Linear Regression and Correlation: Testing the Significance of the Correlation Coefficient (modified R. Bloom)*

Roberta Bloom

Based on *Linear Regression and Correlation: Facts About the Correlation Coefficient for Linear Regress*

Susan Dean

Barbara Illowsky, Ph.D.

Abstract

This module provides an overview of Testing the Significance of the Correlation Coefficient for Roberta Bloom's Custom Collection of Collaborative Statistics col10617. It has been modified from the original module m17077, Facts About the Correlation Coefficient for Linear Regression, which is part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean. The test of significance is presented as a hypothesis test both using the p-value and using a table of critical values. Some of the material from the original module m17077 has been moved to module m33269 in Bloom's custom collection of Collaborative Statistics.

1 Testing the Significance of the Correlation Coefficient

The correlation coefficient, r, tells us about the strength of the linear relationship between x and y. However, the reliability of the linear model also depends on how many observed data points are in the sample. We need to look at both the value of the correlation coefficient r and the sample size n, together.

We perform a hypothesis test of the **"significance of the correlation coefficient"** to decide whether the linear relationship in the sample data is strong enough and reliable enough to use to model the relationship in the population.

The sample data is used to compute r, the correlation coefficient for the sample. IF we had data for the entire population, we could find the population correlation coefficient. But because we only have sample

*Version 1.2: Nov 15, 2010 1:54 pm -0600
†http://cnx.org/content/m17077/1.7/
‡http://creativecommons.org/licenses/by/3.0/

data, we can not calculate the population correlation coefficient. The sample correlation coefficient, r, is our estimate of the unknown population correlation coefficient.

The symbol for the population correlation coefficient is ρ, the Greek letter "rho".
ρ = population correlation coefficient (unknown)
r = sample correlation coefficient (known; calculated from sample data)

The hypothesis test lets us decide whether the value of the population correlation coefficient ρ is "close to 0" or "significantly different from 0". We decide this based on the sample correlation coefficient r and the sample size n.

If the test concludes that the correlation coefficient is significantly different from 0, we say that the correlation coefficient is "significant".

- Conclusion: "The correlation coefficient IS SIGNIFICANT"
- What the conclusion means: We believe that there is a significant linear relationship between x and y. We can use the regression line to model the linear relationship between x and y in the population.

If the test concludes that the correlation coefficient is not significantly different from 0 (it is close to 0), we say that correlation coefficient is "not significant".

- Conclusion: "The correlation coefficient IS NOT SIGNIFICANT."
- What the conclusion means: We do NOT believe that there is a significant linear relationship between x and y. Therefore we can NOT use the regression line to model a linear relationship between x and y in the population.

NOTE:

- If r is significant and the scatter plot shows a reasonable linear trend, the line can be used to predict the value of y for values of x that are within the domain of observed x values.
- If r is not significant OR if the scatter plot does not show a reasonable linear trend, the line should not be used for prediction.
- If r is significant and if the scatter plot shows a reasonable linear trend, the line may NOT be appropriate or reliable for prediction OUTSIDE the domain of observed x values in the data.

PERFORMING THE HYPOTHESIS TEST

SETTING UP THE HYPOTHESES:

- **Null Hypothesis: Ho: ρ=0**
- **Alternate Hypothesis: Ha: $\rho\neq$0**

What the hypotheses mean in words:

- **Null Hypothesis Ho:** The population correlation coefficient IS NOT significantly different from 0. There IS NOT a significant linear relationship(correlation) between x and y in the population.
- **Alternate Hypothesis Ha:** The population correlation coefficient IS significantly DIFFERENT FROM 0. There IS A SIGNIFICANT LINEAR RELATIONSHIP (correlation) between x and y in the population.

DRAWING A CONCLUSION:

There are two methods to make the decision. Both methods are equivalent and give the same result.
Method 1: Using the p-value

Method 2: Using a table of critical values

In this chapter of this textbook, we will always use a significance level of 5%, $\alpha = 0.05$

Note: Using the p-value method, you could choose any appropriate significance level you want; you are not limited to using $\alpha = 0.05$. But the table of critical values provided in this textbook assumes that we are using a significance level of 5%, $\alpha = 0.05$. (If we wanted to use a different significance level than 5% with the critical value method, we would need different tables of critical values that are not provided in this textbook.)

