Section 4H – Quantitative Relationships: The Correlation Test

We saw in the last section, that two quantitative samples are related if their correlation coefficient (r) is close to 1 or −1. When the correlation coefficient (r) is close to zero, the two quantitative samples are not related. How does this apply to populations? What if we want to determine if there is a relationship between two quantitative variables in a population? For this, we will need to look at the correlation hypothesis test.

The Correlation Hypothesis Test

If the sample correlation coefficient (r) is zero, tells us that the two quantitative samples are not related. For populations, we need to look at the population correlation coefficient "rho" (ρ). While this looks like a "p", it is not. It is the Greek letter "rho" and represents the population correlation coefficient.

If you recall from the last section, the correlation coefficient is related to the slope of the regression line.

Sample Slope $b_1 = \frac{(correlation\ coefficient\ times\ standard\ deviation\ of\ the\ y\ values\ )}{standard\ deviation\ of\ the\ x\ values} = \frac{(r \times S_y)}{S_x}$

So if there is no correlation between variables the correlation coefficient and the slope both go to zero. This principle applies to populations as well. As the population correlation coefficient "rho" (ρ) goes to zero, the population slope "Beta 1" (β_1) also goes to zero.

Correlation Test Null and Alternative Hypothesis

There are several ways of writing the null and alternative hypothesis for a correlation hypothesis test. We can use the population correlation coefficient "rho" (ρ) or the population slope "Beta 1" (β_1). We can also specify positive or negative correlation. Remember the correlation coefficient and the slope always have the same sign.

To show positive correlation the correlation coefficient should be close to +1 (greater than zero) and the slope should also be significantly positive (greater than zero). A positive relationship is also called a "direct" relationship. As the X variable increases, the Y variable also tends to increase. As the X variable decreases, the Y variable also tends to decrease.

To show negative correlation the correlation coefficient should be close to −1 (less than zero) and the slope should also be significantly negative (less than zero). A negative relationship is also called an "indirect" or "inverse" relationship. As the X variable increases, the Y variable also tends to decrease. As the X variable decreases, the Y variable also tends to increase.

Note about Statcato: Statcato only has the option for the two-tailed correlation test and cannot specify positive correlation (right-tailed) or negative correlation (left-tailed) hypothesis tests.

Two-Tailed Correlation Test: *For determining if variables are related or not. Does not specify if the direction is positive or negative.*

H_0 : ρ = 0 *(The two quantitative variables in the population are not related.)*
H_A : ρ ≠ 0 *(The two quantitative variables in the population are related.)*

OR

H_0 : β_1 = 0 *(The two quantitative variables in the population are not related.)*
H_A : β_1 ≠ 0 *(The two quantitative variables in the population are related.)*

Right-Tailed Correlation Test: *For determining if variables have a positive (or direct) relationship or not. Notice the alternative hypothesis symbol ">" points to the right.*

H_0 : ρ = 0 *(The two quantitative variables in the population are not related.)*
H_A : ρ > 0 *(The two quantitative variables in the population have a positive (direct) relationship.)*

OR

H_0 : β_1 = 0 *(The two quantitative variables in the population are not related.)*
H_A : β_1 > 0 *(The two quantitative variables in the population have a positive (direct) relationship.)*

Left-Tailed Correlation Test: *For determining if variables have a negative (or inverse) relationship or not. Notice the alternative hypothesis symbol "<" points to the left.*

H_0 : ρ = 0 *(The two quantitative variables in the population are not related.)*
H_A : ρ < 0 *(The two quantitative variables in the population have a negative (inverse) relationship.)*

OR

H_0 : β_1 = 0 *(The two quantitative variables in the population are not related.)*
H_A : β_1 < 0 *(The two quantitative variables in the population have a negative (inverse) relationship.)*

T-test statistic

The relationship between correlation and the slope of the regression line is highlighted in the test statistic. For a correlation test, you can use either the correlation coefficient "r" or a T-test statistic. I prefer the T-test statistic. The null hypothesis is that there is not a relationship between the quantitative variables. This would indicate that that the correlation coefficient and the slope would be close to zero. So the T-test statistic counts how many standard error the slope is from zero. If the T-test statistic is positive, the slope will be above zero and if the T-test statistic is negative, the slope will be below zero.

T-test statistics for correlation = $\frac{(sample\ slope - 0)}{standard\ error\ for\ slope}$

T-test statistics sentence for Correlation: The number of standard errors that the slope of the regression line is above or below zero.

As with all test statistics, we will want to see if the T-test statistic falls in a tail determined by the critical value or values. If so, the sample data significantly disagrees with the null hypothesis and the slope is significantly different from zero. If the T-test statistic does not fall in a tail determined by the critical value or values, then the sample data does not significantly disagree with the null hypothesis and the slope is not significantly different from zero.

Residual Errors

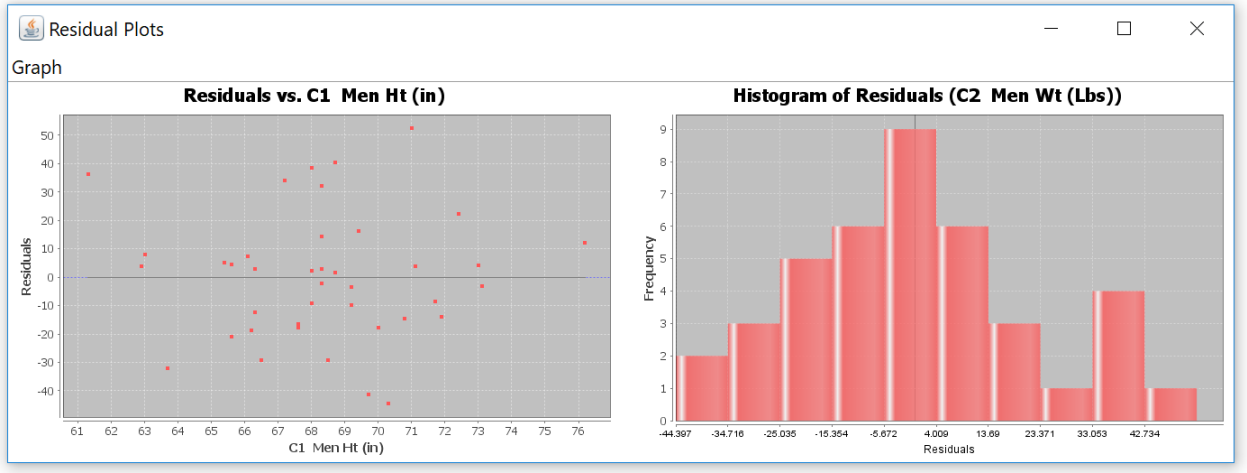
The correlation test has many assumptions. Some of the assumptions are centered on the understanding of "residuals" or "residual errors". We learned in the last section that a residual is the vertical distance between each point in the scatterplot and the regression line. To calculate a residual, the computer subtracts the actual y coordinate of the point minus the predicted ŷ value on the regression line. We also saw that the average of all the residuals is called the standard deviation of the residual errors (S_e). This tells us the average vertical distance that the data is from the regression line and the average prediction error.

