Chapter 16: Statistics 3,  
Correlation and Chi2

## Learning Objectives

1. Understand when and how to use a correlation analysis
2. Interpret and understand ranges for Pearson’s and Spearman’s *r*
3. Understand when and how to do a X2 analysis with categorical data
4. Understand how to draw inferences from X2 analysis.

# Correlation Analysis

For non-experimental research, simple percentages may be computed to describe the percentage of people who engaged in some behavior or held some belief. But more commonly non-experimental research involves computing the correlation between two variables. A correlation coefficient describes the strength and direction of the relationship between two variables. The values of a correlation coefficient can range from −1.00 (the strongest possible negative relationship) to +1.00 (the strongest possible positive relationship). A value of 0 means there is no relationship between the two variables. Positive correlation coefficients indicate that as the values of one variable increase, so do the values of the other variable. A good example of a positive correlation is the correlation between height and weight, because as height increases weight also tends to increase. Negative correlation coefficients indicate that as the value of one variable increase, the values of the other variable decrease. An example of a negative correlation is the correlation between stressful life events and happiness; because as stress increases, happiness is likely to decrease.

The phrase Correlation is not Causality is so common and important. Mathematically, a correlation analysis simply assesses the strength of the relationship between two continuous variables. In theory, if the independent variable was manipulated across a continuous range by the experimenter, we could draw perfectly reasonable causal inference from a correlation coefficient. In practice, this is vanishingly rare, so it is generally safe to use the heuristic that if you see a correlation coefficient reflecting a correlation analysis, the research to which it is being applied is likely also correlational research and non-experimental.

## Correlations Between Quantitative Variables

Correlations between quantitative variables are often presented using scatterplots. An example is shown below based on hypothetical data on the relationship between the amount of stress people are under and the number of physical symptoms they have. Each point in the scatterplot represents one person’s score on both variables. For example, the circled point in red represents a person whose stress score was 10 and who had five physical symptoms. The orange circled point is a participant with a stress score of 20 and twelve physical symptoms. Taking all the points into account, one can see that people under more stress tend to have more physical symptoms. This is a good example of a positive relationship, in which higher scores on one variable tend to be associated with higher scores on the other. In other words, they move in the same direction, either both up or both down. A negative relationship is one in which higher scores on one variable tend to be associated with lower scores on the other. In other words, they move in opposite directions. There is a negative relationship between stress and immune system functioning, for example, because higher stress is associated with lower immune system functioning.

 The strength of a correlation between quantitative variables is typically measured using a statistic called Pearson’s Correlation Coefficient (or Pearson's r). Pearson’s r ranges from −1.00 (the strongest possible negative relationship) to +1.00 (the strongest possible positive relationship). A value of 0 means there is no relationship between the two variables. When Pearson’s r is 0, the points on a scatterplot form a shapeless “cloud.” As its value moves toward −1.00 or +1.00, the points come closer and closer to falling on a single straight line. Correlation coefficients near ±.10 are considered small, values near ± .30 are considered medium, and values near ±.50 are considered large. Notice that the sign of Pearson’s *r* is unrelated to its strength. Pearson’s r values of +.30 and −.30, for example, are equally strong; it is just that one represents a moderate positive relationship and the other a moderate negative relationship. With the exception of reliability coefficients, most correlations that we find in Psychology are small or moderate in size. The scatterplot above has a correlation coefficient between the hypothetical data of 0.55, a fairly strong positive relationship.

An analysis that produces a correlation coefficient is expressed with the statistical parameter, *r*, which like other statistical parameters (t, F) reflects the strength of the relationship between the variables. It is also associated with a p-value, which always has the same definition, the probability of observing this relationship by chance if the null hypothesis was correct. For a correlation analysis, the null hypothesis is that there is no relationship between variables which would produce an *r* = 0.00.

