12.1.2: Hypothesis Test for a Correlation

One should perform a hypothesis test to determine if there is a statistically significant correlation between the independent and the dependent variables. The population correlation coefficient ρ (this is the Greek letter rho, which sounds like "row" and is not a p) is the correlation among all possible pairs of data values (x, y) taken from a population.

We will only be using the two-tailed test for a population correlation coefficient ρ. The hypotheses are:

$H_0 : \rho = 0$
$H_1 : \rho \neq 0$

The null-hypothesis of a two-tailed test states that there is no correlation (there is not a linear relation) between x and y. The alternative-hypothesis states that there is a significant correlation (there is a linear relation) between x and y.

The t-test is a statistical test for the correlation coefficient. It can be used when x and y are linearly related, the variables are random variables, and when the population of the variable y is normally distributed.

The formula for the t-test statistic is $t = r\sqrt{\left(\frac{n-2}{1-r^2}\right)}$.

Use the t-distribution with degrees of freedom equal to $df = n - 2$.

Note the $df = n - 2$ since we have two variables, x and y.

Test to see if the correlation for hours studied on the exam and grade on the exam is statistically significant. Use α = 0.05.

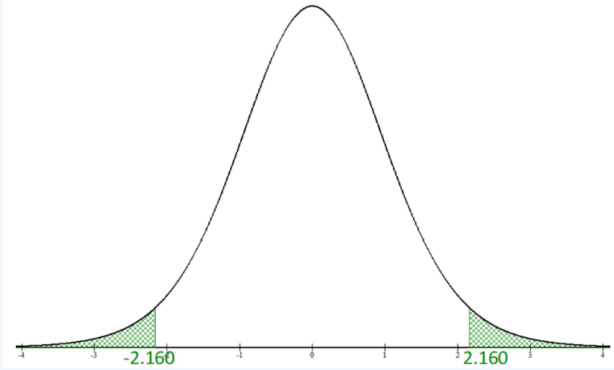
Hours Studied for Exam 20 16 20 18 17 16 15 17 15 16 15 17 16 17 14 Grade on Exam 89 72 93 84 81 75 70 82 69 83 80 83 81 84 76

Solution

The hypotheses are:

$H_0 : \rho = 0$
$H_1 : \rho \neq 0$

Find the critical value using $df = n - 2 = 13$ for a two-tailed test $\alpha = 0.05$ inverse t-function to get the critical values ± 2.160. Draw the sampling distribution and label the critical values as shown in Figure 12-5.



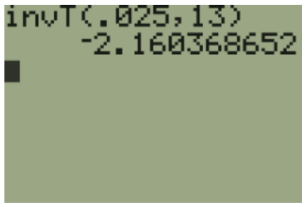


Figure 12-5: Sampling distribution of t-function with critical values labeled.

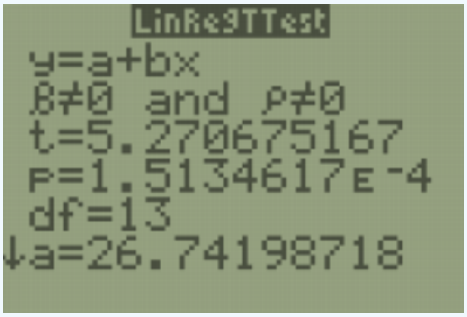
Next, find the test statistic $t = r\sqrt{\left(\frac{n-2}{1-r^2}\right)} = 0.8254\sqrt{\left(\frac{13}{1-0.8254^2}\right)} = 5.271$, which is greater than 2.160 and in the rejection region.

Summary: At the 5% significance level, there is enough evidence to support the claim that there is a statistically significant linear relationship (correlation) between the number of hours studied for an exam and exam scores.

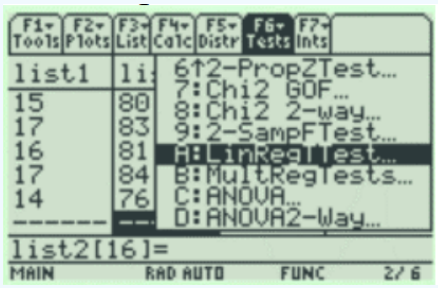
The p-value method could also be used to find the same decision. We will use technology shortcuts for the p-value method. The p-value = $2 \cdot \mathrm{P}(t \geq 5.271 | H_0 \text{ is true}) = 0.000151$, which is less than α = 0.05; therefore we reject H_0.