Residual = $y - \hat{y}$

$S_e = \sqrt{\frac{\sum(y - \hat{y})^2}{n-2}}$

Besides the standard deviation of the residual errors, there are also residual graphs that statisticians often like to examine when doing a correlation test. We will only look at two. They are the "histogram of the residuals errors" and the "residual plot verses the x-values".

Here is an example. These graphs were created with Statcato. The explanatory variable (X) was the height of men and the response variable (Y) was the weight of men.



Residual Plot

The graph on the left is called the "residual plot verses the x-variable". This graph shows the vertical distances that each point is from the regression line. A point that is 40 above the line will have a residual of +40. A point that is 19 below the line will have a residual of −19. The zero line represents the regression line since points on the regression line have a residual of zero. We want the residual plot to be evenly spread out. When the points are evenly spread out, our standard deviation of the residuals is a consistent measure of spread. When the residual plot is not evenly spread out, you will see parts of the x-axis where all the points are very close and other parts of the x-axis where the points are very far away. This is an uneven spread (or fan shaped). We want the standard deviation to be a consistent measure of spread for all x value in the scope. If the points are close for some x values, then the standard deviation will be an overestimate of the variability for those x values. Similarly, if the points are far away for other x values, then the standard deviation will be an underestimate of the variability for those x values. Residual plots can be very difficult to read. I tell my intro students to put all the points on the left side of the graph between your fingers. Now put all the points on the right side of the graph between your fingers. If your fingers are about the same width on both the left and right side, you are probably ok. The data is evenly spread out and the standard deviation is a consistent measure of spread (variability). If your fingers are much closer on one side than the other, that may indicate a fan shape or uneven spread. In that case, the standard deviation is not a consistent measure of variability. Notice that points with an x-value greater than 72 are much closer to the regression line than those below 72. This could indicate an uneven spread (fan shape). This also could indicate that the regression line predictions are more accurate for taller men in the data (over 72 inches) and less accurate for shorter men in the data.

Histogram of the Residuals

The graph on the right is called the "histogram of the residuals". Remember that the calculation of the regression line uses the mean and standard deviation. If you remember from previous chapters, the mean and standard deviations are only accurate for normal data. We could check the shape of each data set separately, but instead we prefer to check the shape of the residuals. The histogram of the residuals should be normal (bell shaped). It should also be centered close to zero. Statcato gives a dark vertical line at zero for this purpose. This line should be close to the highest bar in the histogram. This histogram above passes both criteria.
Let us look at the assumptions for a correlation test.

Correlation Test Assumptions

1. The quantitative ordered pair data should be collected randomly or be representative of the population. *(The two samples usually have different units, but must have a one-to-one pairing).*

2. Data values within the sample should be independent of each other. *(The two samples are not independent since they are ordered pair. The individual data values within each sample should be independent. If you have small simple random sample from a large population, then the data values are probably not related.)*

3. The sample size should be at least 30. *(There should be 30 or more ordered pairs.)*

4. The scatterplot and correlation coefficient (r) should show some linear pattern. *(The correlation coefficient (r) should not be close to zero.)*

5. There should be no influential outliers in the scatterplot. *(If your correlation coefficient is close to 1 or −1, then you probably have no influential outliers. Remember to look for outliers on the scatterplot. A residual plot magnifies the distances, so everything looks like an outlier in a residual plot.)*

6. The histogram of the residuals should be nearly normal.

7. The histogram of the residuals should be centered close to zero. *(The zero line should be touching the highest bar in the histogram, or at least very close to the highest bar.)*

8. The residual plot verses the x variables should be evenly spread out with no fan shape or sideways "V" pattern. *(Put all the points in the residual plot between your fingers on the left side of the graph. Now put all the points in between your fingers on the right side. If your fingers are about the same width apart on the left and right side, the graph is close to evenly spread out.)*

Correlation Test Example 1

Let us use Statcato and the random "Health" data at www.matt-teachout.org to test the claim that there is no relationship between the age of a man and his cholesterol. In the Health data, we have the ages and cholesterol of forty randomly selected men. We will designate the age to be the explanatory variable (X) and the cholesterol to be the response variable (Y). Let us use a 5% significance level.

We can write the null and alternative hypothesis in one of two ways. We can use the population correlation coefficient "rho" (ρ) or the population slope "beta 1" (β_1). Remember our claim is "not related" so that must be the null hypothesis. Since positive or negative relationship was not mentioned, we will assume this is the general two-tailed test.

