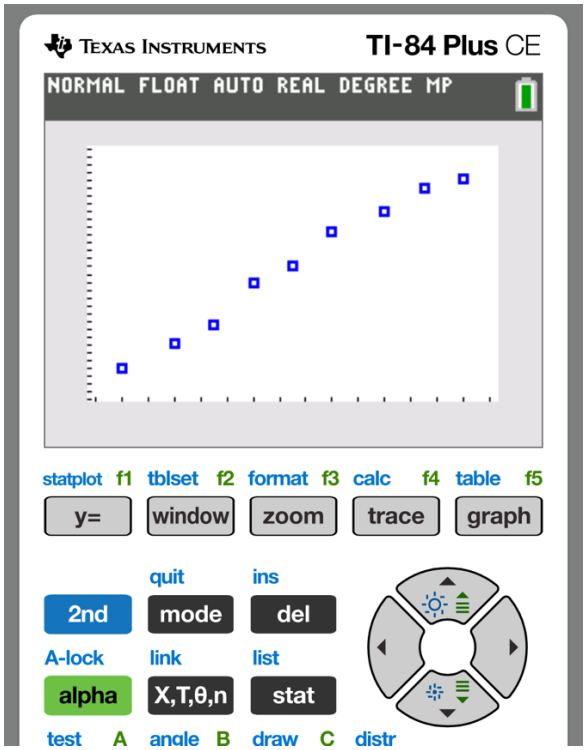
Enter Bivariate Data



Create a Scatter Plot

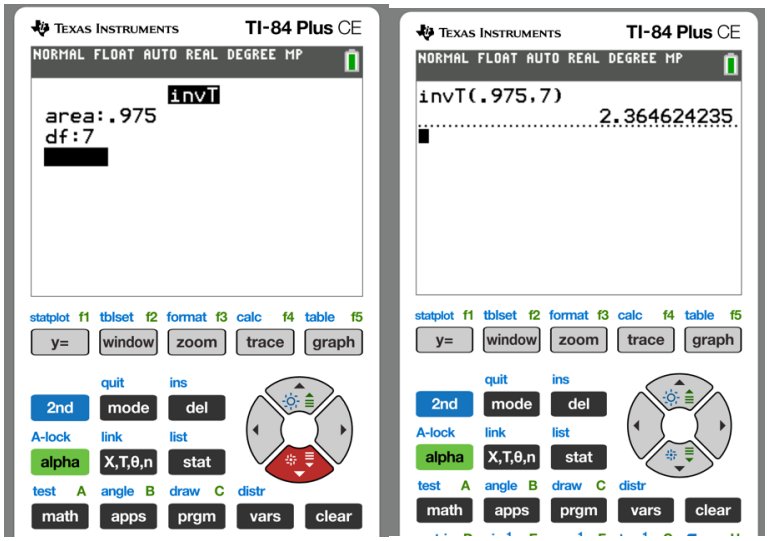


Scatter Plot



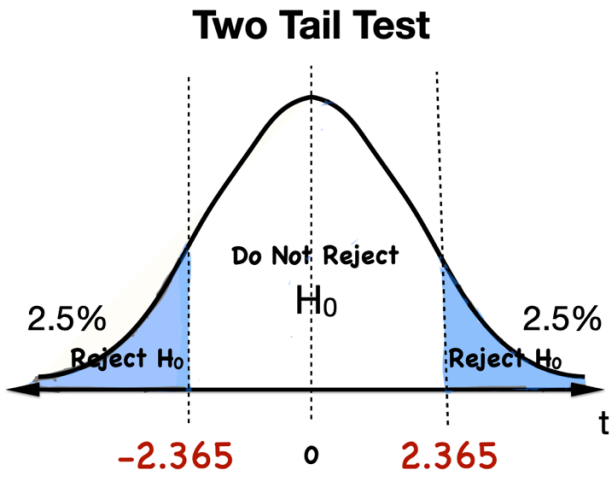
Looks Like a Strong Positive Linear Correlation!

Create Your Formal Hypothesis Test Regarding $\rho \neq 0$ Claiming There is a Linear Correlation $\alpha = 5\%$ and $df = 7$