METHOD 1: Using a p-value to make a decision

The linear regression t-test LinRegTTEST on the TI-83+ or TI-84+ calculators calculates the p-value.
On the LinRegTTEST input screen, on the line prompt for β or ρ, highlight "$\neq$ **0**"
The output screen shows the p-value on the line that reads "p=".
(Most computer statistical software can calculate the p-value.)

If the p-value is less than the significance level ($\alpha = 0.05$):

- Decision: REJECT the null hypothesis.
- Conclusion: "The correlation coefficient IS SIGNIFICANT."
- We believe that there IS a significant linear relationship between x and y. because the correlation coefficient is significantly different from 0.

If the p-value is NOT less than the significance level ($\alpha = 0.05$)

- Decision: DO NOT REJECT the null hypothesis.
- Conclusion: "The correlation coefficient is NOT significant."
- We believe that there is NOT a significant linear relationship between x and y. because the correlation coefficient is NOT significantly different from 0.

Calculation Notes:

You will use technology to calculate the p-value. The following describe the calculations to compute the test statistics and the p-value:

The p-value is calculated using a t-distribution with n-2 degrees of freedom.

The formula for the test statistic is $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$. The value of the test statistic, t, is shown in the computer or calculator output along with the p-value. The test statistic t has the same sign as the correlation coefficient r.

The p-value is the probability (area) in both tails further out beyond the values -t and t.

For the TI-83+ and TI-84+ calculators, the command 2*tcdf(abs(t),10^99, n-2) computes the p-value given by the LinRegTTest; abs(t) denotes absolute value: $|t|$

THIRD EXAM vs FINAL EXAM EXAMPLE: p value method

- Consider the third exam/final exam example.
- The line of best fit is: $\hat{y} = -173.51 + 4.83x$ with $r = 0.6631$ and there are n = 11 data points.
- Can the regression line be used for prediction? **Given a third exam score (x value), can we use the line to predict the final exam score (predicted y value)?**

Ho: $\rho = 0$
Ha: $\rho \neq 0$
$\alpha = 0.05$
The p-value is 0.026 (from LinRegTTest on your calculator or from computer software)

The p-value, 0.026, is less than the significance level of $\alpha = 0.05$
Decision: Reject the Null Hypothesis Ho
Conclusion: The correlation coefficient IS SIGNIFICANT.
Because r is significant and the scatter plot shows a reasonable linear trend, the regression line can be used to predict final exam scores.

METHOD 2: Using a table of Critical Values to make a decision
The 95% Critical Values of the Sample Correlation Coefficient Table[1] at the end of this chapter (before the Summary[2]) may be used to give you a good idea of whether the computed value of r is significant or not. Compare r to the appropriate critical value in the table. If r is not between the positive and negative critical values, then the correlation coefficient is significant. If r is significant, then you may want to use the line for prediction.

Example 1
Suppose you computed $r = 0.801$ using $n = 10$ data points. df $= n - 2 = 10 - 2 = 8$. The critical values associated with df $= 8$ are -0.632 and + 0.632. If $r<$ negative critical value or $r >$ positive critical value, then r is significant. Since $r = 0.801$ and $0.801 > 0.632$, r is significant and the line may be used for prediction. If you view this example on a number line, it will help you.

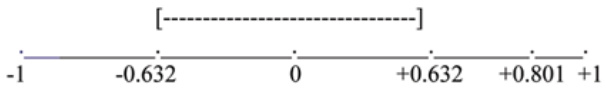

Figure 1: r is not significant between -0.632 and +0.632. $r = 0.801 > + 0.632$. Therefore, r is significant.

Example 2
Suppose you computed $r = -0.624$ with 14 data points. df $= 14 - 2 = 12$. The critical values are -0.532 and 0.532. Since $-0.624<-0.532$, r is significant and the line may be used for prediction

Figure 2: $r = -0.624<-0.532$. Therefore, r is significant.

Example 3
Suppose you computed $r = 0.776$ and $n = 6$. df $= 6 - 2 = 4$. The critical values are -0.811 and 0.811. Since $-0.811<0.776<0.811$, r is not significant and the line should not be used for prediction.

[1] "Linear Regression and Correlation: 95% Critical Values of the Sample Correlation Coefficient Table" <http://cnx.org/content/m17098/latest/>
[2] "Linear Regression and Correlation: Summary" <http://cnx.org/content/m17081/latest/>

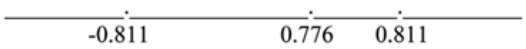

Figure 3: $-0.811<r=0.776<0.811$. Therefore, r is not significant.