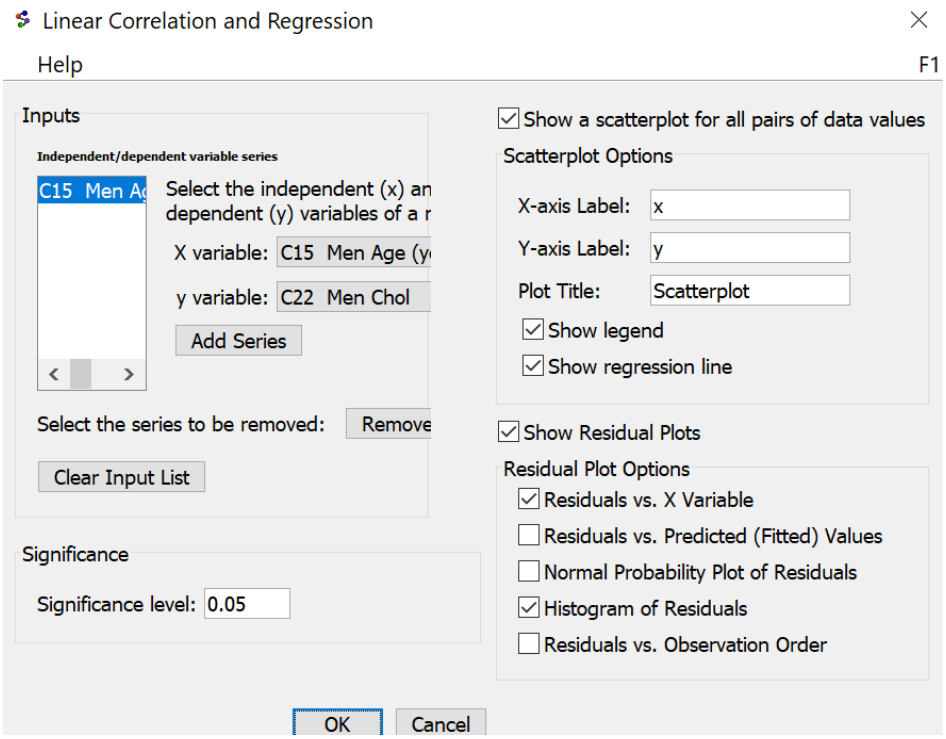
H_0 : ρ = 0 *(The age and cholesterol of men are not related.) CLAIM*
H_0 : ρ ≠ 0 *(The age and cholesterol of men are related.)*

OR

H_0 : β_1 = 0 *(The age and cholesterol of men are not related.) CLAIM*
H_0 : β_1 ≠ 0 *(The age and cholesterol of men are related.)*

Copy and paste the men's age and cholesterol data into two columns of Statcato. Go to the "statistics" menu, click on "correlation and regression" and then click on "linear". Click on the men's age to be the X-variable and the men's cholesterol to be the Y-variable and then push "add series". Check the box that says "show scatterplot" and the box that says "show regression line". Statcato also has the capability of making residual plots. Check the box that says, "Show residual plots", the box that says "residuals vs x-variable", and the box that says "histogram of the residuals". Now push "OK". Here is the Statcato printout, with the test statistic, P-value, correlation coefficient and all of the graphs.

Note: Some versions of Statcato do not have residual plots.



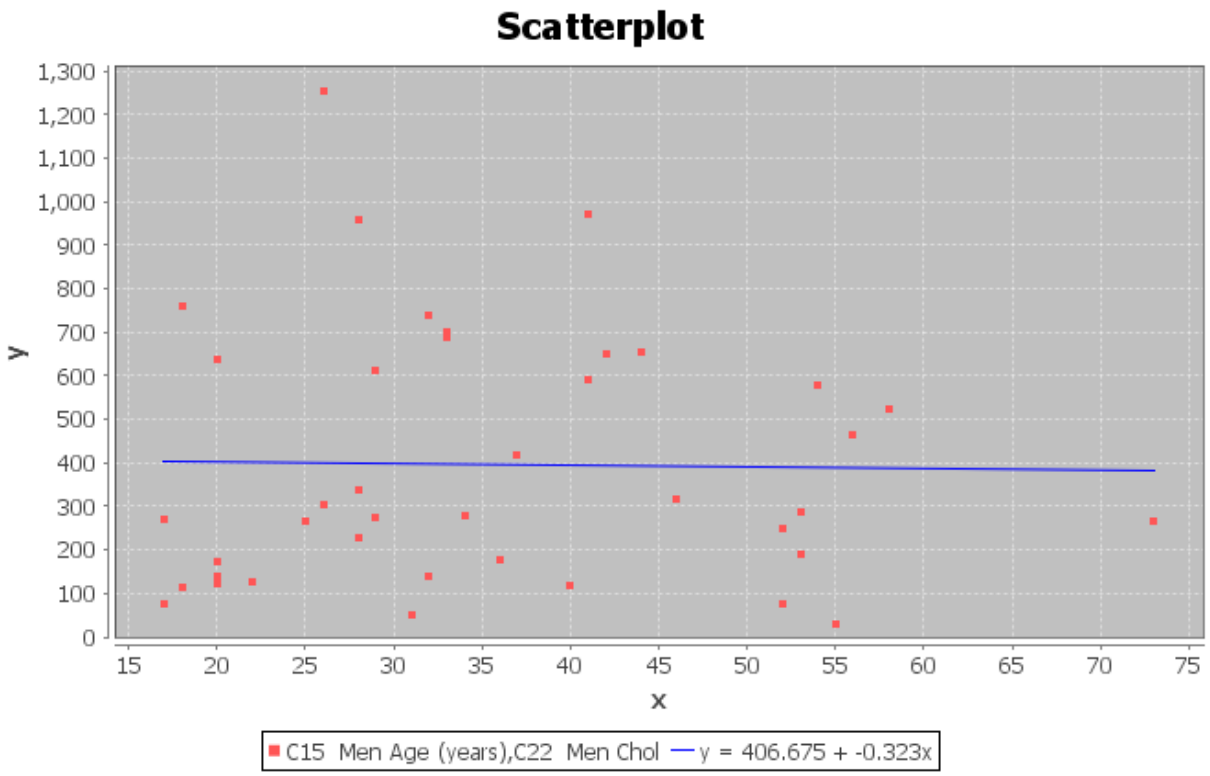
Correlation and Regression: Significance level = 0.05

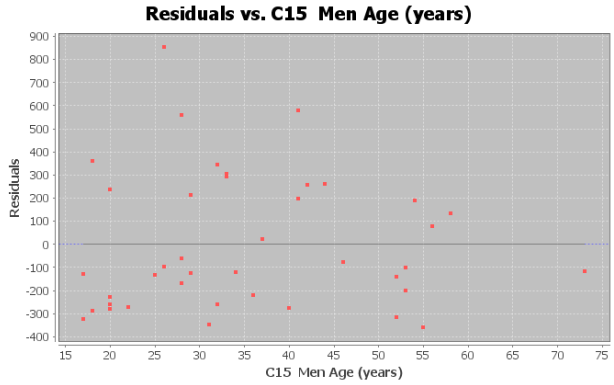
Series: C15 Men Age (years),C22 Men Chol
x = C15 Men Age (years)
y = C22 Men Chol
Sample size n = 40
Degrees of freedom = 38