Alternatively, we could test to see if the slope was equal to zero. If the slope is zero then the correlation will also be zero. The setup of a test is a little different, but we get the same results. Most software packages report the test statistic and p-value for a slope. This test is introduced in the next section.

TI-84: Enter the data in L_1 and L_2. Press the [STAT] key, arrow over to the [TESTS] menu, arrow down to the option [LinRegTTest] and press the [ENTER] key. The default is Xlist: L_1, Ylist: L_1, Freq:1, β and $\rho : \neq 0$. Arrow down to Calculate and press the [ENTER] key. The calculator returns the t-test statistic, p-value and the correlation coefficient = r. Note the p-value = 0.0001513, is less than α = 0.05; therefore reject H_0, as there is a significant correlation.

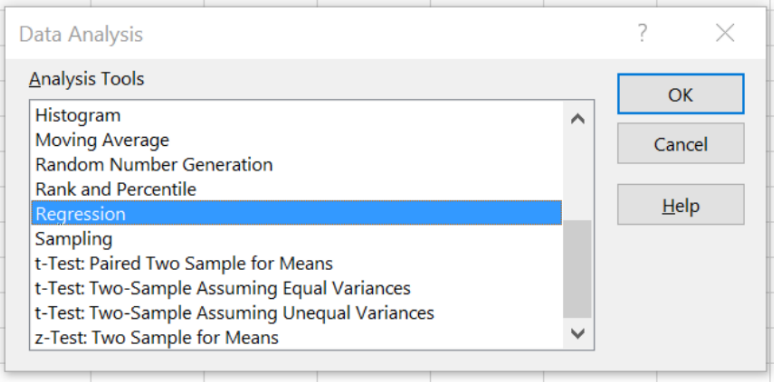


TI-89: Enter the data in List1 and List2. In the Stats/List Editor select F6 for the Tests menu. Use cursor keys to select A:LinRegTTest and press [Enter]. In the "X List" space type in the name of your list with the x variable without space, for our example "list1" or use [2nd] [Var-Link] and highlight list1. In the "Y List" space type in the name of your list with the y variable without space, for our example "list2" or use [2nd] [Var-Link] and highlight list2. Under the "Alternate Hyp" menu select the β and $\rho : \neq 0$ option, which is the same as the question's alternative hypothesis statement, then press the [ENTER] key, arrow down to [Calculate] and press the [ENTER] key. The calculator returns the t-test statistic, p-value, and the correlation = r.



Excel: Type the data into two columns in Excel. Select the Data tab, then Data Analysis, then choose Regression and select OK.

	A	B
1	Hours Studied for Exam	Grade on Exam
2	20	89
3	16	72
4	20	93
5	18	84
6	17	81
7	16	75
8	15	70
9	17	82
10	15	69
11	16	83
12	15	80
13	17	83
14	16	81
15	17	84
16	14	76



Be careful here. The second column is the y range, and the first column is the x range. Only check the Labels box if you highlight the labels in the input range. The output range is one cell reference where you want the output to start, and then select OK.