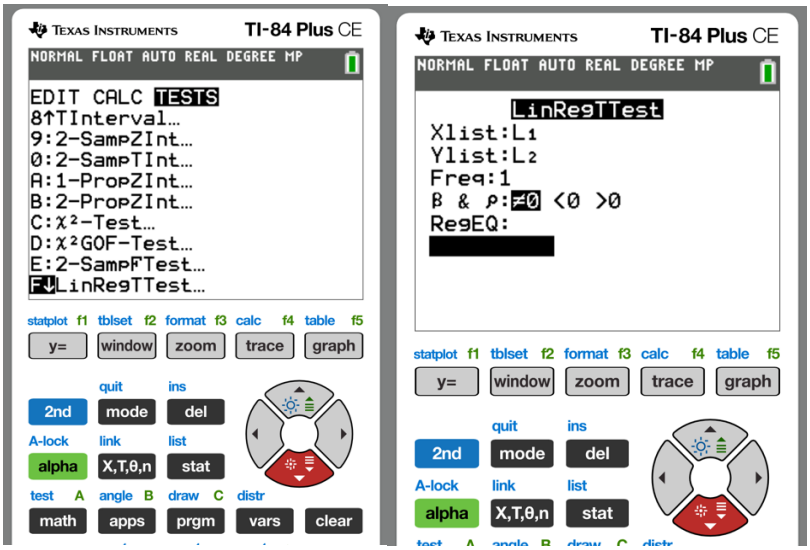


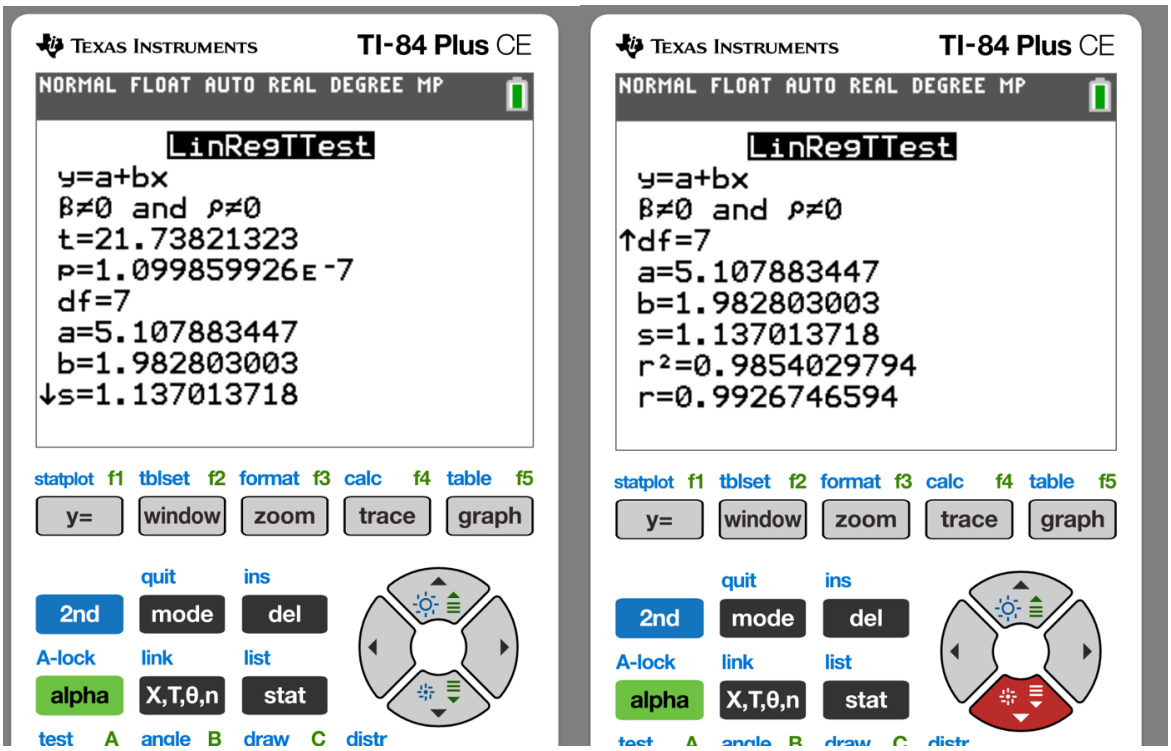
$H_0: \rho = 0$ (No Linear Correlation)

$H_1: \rho \neq 0$ (Linear Correlation) $Claim$



LinRegTTest





$t \approx 21.74$ so Reject H_0

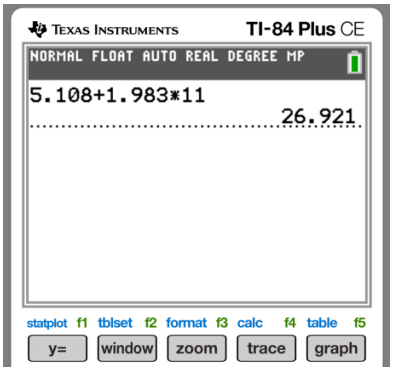
There is a linear Correlation between x and y

$r \approx 0.993$ which is very close to 1; **Strong Positive Linear Correlation**

If there is a linear correlation between the variables x and y , what is the regression line?

$$y = 5.108 + 1.983x$$

If there is a linear correlation and $x = 11$, then what is the value of y?