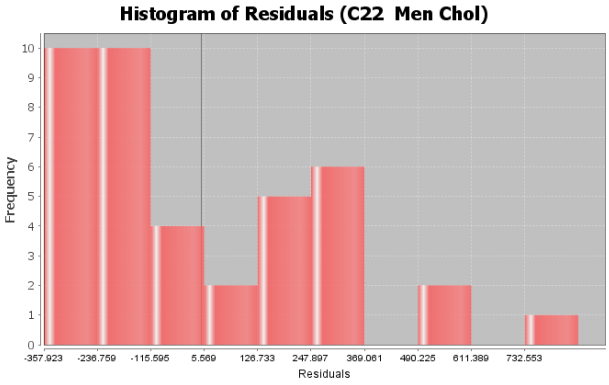
Correlation:
H_0: $\rho = 0$ (no linear correlation)
H_1: $\rho \neq 0$ (linear correlation)

	Test Statistic	Critical Value
r	-0.0154	±0.3120
t	-0.0948	±2.0244

p-Value = 0.9250







Let us start by checking the assumptions for the men's age and cholesterol problem. Notice that this data fails many of the assumptions. That means our hypothesis test is compromised. We should also not use this regression line to make predictions about men's cholesterol.

1. Two quantitative ordered pair random samples. Yes. Age and cholesterol are both quantitative. The data had randomly selected men with the age and cholesterol of each man. It is ordered pair data.

2. Data values within each sample should be independent of each other. Yes. Since there is only forty randomly selected men out of millions of men in the population, the men are not likely to be related.

3. The sample size should be at least 30. Yes. There was forty men in the data. This is greater than thirty.

4. The scatterplot and correlation coefficient (r) should show some linear pattern. No. The regression line does not seem to fit the points in the scatterplot at all and the correlation coefficient r is very close to zero.

5. There should be no influential outliers in the scatterplot. No. There seem to be many influential outliers in the scatterplot and the correlation coefficient r is very close to zero.

6. The histogram of the residuals should be nearly normal. No. The histogram of the residuals is skewed right and not normal.

7. The histogram of the residuals should be centered close to zero. No. The histogram of the residuals seems to be centered to the left of zero. The zero line is not touching the highest bar in the histogram.

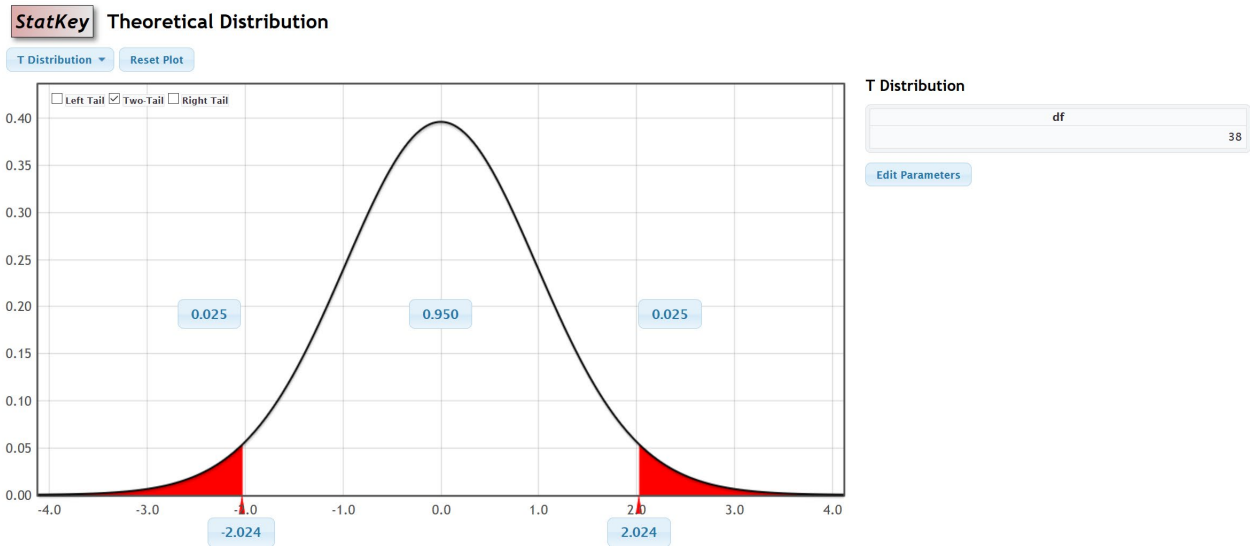
8. The residual plot verses the x variables should be evenly spread out. No. The residual plot seems to show a distinct fan shape and is not evenly spread out. The points on the left side of the graph seem to have a very wide spread while the points on the right side of the graph seem to be very close.

Test Statistic: T = −0.0948

Sentence: The slope of the regression line is 0.0948 standard errors below zero.

Our T-test statistic is −0.0948 and does not fall in either of the tails determined by the critical values. Our random sample data does not significantly disagree with the null hypothesis. This also indicates the slope is not significantly different from zero.

We put the degrees of freedom 38 into the theoretical T distribution calculator in StatKey to get the following picture.



↑
T = −0.0948

P-value = 0.9250

Sentence: If the null hypothesis is true and there is no relationship between the age and cholesterol for men, then there is a 92.5% probability of getting the sample data or more extreme because of sampling variability.

Notice the P-value is greater than our 5% significance level. This indicates that the sample data or more extreme could have occurred because of sampling variability if the null hypothesis was true. Since sampling variability cannot be ruled out, we must fail to reject the null hypothesis.

Fail to reject the Null Hypothesis.

We have a high P-value and the null hypothesis is the claim. The sample data did not pass all of the assumptions for the correlation test.

Conclusion: There is not significant evidence to reject the claim that the age and cholesterol of men is not related.

Age and cholesterol of men are probably not related. This sample data did not provide evidence since the P-value was high and it failed many of the assumptions for the correlation test.

Example 2

We can also use randomized simulation on StatKey to determine significance and calculate the P-value. StatKey can calculate the scatterplot and the correlation coefficient and slope, but does not calculate any of the residual graphs.

