Correlation (Part 2)

Ray Block, Jr.
PS #585 (Research Methods)
Fall 2003

Today's Blueprint

Last Class

- Correlation (Part 1)
  - The Big Picture
  - Measuring Correlation

Today's Class

- Correlation (Part 2)
  - Measuring Correlation (A Recap)
  - Interpreting Correlation (contd.)

Measuring Correlation (A Recap)

[Graphical Illustration of What We Learned Last Class]

Interpreting Correlation

Strong *vs*. Meaningful Relationships:

- Strong correlations = significant correlations
- However, statistical Correlation does not always mean meaningful correlation
- A low Pearson r does not always indicate a insignificant correlation
  - If your sample is large enough, even a weak correlation is statistically significant
- A high Pearson r does not always mean a meaningful correlation
  - For example, a significant correlation between ice cream consumption and crime in a city might correlate highly, but the relationship itself is suspect
- Just "eyeballing" the correlation coefficient is not enough
- There are other, more sound ways of judging the meaningfulness of a correlation
  - The coefficient of determination
  - Hypothesis testing

Interpreting Correlation

Coefficient of Determination (r_{XY}^2):

- The coefficient of determination (r_{XY}^2) is the amount of variance that is accounted for in one variable by another variable
- It allows you to estimate the amount of variance that can be accounted for in one variable by examining the variance in another variable
- Example: Recall from last class that the correlation between education level and level of prejudice is very strong ($r_{XY} = -.92$)
- But is it meaningful?
  - The coefficient of determination would be $(-.92)^2 = .8464 \approx .85$
- This means that:
  - 85% of the variance in one variable is explained by the variance in the other variable
  - 15% (or 100% - 85%) of the variance is unexplained
  - this portion of unexplained variance is often referred to as the *coefficient of alienation*

Remember:

- The more variance explained, the better ☺
- The less variance left unexplained, the better ☺
- Therefore, the correlation between education and prejudice is a meaningful one because it the variance in one education explains most of the variance in prejudice and vice-versa (this leaves little variance between them unexplained)

Hypothesis Testing:

- Do a hypothesis test to determine whether the correlation between X and Y is a meaningful one
- The claim we are testing is: "There is a significant correlation"
- You want to determine whether:
  - The association between X and Y exists in the population (true correlation)
- Or whether:
  - The correlation is merely due to sampling error (false correlation)

- H_0: There is no relationship between X and Y
  - The null states that the population correlation between X and Y (rho_{XY}) is zero
  - H_0: rho = 0

- H_1: A relationship exists between X and Y
  - The alternative hypothesis states that the population correlation between X and Y is not zero (H_1: rho $\neq$ 0)
  - The correlation is either positive (H_1: rho > 0) or negative (H_1: rho < 0)

- To do hypothesis tests, you need to answer the following questions:
  - What are the degrees of freedom?
  - How confident do you want your test to be?

What are Degrees of Freedom (df)?

- df = the number of observation that are free to vary
- As a general equation, df = N -K
- Where:
  - N = # of observations
  - K = # of parameters to be estimated

- If we take a sample of N observations, then they are free to vary in any way (take on any values)
- However, if we use the sample of N observations to calculate the standard deviation, we use the sample mean as an estimate of the population mean

- Therefore, we hold one parameter constant
- With this parameter fixed, only N – 1 observations are free to vary

- Now, since we got the means for two variables (X and Y), there are two parameters being held constant
- If 2 parameters are fixed, then there are only N – 2 observations that are free to vary

- Therefore, when doing bivariate correlation, we use the following formula to calculate the degrees of freedom: df = N – 2
- From the education/prejudice example, we know that N = 10, so N – 2 = 8

How confident do you want the test to be?

- Usually, when people test hypotheses, they want to be at least 95% confident that they got it right
- The level of significance (represented by alpha "α") of a hypothesis test tells you how confident you can be about being right
- Most people test their hypotheses at the significance level of .05 (α = .05)
- This means that we only have a 5% chance (100% - 95% = 5%) of making a Type I error (rejecting the null hypothesis when it is true)
- With a .05 alpha level, the odds of being wrong about whether X relates to Y are 20 to 1

- Now that we have the necessary information, we can actually do the hypothesis test
  - r_{XY} = -.92

- N = 10
- df = 8
- α = .05

- Most statistics textbooks have a table where you can find the a list of significant values of Pearson's r for the .05 and .01 levels of significance with the number of degrees of freedom
- Based on the table, the critical r = .6319
- In order to reject the null that rho = 0 at the α = .05 level, our calculated Pearson's r must exceed .6319
- Since our Pearson's r is |.92| (disregarding the negative sign), we can reject the null

References

- FYI:

  - Levin, Jack and James Alan Fox. 2003. Elementary Statistics in Social Research, 9th Edition. Boston, MA: Pearson Education Group, Inc.
  - Salkind, Neil J. 2003. Exploring Research, 5th Edition. Upper Saddle River, NJ: Prentice Hall.
  - Kranzler, John H. 2003. Statistics for the Terrified, 3rd Edition. Upper Saddle River, NJ: Prentice Hall.