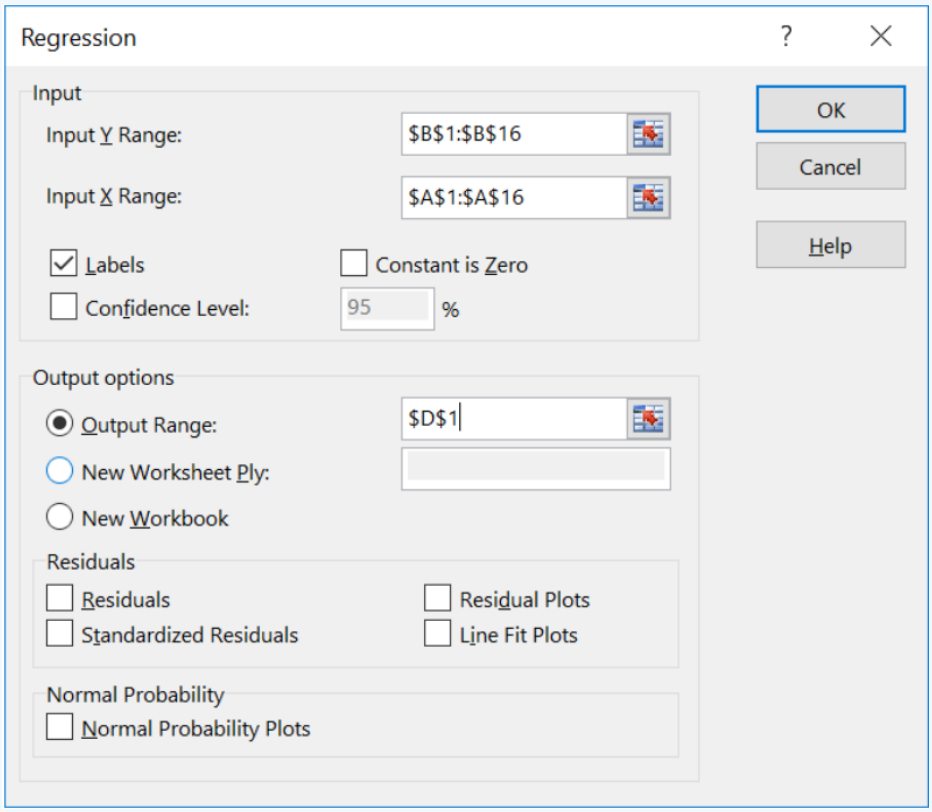


Figure 12-6 shows the regression output.

Regression Statistics	
Multiple R	0.825358
R Square	0.681216
Adjusted R Square	0.656695
Standard Error	3.935892
Observations	15

Absolute value of the correlation coefficient |r|

Sample size n

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	430.3471	430.3471	27.78002	0.000151
Residual	13	201.3862	15.49125		
Total	14	631.7333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	26.74199	10.18074	2.626725	0.020917
Hours Studied for Exam	3.216346	0.610234	5.270675	0.000151

t-test statistic

p-value

Figure 12-6: Excel-generated regression output.

When you reject H_0, the slope is significantly different from zero. This means there is a significant relationship (correlation) between x and y, and you can then find a regression line to use for prediction which we explore in the next section, called Simple Linear Regression.

Correlation is Not Causation

Just because two variables are significantly correlated does not imply a cause and effect relationship. There are several relationships that are possible. It could be that x causes y to change. You can actually swap x and y in the fields and get the same r value and y could be causing x to change. There could be other variables that are affecting the two variables of interest. For instance, you can usually show a high correlation between ice cream sales and home burglaries. Selling more ice cream does not "cause" burglars to rob homes. More home burglaries do not cause more ice cream sales. We would probably notice that the temperature outside may be causing both ice cream sales to increase and more people to leave their windows open. This third variable is called a **lurking variable** and causes both x and y to change, making it look like the relationship is just between x and y.

There are also highly correlated variables that seemingly have nothing to do with one another. These seemingly unrelated variables are called spurious correlations.

The following website has some examples of spurious correlations (a slight caution that the author has some gloomy examples): http://www.tylervigen.com/spurious-correlations. Figure 12-7 is one of their examples:

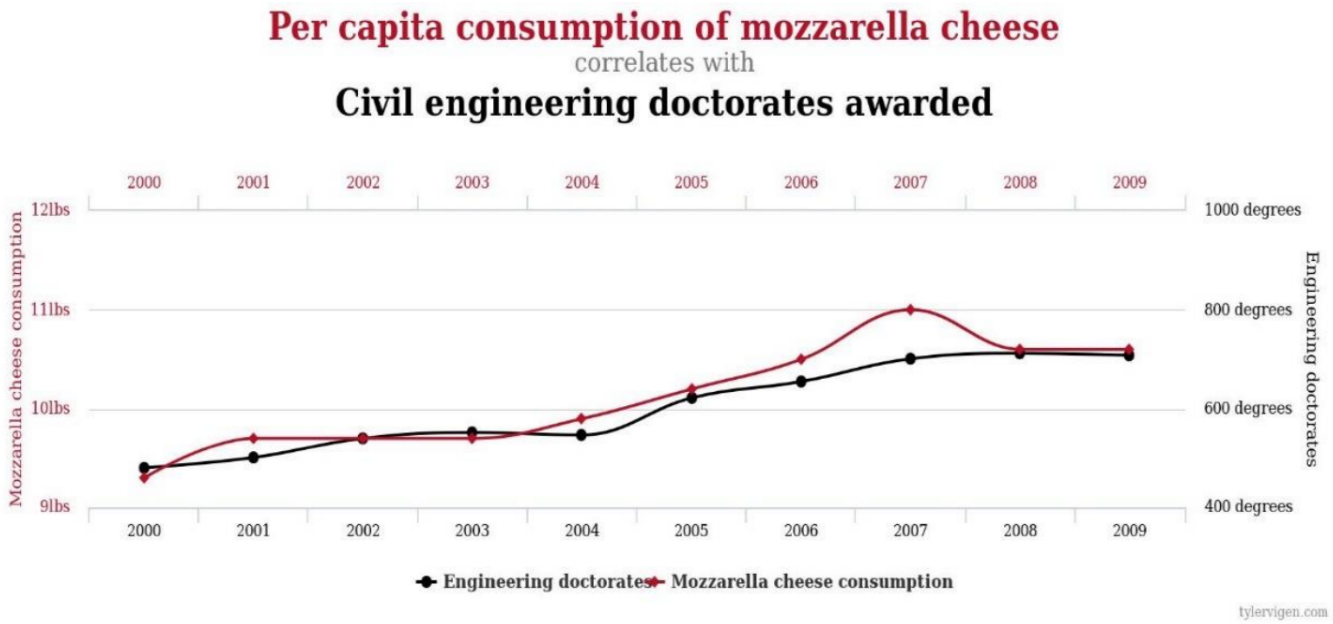


Figure 12-7: Example of spurious correlations. (6/25/2020) Retrieved from http://tylervigen.com/view_correlation?id=28726.

If we were to take out each pair of measurements by year from the time-series plot in Figure 12-7, we would get the following data.

Year	Engineering Doctorates	Mozzarella Cheese Consumption
2000	480	9.3
2001	501	9.7
2002	540	9.7
2003	552	9.7
2004	547	9.9
2005	622	10.2
2006	655	10.5
2007	701	11
2008	712	10.6
2009	708	10.6

Using Excel to find a scatterplot and compute a correlation coefficient, we get the scatterplot shown in Figure 12-8 and a correlation of $r = 0.9586$.