We are going to be using the "mpg weight horsepower" data on www.matt-teachout.org to test the claim that there is a negative (inverse) relationship between the weight of a car and the miles per gallon of gas (mpg). We will be using a 5% significance level and assume the data met all of the assumptions.

H_0 : ρ = 0 *(The weight and mpg of a car are not related.)*
H_A : ρ < 0 *(The weight and mpg of a car have a negative (inverse) relationship.) CLAIM*

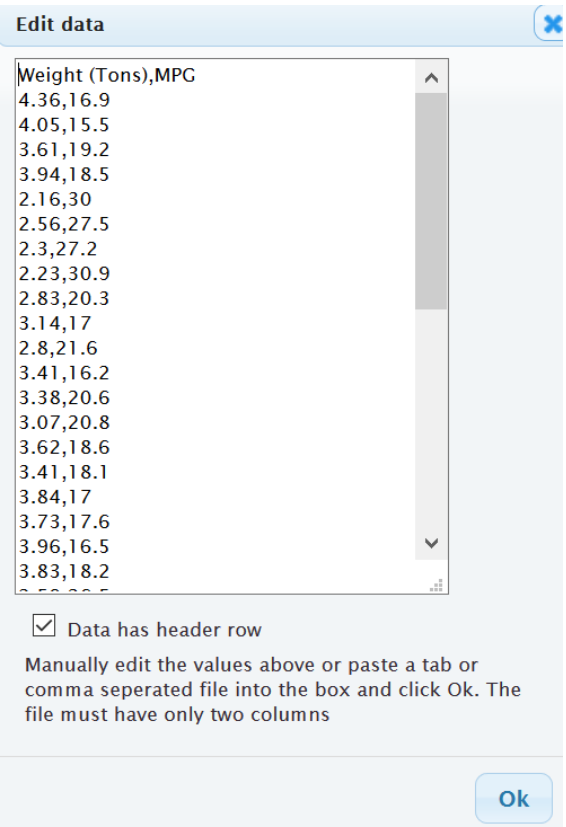
OR

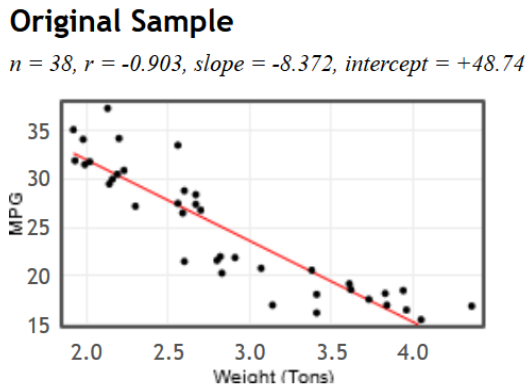
H_0 : β_1 = 0 *(The weight and mpg of a car are not related.)*
H_A : β_1 < 0 *(The weight and mpg of a car have a negative (inverse) relationship.) CLAIM*

We will designate the weight of the car as the explanatory variable (X) and the miles per gallon as the response variable (Y). Copy and paste the weight of the cars and mpg into a fresh excel spreadsheet. Put the weight data on the left and the mpg on the right. Now copy both columns together.

Weight (Tons)	MPG
4.36	16.9
4.05	15.5
3.61	19.2
3.94	18.5
2.16	30
2.56	27.5
2.3	27.2
2.23	30.9

Go to www.lock5stat.com and open StatKey. Under the "Randomized Hypothesis Tests" menu, click on "Test for Slope, Correlation". Under the "Edit Data" menu, paste in the weight and mpg columns. Since the data sets have titles, check the box that says, "Header has header row" and push OK. Under "Original Sample" we see the scatterplot, correlation coefficient (r) and the sample slope (b_1).





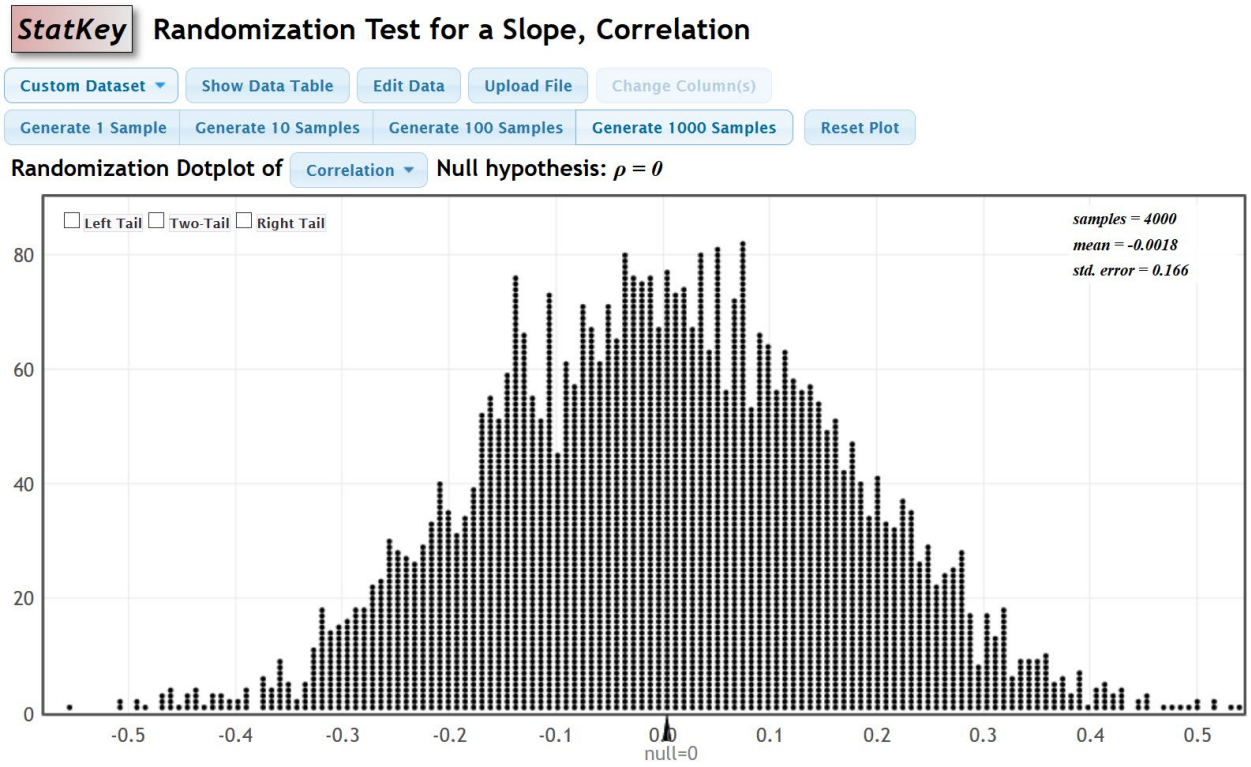
Let us give a quick analysis of the sample data as we did in the last section. We see that the scatterplot and the correlation coefficient (r) show a strong negative relationship between the samples. Notice that r = −0.903 and is close to −1. The points in the scatterplot seem to be close to the regression line and there does not appear to be any influential outliers.