Additional correlation examples:

|  |  |
| --- | --- |
|  |  |
| *r* = 0.2, very small effect | *r* = 0.3, small effect |
|  |  |
| *r* = 0.5, moderate to large effect | *r* = 0.8, large effect |
|  | *Scatterplots illustrating a variety of strengths of relationships between the x and y variables* |
| *r* = -0.5, moderate to large negative effect |  |

There are two common situations in which the value of Pearson’s r can be misleading. Pearson’s r is a good measure only for linear relationships, in which the points are best approximated by a straight line. It is not a good measure for curvilinear relationships, in which the points are better approximated by a curved line. The figure below, for example, shows a hypothetical relationship between the amount of sleep people get per night and their level of depression. In this example, the line that best approximates the points is a curve—a kind of upside-down “U”—because people who get about eight hours of sleep tend to be the least depressed. Those who get too little sleep and those who get too much sleep tend to be more depressed. Even though the figure shows a fairly systematic relationship between depression and sleep, Pearson’s r would be close to zero because the points in the scatterplot are not well fit by a single straight line (flat trend line shown). This means that it is important to make a scatterplot and confirm that a relationship is approximately linear before using Pearson’s r. Curvilinear relationships are fairly common in psychology, but measuring their strength is beyond the scope of this book (this kind of curve would be fit by polynomial regression and show a strong quadratic fit).

The other common situations in which the value of Pearson’s r can be misleading is when one or both of the variables have a limited range in the sample relative to the population. This problem is referred to as restriction of range. Assume, for example, that there is a strong negative correlation between people’s age and their enjoyment of hip hop music as shown by the scatterplot across age ranges from 18 to 80. However, if data were collected from a restricted range sample, e.g., 18 to 24, the relationship might not be visible. This is yet another example of why we cannot confidently draw conclusions from null results. It is also a reminder that calculation of a correlation coefficient based on Pearson’s r depends on having data sampled across a reasonably wide rand and also assumes that the distribution of both the x and y variables are roughly normal (gaussian).

A tool to be aware for conditions in which the data are not normally distributed is the **Spearman’s rank correlation**. This also results in calculating an *r* statistic that acts just like the Pearson’s correlation. Spearman’s correlation can be used when the observed data has a number of notable outliers that would not be expected in a normally distributed dataset. This correlation coefficient is calculated based on ranking the data such that the lowest value is recoded as 1 and each higher value is one more so that the highest value in the data set is the number of total participants. This reduces the distorting impact of extreme outliers that can reduce the effectiveness of a more typical Person’s correlation. An example of where this tool can be used effectively is in the analysis of reaction time (RT) data where most of the responses cluster around some average speed but there are a few extremely slow responses (a very common shape of RT data). This produces a highly skewed distribution that is not gaussian (normal). A rank correlation, Spearman’s, analysis enables analysis of these types of data without problems caused by the violation of the assumption of normality. As with all tools that allow us to carry out statistical analysis when data violate assumptions of normality, this should be used with caution and some thought towards why the data are not normally distributed.

# Analysis of Categorical Data

In our standard model of experimental design, we use a manipulated (experimental) independent variable and measure a dependent variable. The IV (or factors) typically have a small number of levels, often 2, among which participants are assigned. We can think of our IV as being defined by a categorical variable in that participants are assigned to one condition or the other. The DV is a continuous variable that we can then look for differences in the average score across conditions. The correlation analysis described above is one variation from this model where both the IV and DV are continuous measures. It can also be the case that the DV sometimes needs to be a categorical variable.

The canonical examples of categorical variables can be captures in the memorable phrase, “you can’t be a little bit pregnant or a little bit dead.” These are events for which there are only two outcomes: you are, or you aren’t. Measures of these kinds of variables are ‘binary’ in that there are two possibilities. It is also possible to have categorical variables for which there are more than two alternatives. In general, if the alternatives can be ordered in a systematic, ranked way, these will often be coded as a familiar continuous variable. But there are plenty of cases where there is a range of options that are each independent choices. For example, one might look at some aspect of high school education and what college within Northwestern a student applied with the possible outcomes being WCAS, McCormick, Medill, or the Bienen School of Music. Here the outcome variable is categorical across four possibilities.

These approaches can very well be experimental, and the same concepts drive our ability to draw inferences from the data: did the IV affect the DV? But now we need a statistical tool that allows us to characterize how the shift in categorical choices was affected by the manipulated IV. The general approach for analyzing these data is to organize the outcomes into a **contingency table**.