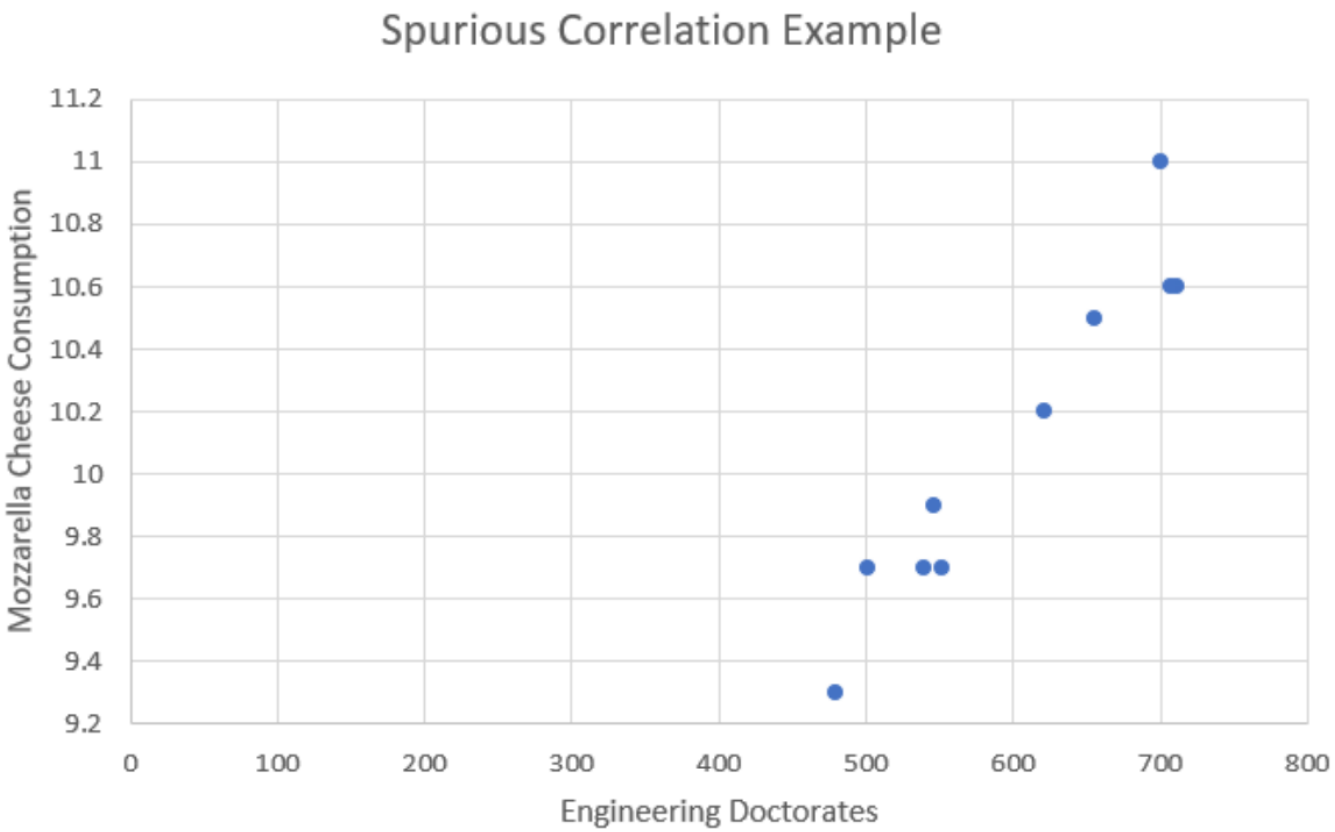


Figure 12-8: Scatterplot for spurious correlation example.

With $r = 0.9586$, there is strong correlation between the number of engineering doctorate degrees earned and mozzarella cheese consumption over time, but earning your doctorate degree does not cause one to go eat more cheese. Nor does eating more cheese cause people to earn a doctorate degree. Most likely these items are both increasing over time and therefore show a spurious correlation to one another.

> When two variables are correlated, it does not imply that one variable causes the other variable to change.

"Correlation is causation" is an incorrect assumption that because something correlates, there is a causal relationship. Causality is the area of statistics that is most commonly misused, and misinterpreted, by people. Media, advertising, politicians and lobby groups often leap upon a perceived correlation and use it to "prove" their own agenda. They fail to understand that, just because results show a correlation, there is no proof of an underlying causality. Many people assume that because a poll, or a statistic, contains many numbers, it must be scientific, and therefore correct. The human brain is built to try and subconsciously establish links between many pieces of information at once. The brain often tries to construct patterns from randomness, and may jump to conclusions, and assume that a cause and effect relationship exists. Relationships may be accidental or due to other unmeasured variables. Overcoming this tendency to jump to a cause and effect relationship is part of academic training for students and in most fields, from statistics to the arts.

Summary

When looking at correlations, start with a scatterplot to see if there is a linear relationship prior to finding a correlation coefficient. If there is a linear relationship in the scatterplot, then we can find the correlation coefficient to tell the strength and direction of the relationship. Clusters of dots forming a linear uphill pattern from left to right will have a positive correlation. The closer the dots in the scatterplot are to a straight line, the closer r will be to 1. If the cluster of dots in the scatterplots go downhill from left to right in linear pattern, then there is a negative relationship. The closer those dots in the scatterplot are to a straight line going downhill, the closer r will be to -1. Use a t-test to see if the correlation is statistically significant. As sample sizes get larger, smaller values of r become statistically significant. Be careful with outliers, which can heavily influence correlations. Most importantly, correlation is not causation. When x and y are significantly correlated, this does not mean that x causes y to change.