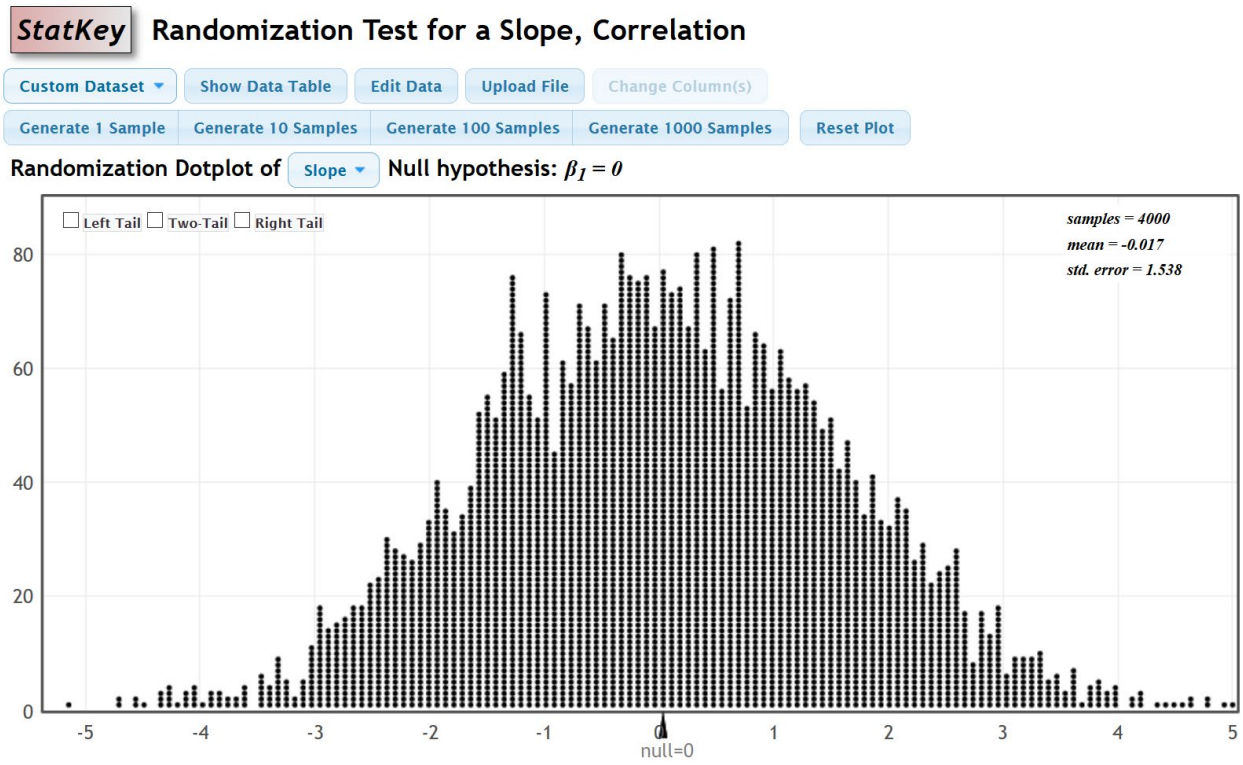
The slope is −8.372. In our last section, we saw that the slope is the amount of increase or decrease in the Y variable per unit of X. Since the slope was negative, it is a decrease. In addition, the X variable is the weight of the car in tons and the Y variable is the gas mileage in miles per gallon.

Sample Slope Sentence: For every 1 ton heavier the car, the average miles per gallon of the cars in the samples are decreasing 8.372 mpg.

Randomized Simulation

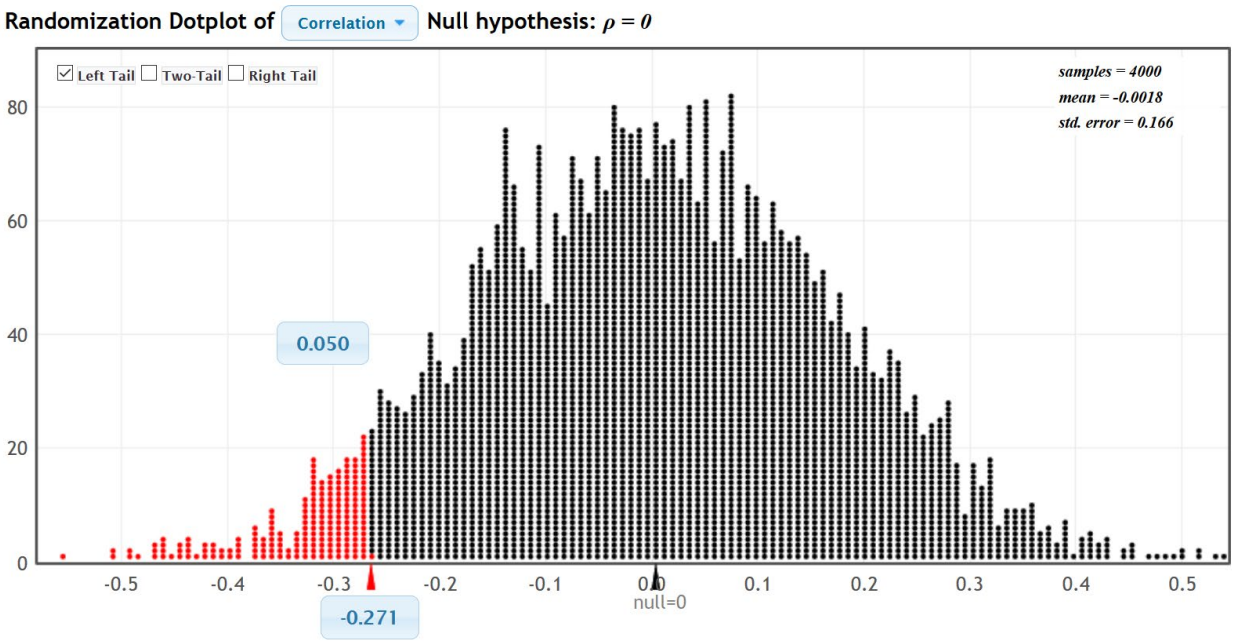
There are two ways to do the randomized simulation. We can have the computer create thousands of random samples and calculate the correlation coefficient for each. Another way is to have the computer create thousands of random samples and calculate the slope for each. At the top of the distribution, you will see we can change the setting to "correlation" or "slope". Notice how the null hypothesis changes to reflect the setting. Click on "Generate 1000 Samples" a few times.