$$y \approx 26.921$$

Problem 1

Study Time and Grade

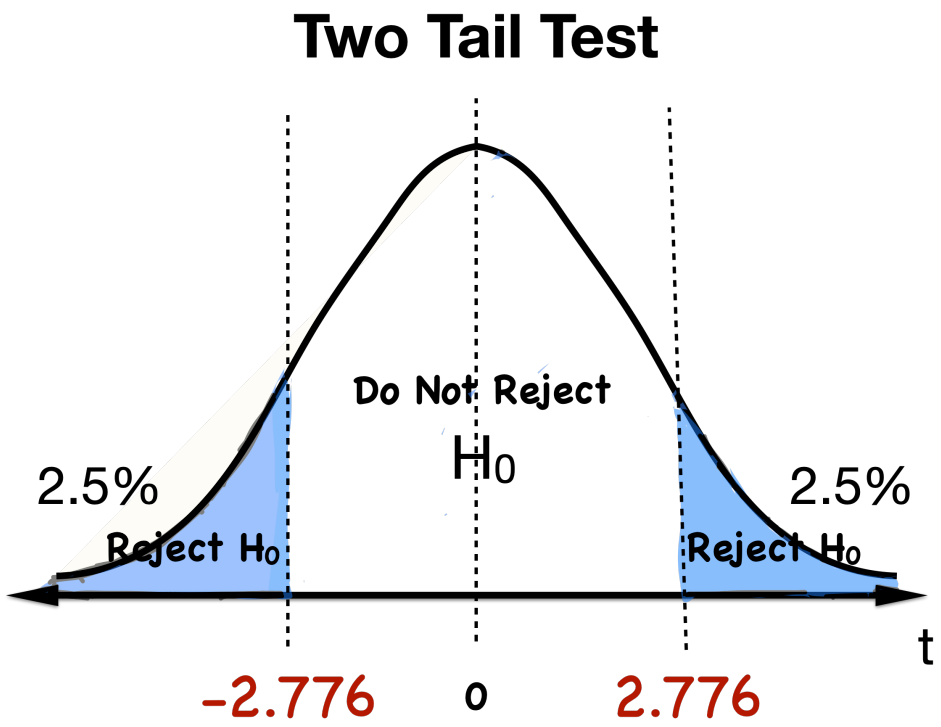
The bivariate sample data below are the study times (minutes) and the grades (percent) earned on an exam of randomly selected grades. Use the 5% level of significance to test the claim that there is a linear correlation between the time studied and the grade earned on an exam.

$\boldsymbol{\alpha = 0.05}$ and $\boldsymbol{df = 4}$ and $\boldsymbol{critical\ values = \pm 2.776}$

Study Time (Min)	Grade (percent)
20	40
40	45
50	70
60	76
80	92
100	95

$H_0: \rho = 0$ (No Linear Correlation)

$H_1: \rho \neq 0$ (Linear Correlation) *Claim*



$r \approx$

$t \approx$

$p \approx$

Conclusion

What is the Best Fit (Regression) line?

Problem 2

Customer Service and Computer Sales

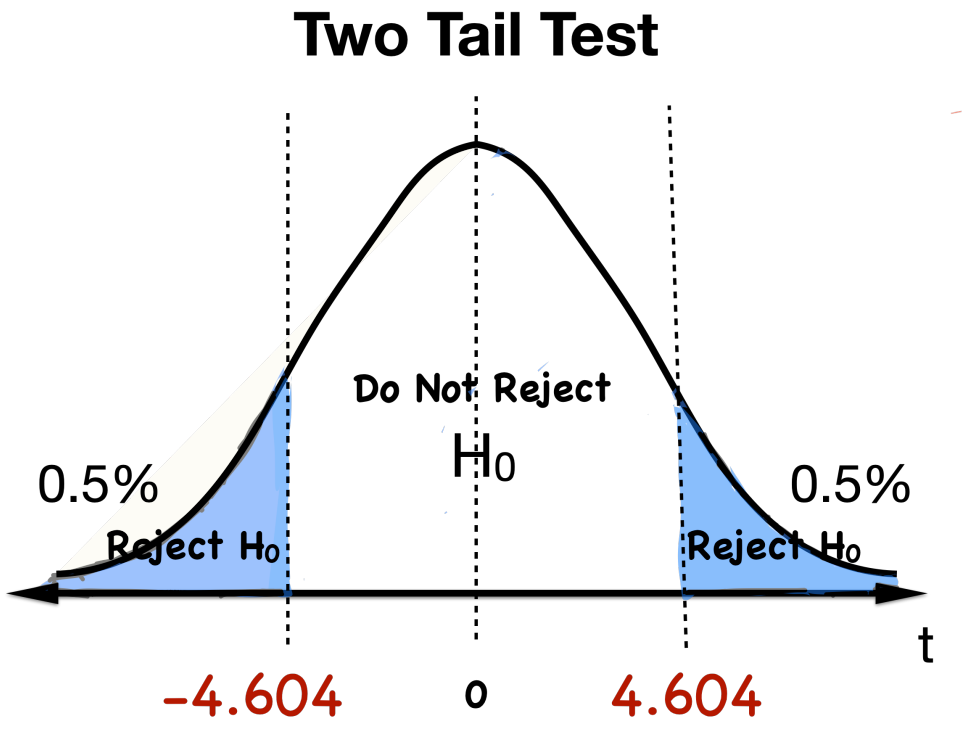
The bivariate sample data below are the number of customer service calls (per week) and the computer sales (thousand dollars) at a computer wholesale store. Use the 1% level of significance to test the claim that there is a linear correlation between the number of calls (week) and the computer sales (one thousand dollars).

