

Chapter 12: Correlation

I. Overview

- a. Correlation coefficients allow researchers to examine the association between two variables.
- b. The Pearson correlation coefficient (r) is the primary focus of this chapter
 - i. Involves associations between two variables measured on interval-ratio scales
- c. Correlation coefficients reveal the strength and direction of the association between the two variables
 - i. Possible range of r from -1.0 to 1.0
- d. There are several other types of correlations mentioned in this chapter, but not described in depth
 - i. Point-biserial: Correlations involving one dichotomous variable and one intervally scaled variable.
 - ii. Phi: Correlations involving two dichotomous variables.
 - iii. Spearman Rho: Correlations among ranked data.

II. What the Pearson Correlation Coefficient Tells Us

- a. The direction of the association between two variables
 - i. Positive: Scores on both variables move in the same direction: As scores on variable X increase, scores on variable Y also increase.
 - ii. Negative: Scores on the two variables move in opposite directions: As scores on variable X increase, scores on variable Y decrease.
- b. The strength of the association between two variables
 - i. The further the r value is from zero, in either direction, the stronger the association between the two variables.
 - 1. $r = -.50$ is a stronger correlation than $r = -.20$.
 - 2. Positive and negative correlations are equally strong.
 - a. $r = -.50$ is the same strength, or magnitude, as $r = .50$
 - ii. In the social sciences, correlation coefficients between $-.20$ and $.20$ are generally considered weak, between $.20$ and $.50$ (positive or negative) are considered moderate, and above $.50$ (positive or negative) are considered strong.
 - 1. This is just a rule of thumb. The specific variables in question and the nature of their association determine whether a specific correlation coefficient should be considered weak, moderate, or strong.
 - iii. Coefficient of Determination
 - 1. By squaring the correlation coefficient, researchers can calculate the coefficient of determination (r^2).
 - 2. This statistic reveals how much of the variance in one variable is explained by the second variable in the correlation analysis.
 - 3. This idea of *explained* or *shared variance* between two variables is a key concept for later statistics with multiple predictor variables (e.g., factorial ANOVA, multiple regression) and is a common measure of effect size (R^2 and eta-squared).

- c. Correlation *does not* tell us whether the association between two variables is a causal one.
 - i. E.g., if the correlation between happiness and number of movies watched per year is .30, that does not necessarily mean that watching movies increases happiness. Correlation \neq causation.
- d. Certain characteristics of the data or the association between the two variables can create distorted perceptions of the strength of the association between the two variables.
 - i. Curvilinear associations: When the association between two variables is positive at some values and negative at others (e.g., age and mental sharpness), the overall correlation can seem weak even when the association is quite strong.
 - ii. Truncated range: When there are ceiling or floor effects on one or both variables, the correlation between the two variables can appear weaker than it actually is.

III. How the Pearson r is calculated.

- a. The two variables are paired, such that for each case in the sample or population, the score on the first variable is paired with the score on the second variable.
- b. The scores on each of the variables in the analysis are standardized.
- c. Each pair of standardized scores is multiplied together, and these products are then summed.
- d. This sum is then divided by the number of cases in the distribution, i.e., the number of pairs of scores.
- e. This formula produces the average standardized cross-product (r).

IV. Testing for Statistical Significance

- a. Researchers often want to know whether a correlation coefficient calculated with sample data (r) represents a real, i.e., significant, correlation between these two variables in the population.
- b. The null hypothesis is that the population correlation coefficient (ρ) is zero.
 - i. Therefore, the t test formula is $(r - 0)/\text{standard error of } r$.
 - 1. Note that a shortcut formula can be used to avoid having to calculate the standard error of r .
 - ii. The resulting t value, with $n - 2$ degrees of freedom, can be looked up in Appendix B to determine whether it is statistically significant.

V. Summary

- a. Correlation coefficients are the basic statistical measure of the association between two variables.
 - i. It forms the basis for many more advanced statistics, such as regression, factor analysis, and structural equation modeling.
- b. There are several different types of correlation coefficients for different kinds of variables (e.g., interval, nominal).

- i. In this chapter we focused primarily on the Pearson r , used with two variables measured with interval/ratio scales.
- c. The strength, direction, and effect size can all be determined from the correlation coefficient.
- d. t tests can be calculated to determine whether a sample correlation coefficient is statistically significant.