Simulating with the Correlation Coefficient

Let us start with looking at the correlation coefficient simulation. These are thousands of correlation coefficients. When the setting is on "Correlation", we will need to use the "Original Sample" correlation coefficient (r) to determine significance and calculate the P-value. Since the alternative hypothesis was less than "<", this was a left-tailed test. Click on left tail. Since we are using a 5% significance level, we will put in 0.05 in the left tail proportion. Notice the simulation indicates that our "Original Sample" correlation coefficient (r) needs to be −0.271 or less to be significant. Our "Original Sample" correlation coefficient (r) is −0.903 and definitely falls in the left tail.

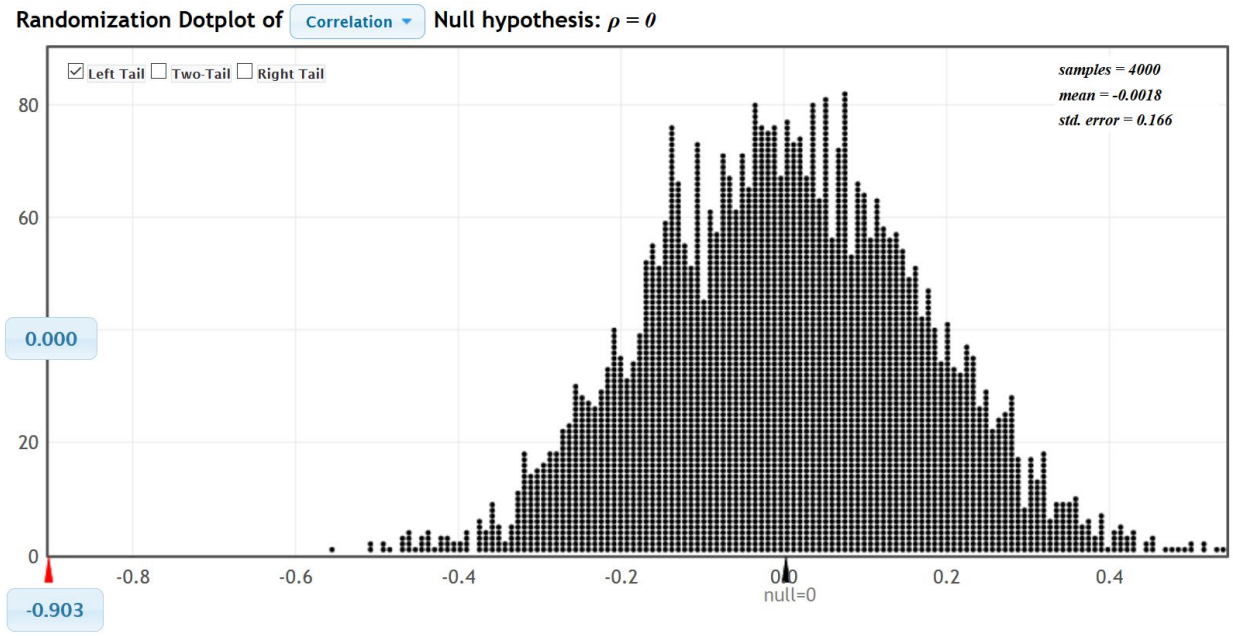


↑
r = −0.903

Since our correlation coefficient r falls in the left tail of the simulation, the sample data significantly disagrees with the null hypothesis.