$\boldsymbol{\alpha = 0.01}$ and $\boldsymbol{df = 4}$ and $\boldsymbol{critical\ values = \pm 4.604}$

# of calls (week)	Cost (1K Dollars)
10	37
15	43
12	37
20	49
25	54
17	45

$H_0: \rho = 0$ (No Linear Correlation)

$H_1: \rho \neq 0$ (Linear Correlation) $Claim$



$r \approx$

$t \approx$

$p \approx$

Conclusion

What is the Best Fit (Regression) line?

Problem 3

Police Officers and Muggings

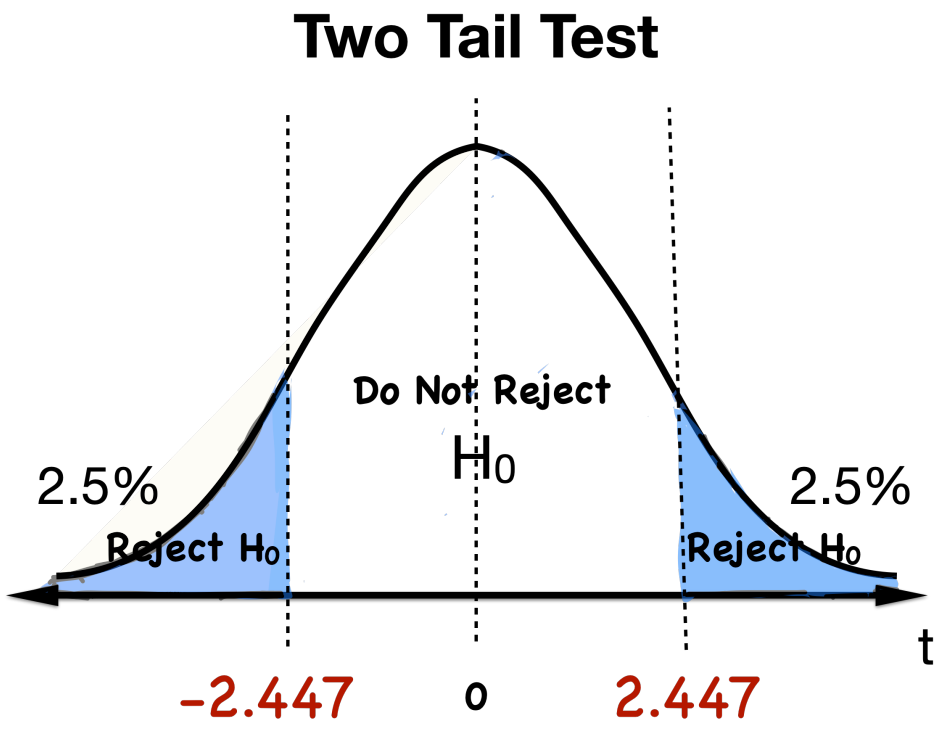
The bivariate sample data below is a sample of the number of police officers on patrol in a small city and the number of muggings in the same sized city. Use the 5% level of significance to test the claim that there is a linear correlation between the number of police officers on patrol and the number of muggings.

$\boldsymbol{\alpha = 0.05}$ and $\boldsymbol{df = 6}$ and $\boldsymbol{critical\ values = \pm 2.447}$

Police Officers	Muggings
20	8
12	10
18	12
15	9
22	6
10	15
20	7
12	18

$H_0: \rho = 0$ (No Linear Correlation)

$H_1: \rho \neq 0$ (Linear Correlation) *Claim*



$r \approx$

$t \approx$

$p \approx$

Conclusion

What is the Best Fit (Regression) line?

