The Center for Research and Outreach
REACH
Leveraging Evidence and Science into Practice

UNDERSTANDING CORRELATION

A HOW-TO GUIDE

USDA

UNIVERSITY OF MINNESOTA
Driven to Discover℠

Developed in collaboration with the Department of Defense's Office of Family Policy, the National Institute of Food and Agriculture, and the U.S. Department of Agriculture under Award No. 2009-48667-05833

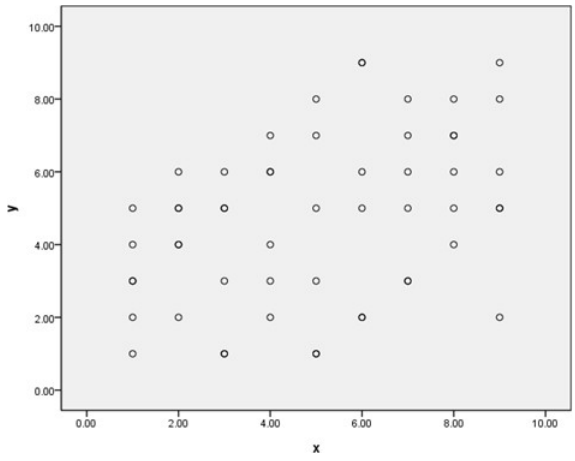
Understanding Correlation: A How-To Guide

Introduction

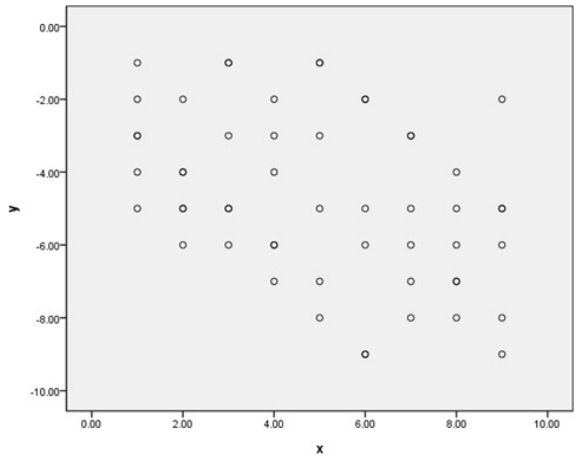
Perhaps one of the most basic and foundational statistical analysis techniques is the correlation. The proper name for correlation is the Pearson Product-Moment Correlation; however, it is rarely described as such and instead more commonly referred to as a correlation, Pearson correlation, or Pearson *r*. In this How-To Guide we will cover the basics of correlation as well as provide examples of how correlation is used in academic research.

What is a Correlation?

A correlation can be defined as the association between two *variables*[1]. It is the degree to which two variables share a common relationship. Consider for example two variables, one called 'X' and one called 'Y'. For each variable there is a distribution of scores, some scores are high and some scores are low and others are in between. For each variable there is a range of scores that increases from a low score to a high score. If 'Y' increases at a similar rate as 'X,' these two variables would be described as *positively correlated*. The image below on the left provides an example of a positive correlation. Alternatively if 'Y' decreases as 'X' increases then the two variables would be described as *negatively correlated*. The image below on the right provides an example of a negative correlation.



Positive Correlation



Negative Correlation

As another example, imagine that instructor quality was assessed in several youth programs. Some instructors are more effective than others. A correlation could provide useful information about what is associated with students' perception that the instruction is better. Many variables could be measured, but an obvious one would be how much time the instructor spent preparing. One would expect that those teachers who prepare more thoroughly are perceived more favorably. A positive correlation would confirm that these attributes of teaching are related.

[1] A *variable* refers to something that varies. An example might be the age of the participants in a study. The participants may differ in their age and thus will vary between one another. Another example of a variable is something that is measured. For example, in a study on youth programs, instructor quality might be measured. Because some instructors are more effective than others, scores for instructor quality will vary between instructors.