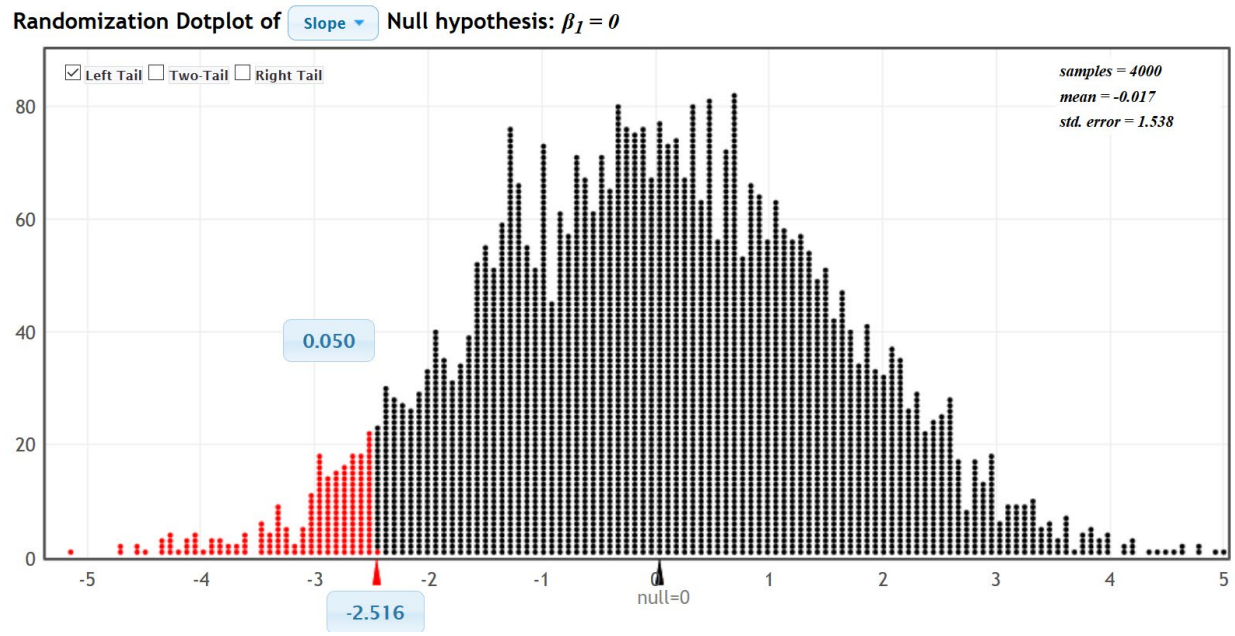
Calculating the P-value

The P-value is the probability of getting the sample data or more extreme by sampling variability if the null hypothesis is true. This simulated distribution is a view of sampling variability if the null is true. We just need to figure out the probability of the sample data or more extreme. Since this simulation created thousands of correlation coefficients, we will enter the real "original sample" correlation coefficient (r = −0.903) in the bottom box of the simulation. The left tail probability will give us the probability we are looking for. In this case, the
P-value was approximately zero.



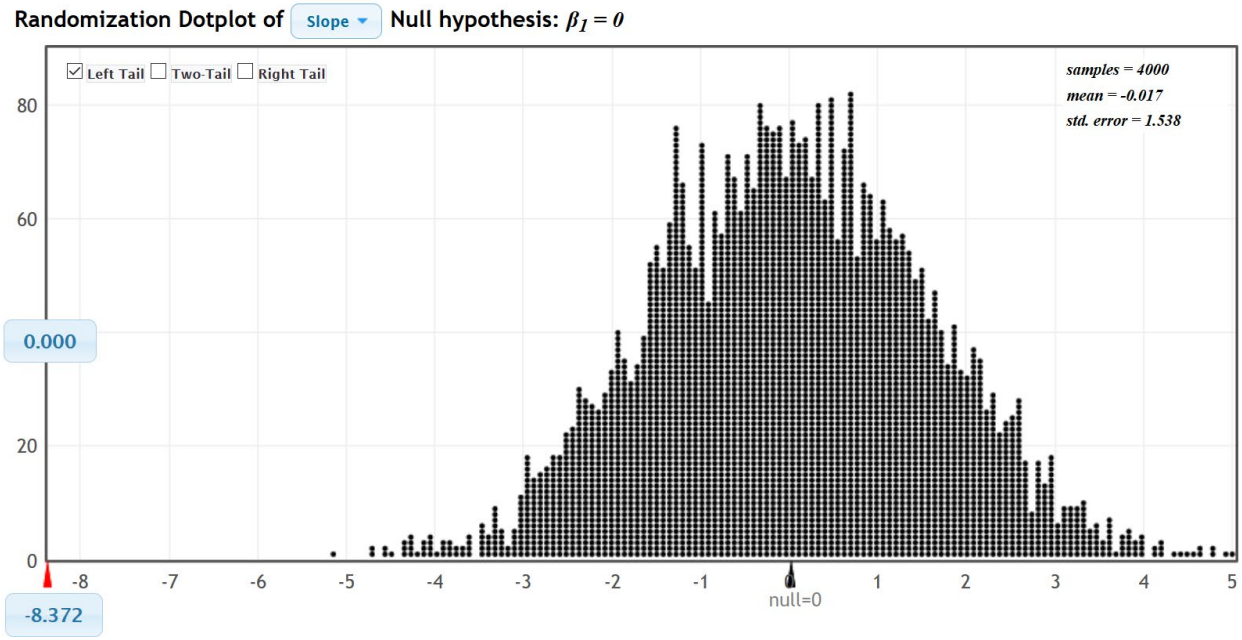
Simulating with the Slope

We can simulate with either the correlation coefficient or the slope. Here is the randomized simulation of thousands of sample slopes. Putting the 5% significance level in the tail, shows us that the real "original sample" slope needs to be −2.516 or less to be in the left tail. So if the real "original sample" slope is less than −2.516, the sample data will significantly disagree with the null hypothesis. The real "original sample" slope is −8.372 so it does fall in the left tail.



↑
Slope = −8.372

Now let us calculate the P-value with the slope. Enter the real "original sample" slope in the bottom box in the left tail. We see that the P-value is zero. Notice this is the same P-value as we got when we simulated with the correlation coefficient.



Notice that the original sample slope or correlation coefficient fell in the tail. So the sample data significantly disagrees with the null hypothesis. The slope is significantly different from zero.

What is the T-test statistic? Remember in a simulation, you do necessarily have to use the test statistic to judge significance. We used the sample correlation coefficient and slope to judge significance. We can calculate the T-test statistic though using the formula. Notice in the slope simulation, the approximate standard error for this simulation is 1.538. The standard error will vary between simulations though.

T-test statistic (for the correlation test) = $\frac{(Slope - Zero)}{Standard\ Error} = \frac{(-8.372 - 0)}{1.538} \approx -5.443$

T-test statistics Sentence: The slope of the regression line is 5.443 standard errors below zero.

P-value ≈ 0

P-value sentence: If the null hypothesis is true and there is no relationship between the weight of a car and the miles per gallon, then there is zero probability of getting this sample data or more extreme by sampling variability.

The P-value also tells us that it is extremely unlikely for this sample data to occur because of sampling variability.

The P-value is less than our 5% significance level, so we will reject the null hypothesis.

Conclusion: There is significant evidence to support the claim that there is a negative (inverse) relationship between the weight of a car and the miles per gallon. This does not imply that a heavy car causes the car to have

Notes

- Remember, if you simulate with the correlation coefficient, then you have to use the real "original sample" correlation coefficient when you calculate the approximate P-value. If you simulate with the slope, then you have to use the real "original sample" slope when you calculate the approximate P-value.
- You do not have to simulate with both the correlation coefficient and the slope. The point is that either simulation gives you approximately the same P-value.
- In all randomized simulations, there is sampling variability. Answers will vary slightly in different simulations.