Problem 4

Weeks and Weight Loss

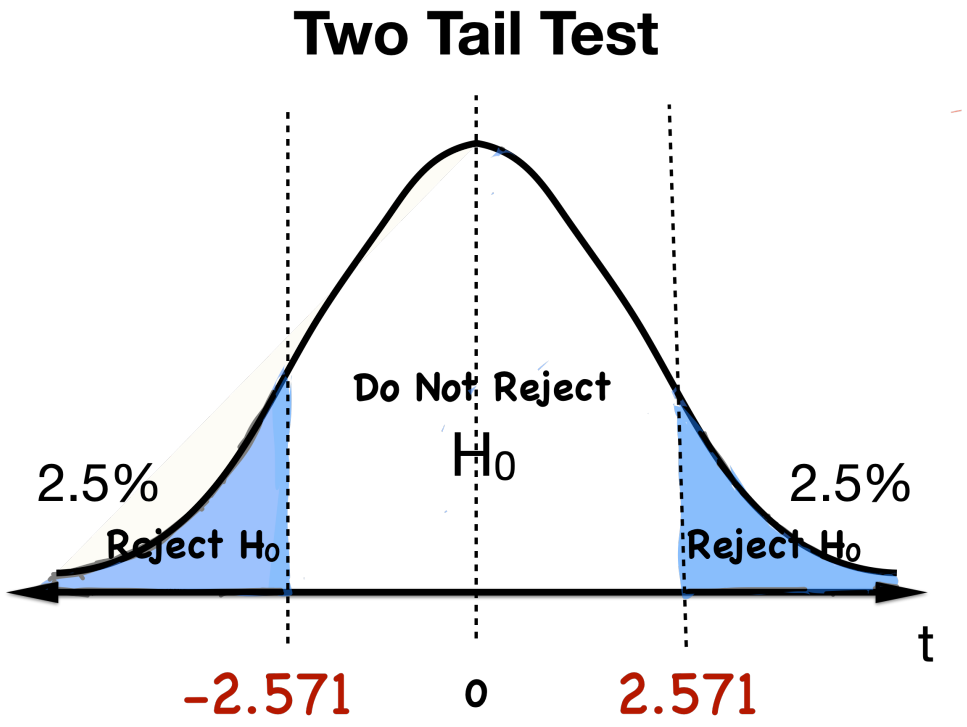
The bivariate sample data below represent the number of weeks a person is on "lose weight quick" pills and the cumulative weight loss in pounds. Use the 5% level of significance to test the claim that there is a linear correlation between the two quantities.

$\boldsymbol{\alpha = 0.05}$ and $\boldsymbol{df = 5}$ and $\boldsymbol{critical\ values = \pm 2.571}$

# of weeks	Cumulative Weight Loss (pounds)
1	2
2	5
3	5
4	6
5	5
6	7
7	9

$H_0: \rho = 0$ (No Linear Correlation)

$H_1: \rho \neq 0$ (Linear Correlation) $Claim$



$r \approx$

$t \approx$

$p \approx$

Conclusion

What is the Best Fit (Regression) line?

Problem 5

Age and Miles Run

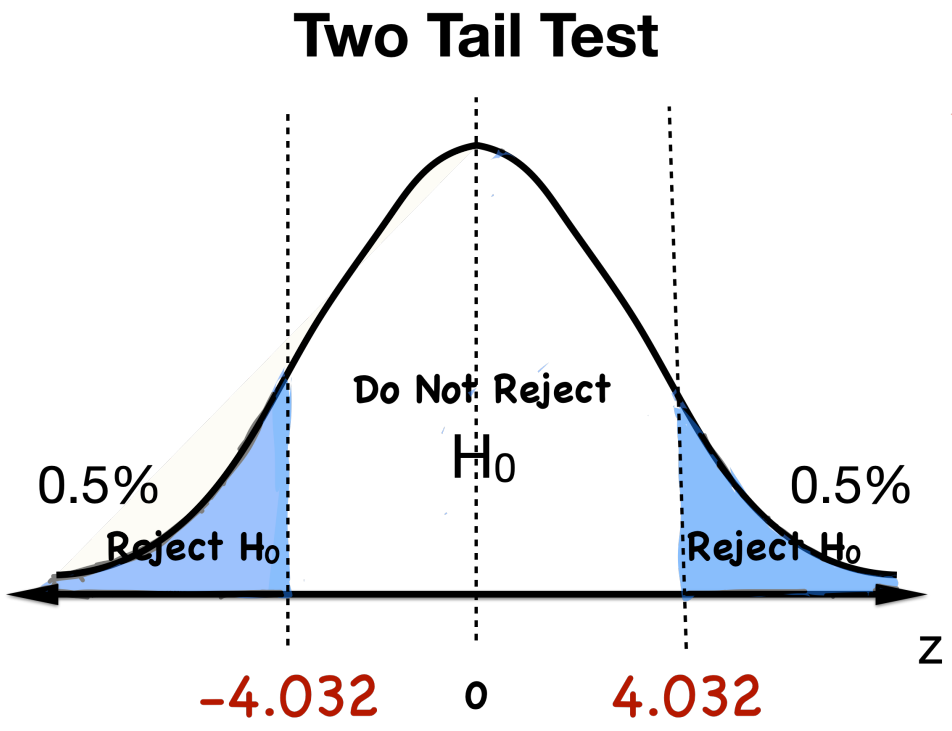
The bivariate sample data below represent the ages of people in years and the number of miles they run per week. Use the 1% level of significance to test the claim that there is a linear correlation between the two quantities.

$\boldsymbol{\alpha = 0.01}$ and $\boldsymbol{df = 5}$ and $\boldsymbol{critical\ values = \pm 4.032}$

Age	Miles Run (per week)
18	6
34	8
50	4
18	4
25	10
62	12
20	6

$H_0: \rho = 0$ (No Linear Correlation)

$H_1: \rho \neq 0$ (Linear Correlation) *Claim*



$r \approx$

$t \approx$

$p \approx$

Conclusion

What is the Best Fit (Regression) line?