Properties of a Correlation

Correlation will often be reported as a number denoted by an '*r*'. The coefficient 'r' can range between -1.0 and +1.0. The plus and minus signs indicate the direction of the relationship. Minus means for example that lots of preparation leads to poorer teaching outcomes.
The size of the correlation coefficient is important because it communicates the degree to which the two variables are associated. High-quality articles will help readers interpret the correlation size by comparing it to other research or describing the consequences, for example, providing the difference in average student rating of instruction between teachers with small and large amounts of time preparing. When this specific information is not available, these guidelines can help interpret correlation size[1]:

- .10 can be considered small, whether it is a positive or negative relationship. It is not very useful to explain one event, such as average student rating of one class, but it could have consequences in the long-run (such as average student ratings over multiple years) or aggregate (such as student ratings of all the teachers in the program)
- .20 is medium, it could help explain single events and events in the long-run
- .30 is large, it could be powerful for explaining both single events and events in the long-run.

In specific fields or for specific types of studies, other classifications may be used.

Correlation and Causation

Correlations do not imply causation. Two variables can be related without one causing the other.
In other words, a correlation between two variables does not suggest that one variable causes another. A correlation can actually be interpreted in three different ways. Take the example of a correlation between instructor quality and instructor preparation. The first way that this correlation could be interpreted is that greater preparation leads to greater instructor quality. This interpretation suggests that instructors that prepare more provide better quality instruction. The second way that this correlation could be interpreted is that high quality instructors spend more time preparing. This interpretation suggests that it may not be the case that greater preparation time leads to better instruction, but rather, that higher quality instructors spend more time preparing. The third interpretation is that a third possibly unknown variable influences both instructor quality and instructor preparation. For example, teaching fewer activities may result in better instructor quality as well as provide more preparation time for the instructor. All three interpretations of the correlation are possible.

Reporting Correlations in Research Articles

There are two primary ways that correlations are reported in research articles: (1) in the text and (2) as a *correlation matrix*[2]. Below is an example of how a correlation might be reported in the text of a research article.

> *Pearson correlations were conducted to determine the association between three variables: instructor quality, instructor preparation, and number of activities taught. All correlations were significant at the p < .05 level. The correlation between instructor quality and instructor preparation was large (r = .31), as was the correlation between instructor preparation and number of teaching activities (r = .29). These are similar in*

[1] Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science, 2*(2), 156-168. doi: 10.1177/2515245919847202

[2] A *correlation matrix* is a simple **and visual** way to present several correlations at once. A correlation matrix will be presented as a table in research articles.

size to previously found correlations between preparation time and public speaking performance (r = .25, Menzel & Carrell, 1994[3]). The correlation between instructor quality and number of teaching activities was small (r = .12).

From this example we learn that the correlations ranged from small to large, that they were all positive, and that all were statistically significant at the $p < .05$ level (see the How-To Guide on *Statistical Language* for more information on p-values and significance testing). This information provides a basis for making judgments about the relative importance of each of the variables. In this example, preparation time had the largest correlation with instructor quality.

Generally, when correlations are conducted between three or more variables, the correlations will be presented in a table as a correlation matrix. Below is an example of what a correlation matrix might look like:

Variable			Correlation	
		1	2	3
1. Instructor Quality		--		
2. Instructor Preparation		.31*	--	
3. Activities Taught		.12*	.29*	--

Note: *p < .05

A correlation matrix is a useful way to summarize a set of correlations and is more commonly used as the number of correlations conducted increases. From a correlation matrix we can observe the size, direction, and significance level of several correlations at once. Each variable is numbered and named in the left column. In the top row, only the number of the variable appears. To determine how highly related instructor preparation (variable 2) is to instructor quality (variable 1), look for the number at the intersection of row 2 (for preparation) and column 1 (for quality). At this spot, one finds the size of the correlation, here it is .31. Note that the hyphens or dash marks indicate no correlation because that is the space denoting a variable's relationship to itself.

Summary

This How-To Guide illustrates how correlations might be presented in research articles. The goal of this guide is to improve skills to critically read scientific research. Readers are encouraged to review our other How-To guides on the website that provide information about Regression, ANOVA, t-Tests, and Statistical Language.

[3] Our correlations on instructional quality are provided as a fictional example. Readers interested in learning about the real relationship between preparation time and public speaking performance can consult Menzel, K. E., & Carrell, L. J. (1994). The relationship between preparation and performance in public speaking. *Communication Education, 43*(1), 17-26.

Developed by
The Arizona Center for Research and Outreach (AZ REACH)

Lynne M. Borden, Principal Investigator
Gabriel Schlomer, Ph.D. (Primary Author)

Updated March 2021

The Center for Research and Outreach (REACH Lab)
University of Minnesota
Email: REACHLab@umn.edu
Website: https://reachfamilies.umn.edu