THIRD EXAM vs FINAL EXAM EXAMPLE: critical value method

- Consider the third exam/final exam example.
- The line of best fit is: $\hat{y}= -173.51 + 4.83$x with $r = 0.6631$ and there are n = 11 data points.
- Can the regression line be used for prediction? **Given a third exam score (x value), can we use the line to predict the final exam score (predicted y value)?**

Ho: $\rho = 0$
Ha: $\rho \neq 0$
$\alpha = 0.05$
Use the "95% Critical Value" table for r with df = n − 2 = 11 − 2 = 9
The critical values are -0.602 and +0.602
Since $0.6631 > 0.602$, r is significant.
Decision: Reject Ho
Conclusion: The correlation coefficient is significant
Because r is significant and the scatter plot shows a reasonable linear trend, the regression line can be used to predict final exam scores.

Example 4: Additional Practice Examples using Critical Values
Suppose you computed the following correlation coefficients. Using the table at the end of the chapter, determine if r is significant and the line of best fit associated with each r can be used to predict a y value. If it helps, draw a number line.

1. $r = -0.567$ and the sample size, n, is 19. The df = $n - 2 = 17$. The critical value is -0.456. $-0.567<-0.456$ so r is significant.
2. $r = 0.708$ and the sample size, n, is 9. The df = $n - 2 = 7$. The critical value is 0.666. $0.708 > 0.666$ so r is significant.
3. $r = 0.134$ and the sample size, n, is 14. The df = $14 - 2 = 12$. The critical value is 0.532. 0.134 is between -0.532 and 0.532 so r is not significant.
4. $r = 0$ and the sample size, n, is 5. No matter what the dfs are, $r = 0$ is between the two critical values so r is not significant.

2 Assumptions in Testing the Significance of the Correlation Coefficient

Testing the significance of the correlation coefficient requires that certain assumptions about the data are satisfied. The premise of this test is that the data are a sample of observed points taken from a larger population. We have not examined the entire population because it is not possible or feasible to do so. We are examining the sample to draw a conclusion about whether the linear relationship that we see between

x and y in the sample data provides strong enough evidence so that we can conclude that there is a linear relationship between x and y in the population.

The regression line equation that we calculate from the sample data gives the best fit line for our particular sample. We want to use this best fit line for the sample as an estimate of the best fit line for the population. Examining the scatterplot and testing the significance of the correlation coefficient helps us determine if it is appropriate to do this.

The assumptions underlying the test of significance are:

- There is a linear relationship in the population that models the average value of y for varying values of x. In other words, the **average of the y values for each particular x value** lie on a straight line in the population. (We do not know the equation for the line for the population. Our regression line from the sample is our best estimate of this line in the population.)
- The y values for any particular x value are normally distributed about the line. This implies that there are more y values scattered closer to the line than are scattered farther away. Assumption (1) above implies that these normal distributions are centered on the line: the means of these normal distributions of y values lie on the line.
- The standard deviations of the population y values about the line are the equal for each value of x. In other words, each of these normal distributions of y values has the same shape and spread about the line.

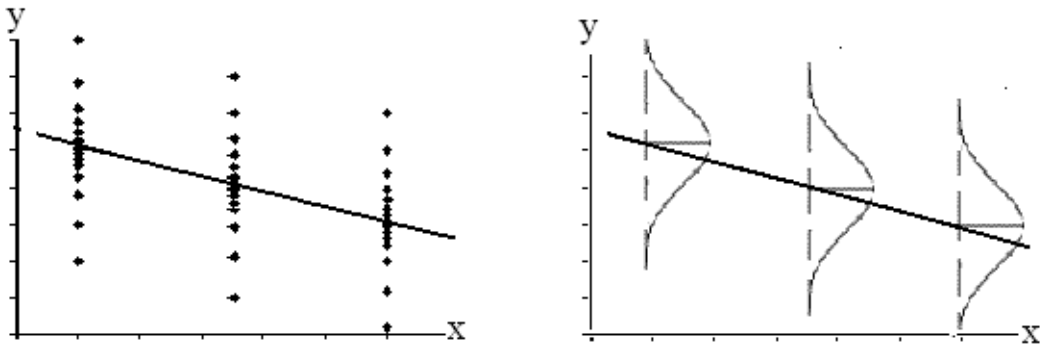

Figure 4: The y values for each x value are normally distributed about the line with the same standard deviation. For each x value, the mean of the y values lies on the regression line. More y values lie near the line than are scattered further away from the line.