10.3: Testing the Significance of the Correlation Coefficient

The correlation coefficient, r, tells us about the strength and direction of the linear relationship between X and Y.

The sample data are 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 have only sample data, we cannot calculate the population correlation coefficient. The sample correlation coefficient, r, is our estimate of the unknown population correlation coefficient.

- 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 zero" or "significantly different from zero". 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 zero, we say that the correlation coefficient is "significant."

- Conclusion: There is sufficient evidence to conclude that there is a significant linear relationship between variables X and Y because the correlation coefficient is significantly different from zero.
- What the conclusion means: There is a significant linear relationship between variables X and Y. If the test concludes that the correlation coefficient is not significantly different from zero (it is close to zero), we say that correlation coefficient is "not significant".

Performing the Hypothesis Test

- **Null Hypothesis:** $H_0 : \rho = 0$
- **Alternate Hypothesis:** $H_a : \rho \neq 0$

What the Hypotheses Mean in Words:

- **Null Hypothesis** H_0: The population correlation coefficient IS NOT significantly different from zero. There IS NOT a significant linear relationship (correlation) between X and Y in the population.
- **Alternate Hypothesis** H_a: The population correlation coefficient is significantly different from zero. There is a significant linear relationship (correlation) between X and Y in the population.

Drawing a Conclusion There are two methods of making the decision concerning the hypothesis. The test statistic to test this hypothesis is:

$$t_{obs} = \frac{r - \rho}{\sqrt{(1 - r^2)/(n - 2)}}$$

OR

$$t_{obs} = \frac{(r - \rho) * \sqrt{n - 2}}{\sqrt{1 - r^2}}$$

Where the second formula is an equivalent form of the test statistic, n is the sample size and the degrees of freedom are $n - 2$. This is a t-statistic and operates in the same way as other t-tests. Calculate the t-value and compare that with the critical value from the t-table at the appropriate degrees of freedom and the level of confidence you wish to maintain. If the calculated (observed) value is in the tail, then reject the null hypothesis that there is no linear relationship between these two independent random variables. If the calculated (observed) t-value is NOT in the tail, then we cannot reject the null hypothesis that there is no linear relationship between the two variables.

A quick shorthand way to test correlations is the relationship between the sample size and the correlation. If:

$$|r| \geq \frac{2}{\sqrt{n}}$$

then this implies that the correlation between the two variables demonstrates that a linear relationship exists and is statistically significant at approximately the 0.05 level of significance. As the formula indicates, there is an inverse relationship between the sample size and the required correlation for significance of a linear relationship. With only 10 observations, the required correlation

for significance is 0.6325, for 30 observations the required correlation for significance decreases to 0.3651 and at 100 observations the required level is only 0.2000.

Correlations may be helpful in visualizing the data, but are not appropriately used to "explain" a relationship between two variables. Perhaps no single statistic is more misused than the correlation coefficient. Citing correlations between health conditions and everything from place of residence to eye color have the effect of implying a cause and effect relationship. This simply cannot be accomplished with a correlation coefficient. The correlation coefficient is, of course, innocent of this misinterpretation. It is the duty of the analyst to use a statistic that is designed to test for cause and effect relationships and report only those results if they are intending to make such a claim. The problem is that passing this more rigorous test is difficult so lazy and/or unscrupulous "researchers" fall back on correlations when they cannot make their case legitimately.