Problem 6

Pizza and Subway Fare (Pizza Connection)

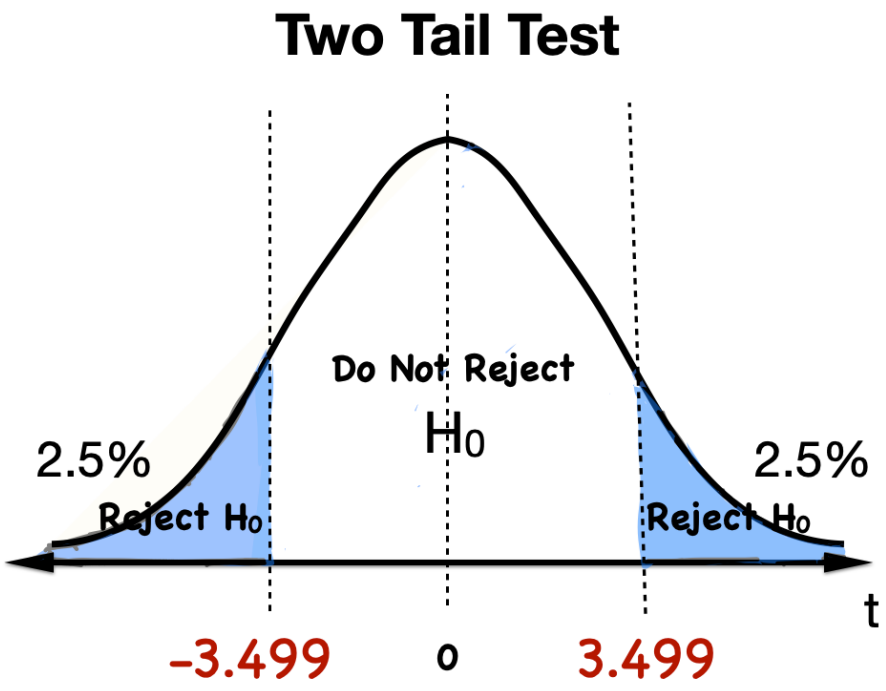
The bivariate sample data below represent the cost of a slice of pizza and the price of a subway fare in New York City. Use the 1% level of significance to test the claim that there is a linear correlation between the two quantities.

$\boldsymbol{\alpha = 0.01}$ and $\boldsymbol{df = 7}$ and $\boldsymbol{critical\ values = \pm 3.500}$

Year	1960	1973	1986	1995	2002	2003	2009	2013	2015
Pizza Cost	0.15	0.35	1	1.25	1.75	2	2.25	2.3	2.75
Subway Fare	0.15	0.35	1	1.35	1.5	2	2.25	2.5	2.75

$H_0: \rho = 0$ (No Linear Correlation)

$H_1: \rho \neq 0$ (Linear Correlation) $Claim$



$r \approx$

$t \approx$

$p \approx$

Conclusion

What is the Best Fit (Regression) line?

Regression Line

Now that we know that there is a linear correlation between the two quantities x and y, what **linear equation** can we use to represent this linear correlation, so that we can be predictive with these two quantities? The linear equation is known as the **Regression Line** and is of the form $\boldsymbol{y = mx = b}$ from beginning algebra. In beginning algebra, the slope value is m, and the y-intercept value is b.

This **Regression Line** is a straight line that "**best fits**" our bivariate data in our scatter plot and consists of an independent variable x and the independent variable y. In statistics, we describe this equation as follows.

$\hat{y} = b_0 + b_1 x$ which comes from a bivariate sample data and is a sample statistic.

$y = \beta_0 + \beta_1 x$ which comes from a census and is a population parameter.

The slope formula

$$b_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

The y-intercept formula

$$b_0 = \frac{\sum y \sum x^2 - \sum x \sum xy}{n\sum x^2 - (\sum x)^2}$$

Note- Round b_0 and b_1 to three significant digits.

Def- Standard Deviation of the x-values is notated $\mathbf{S_x}$ and is given by $s_x = \sqrt{n\sum x^2 - (\sum x)^2}$

Def- Standard Deviation of the y-values is notated $\mathbf{S_y}$ and is given by $s_y = \sqrt{n\sum y^2 - (\sum y)^2}$

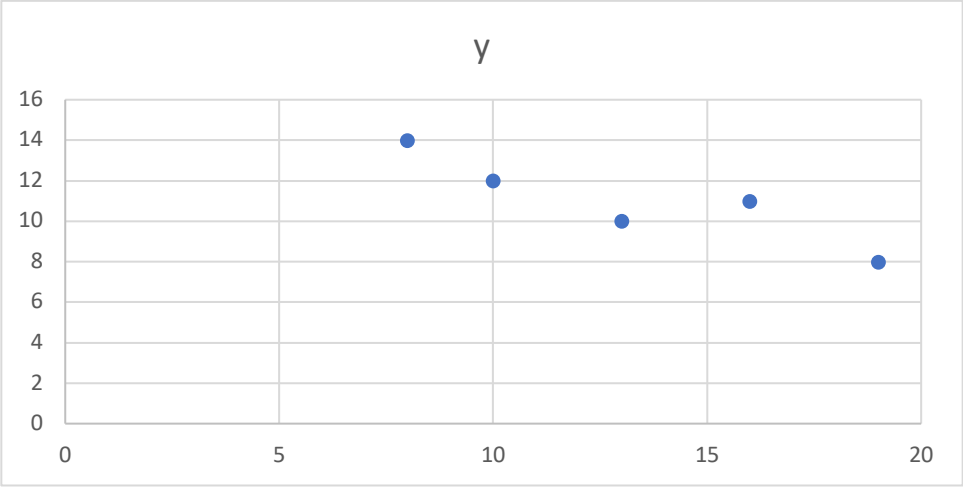
Option 2

The slope can also be represented by $\boldsymbol{b_1 = r\frac{s_y}{s_x}}$ and the y-intercept can also be represented by $\boldsymbol{b_0 = \bar{y} - b_1\bar{x}}$

Example 1 (Excel Spreadsheet) Bivariate data

x	8	10	13	16	19
y	14	12	10	11	8

Determine the scatter plot



Since my data is somewhat linear, determine the sums of the following columns.

	x	y	xy	x^2	y^2
	8	14	112	64	196
	10	12	120	100	144
	13	10	130	169	100
	16	11	176	256	121
	19	8	152	361	64
Sum	66	55	690	950	625

$$r = \frac{\boldsymbol{n\sum xy - \sum x \sum y}}{\sqrt{\boldsymbol{n\sum x^2 - (\sum x)^2}}\sqrt{\boldsymbol{n\sum y^2 - (\sum y)^2}}} \; ; \quad r = \frac{5*690-66*55}{\sqrt{5*950-66^2}\sqrt{5*625-55^2}} \; ;$$

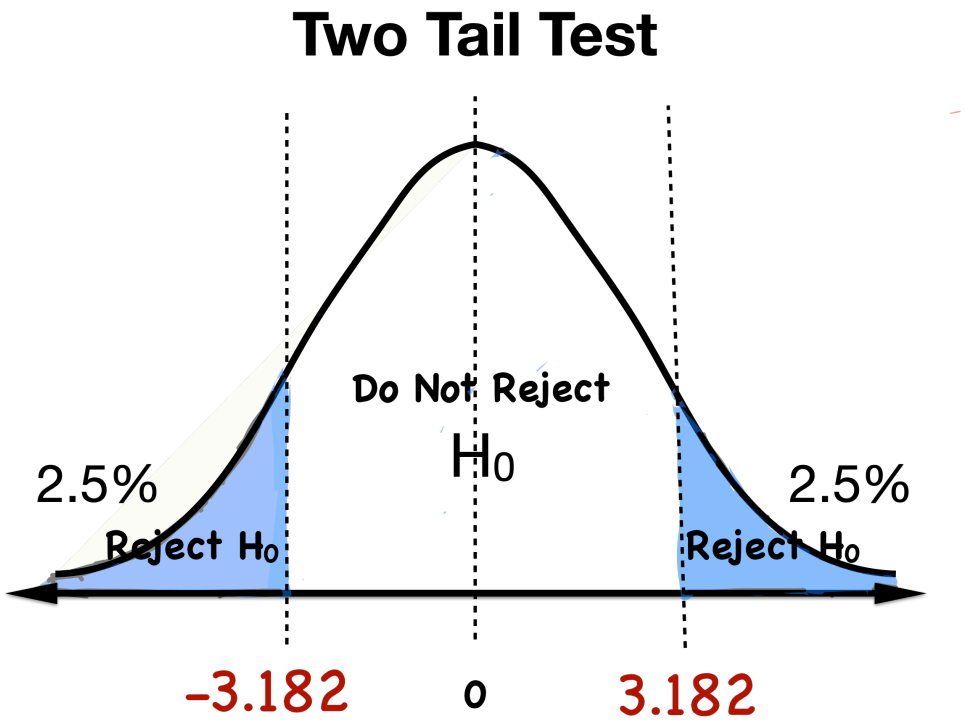
$$r \approx -0.907$$

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} ; \quad t \approx \frac{-.907}{\sqrt{\frac{1-(-.907)^2}{5-2}}} \text{ or } t \approx -3.730$$

Determine whether there is a linear correlation using the 5% level of significance.

$H_0: \rho = 0$ (No Linear Correlation)

Claim $H_1: \rho \neq 0$ (Linear Correlation)



$r \approx$ -0.907

$t \approx$-3.730

Conclusion: Reject the null hypothesis and there is sufficient evidence the sample supports the claim that there is a linear correlation between the two variables.

What is the Regression Line that “best fits” our bivariate data?

$\hat{y} = b_0 + b_1 x$ such that $b_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$ and $b_0 = \frac{\sum y \sum x^2 - \sum x \sum xy}{n\sum x^2 - (\sum x)^2}$

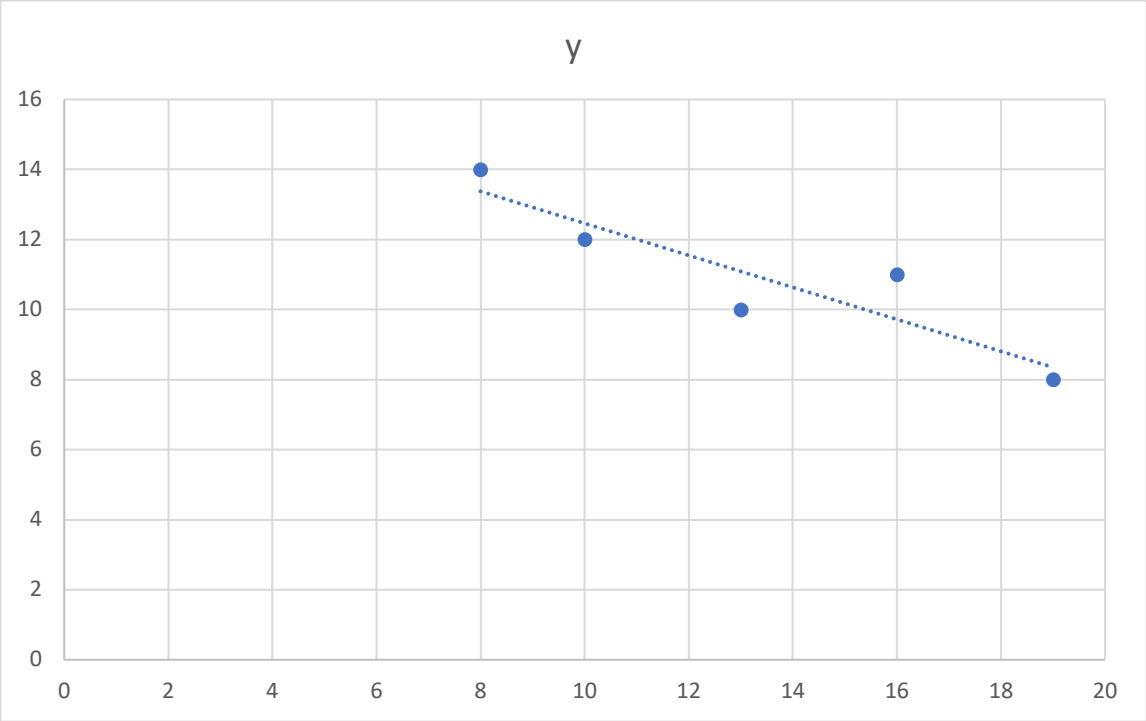
$$b_1 = \frac{5 * 690 - 66 * 55}{5 * 950 - 66^2} \approx -0.457$$

$$b_0 = \frac{55 * 950 - 66 * 690}{5 * 950 - 66^2} \approx 17.030$$

$$\hat{y} = 17.030 - 0.457x$$

x	8	10	13	16	19
y	14	12	10	11	8

The **Regression Line** is $\hat{y} = \mathbf{17.030} - \mathbf{0.457}x$ which can be used to **predict** a y-value given any x-vale.



If we let $x = 9$; $x = 12$; $x = 14$; $x = 22$ we can find the corresponding y values.

$\hat{y} = 17.030 - 0.457 * 9$; $\hat{y} \approx 12.917$

$\hat{y} = 17.030 - 0.457 * 12$; $\hat{y} \approx 11.546$

$\hat{y} = 17.030 - 0.457 * 14$; $\hat{y} \approx 10.632$

$\hat{y} = 17.030 - 0.457 * 22$; $\hat{y} \approx 6.976$

Using Option 2

Option 2

The slope can also be represented by $\boldsymbol{b_1 = r\frac{s_y}{s_x}}$ and the y-intercept can also be represented by $\boldsymbol{b_0 = \bar{y} - b_1\bar{x}}$

x	y	xy	x^2	y^2
8	14	112	64	196
10	12	120	100	144
13	10	130	169	100
16	11	176	256	121
19	8	152	361	64

Mean	13.2	11
SD	4.4385	2.2361
r	-0.907	

$$b_1 = r\frac{s_y}{s_x} \text{ or } b_1 = -0.907\frac{2.2361}{4.4385} \approx -0.457$$

$$b_0 = \bar{y} - b_1\bar{x} \text{ or } b_1 = 11 - (-0.4569)13.2 \approx 170.031$$

so that $\boldsymbol{\hat{y} = 17.031 - 0.457x}$