As an example, consider a non-experimental study that asked athletes if they generally stretch before exercising and if they have had an injury in the past year. Whether or not they stretch is a categorical variable with two possibilities: yes/no. The same is try for whether they have had an injury in the past year. Suppose we had data from 800 athletes. We could organize the results in the following contingency table.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Injury | No Injury | Total |
| Stretches | 55 | 295 | 350 |
| Does not stretch | 231 | 219 | 450 |
| Total | 286 | 514 | 800 |

Looking at the outcome counts, we can see that the number of people with injuries is lower in the group that stretches, but we should also note that there are different numbers of participants in the stretch/no-stretch conditions. To correct for this, we should calculate rates for all the conditions here to see if there is evidence the injury rate is different for the two stretching conditions. The rate of injury for the stretches condition is 55/350 = 15.7%. The rate of injury the not stretching condition is 231/450 = 51.3%. That is clearly a lot higher, but what we have done so far is effectively calculated the descriptive statistics for a categorical design. We need a statistical test to identify if this difference is statistically reliable that will give us a familiar p-value, the probability of observing this pattern of data under the null hypothesis. The null hypothesis here is that stretching or not are associated with injuries at the roughly the same rate overall (and the observed different was somehow just luck).

There are a variety of ways to analyze a contingency table, but we will focus on one tool that is fairly flexible across common experimental designs, the chi-squared analysis or Χ2. ‘Χ’ is the Greek letter ‘chi’ and the analysis is variously referred to by the Greek letter, the term “chi-squared” or the mixed term “chi2”. Conceptually, this is based on looking at the difference between the Observed rate of occurrence from the Expected rate of occurrence in each of the four cells of the contingency table. The Expected rate is the rate of occurrence under the null hypothesis that stretching does not matter. It is useful to note that the injury rate in the dataset does not have to be 50% just because there are two possibilities. We can estimate the average injury rate by looking at the total number of injuries over all the participants ignoring the stretching condition (just as we did with marginal means for main effects in ANOVA). There were a total of 286 injuries out of the 800 participants, which is a rate of 35.8%. The Χ2 formula is essentially telling us that if the average rate of injury is 35.8%, what are the odds that one group would exhibit a 15.7% rate and the other group a 51.3% rate. The number of participants is critical for this calculation, so it is actually done with the numbers expected in each cell not from the rates themselves. The Χ2 is the sum of the difference between the Expected number of participants in each cell (under the null) and the Observed number of participants. For the stretches condition, at a 35.8% injury rate, we should have seen ~125 injuries, yet we only saw 55. Χ2 is actually calculated as the square of this difference divided by the Expected value and so the sum of this across the four cells is:



For this example, we would obtain a Χ2 value of 108.7, which is exceedingly improbable and since p<.05 (the same criterion as always), we can reject the null hypothesis and conclude that stretching did reliably affect the injury rate in this data collection sample.

Nothing about the calculation would be different if this was an experimental study where athletes were assigned to a stretching or no stretching condition and then injury rate was assessed afterwards. Of course, that would be a problematically unethical study since we would have assigned participants to a condition that we thought might cause them to become injured.

The method applies to any design where the DV is a categorical variable and works well even if there are more than two possible outcomes. The general approach is to organize the data into a contingency table of counts of each of the combinations of IV level and DV outcomes. From this, calculate the rates to be able to describe the data as descriptive statistics. The calculation of a Χ2 analysis is simple enough that you could do it with a calculator or spreadsheet, but you can also simply use standard analysis programs like R which also helpfully provide the direct estimate of the p-value.

# General Linear Models, Complex Correlation

Although the more general technique of **regression** is outside the scope of the statistical tools considered here, it is worth noting that correlation is a simplified case of regression where there is one “predictor” (X) variable and one “outcome” (Y) variable. More sophisticated analysis using **multiple regression** are done within an approach called **general linear modeling** (GLM). The approach of using a GLM is the basis of the vast majority statistical modeling of complex datasets except for a handful of special and interesting cases where non-linear relationships are evaluated (although if you work in this area, you learn a variety of approaches for transforming non-linear data to be suitable for linear modeling).

Some common approaches are briefly described here for familiarity with the terms and general concepts for these more complex analytic techniques. However, use and application to complex design are outside the scope of what is covered in this class.

One useful idea that emerges from regression or GLM analysis is the calculation of a statistic, *r2*, which has a useful verbal description of “percentage of variance accounted for” in the data. This is, in fact, the same *r* statistic we use for correlation analysis squared and can help understand what a correlation analysis is telling us. The idea behind the phrase “percent of variance accounted for” is that the data have variance which results from a range of factors related to extraneous variables, participant variances and measurement error. Our predictor variable (IV) accounts for some of the variability in the measured variable (DV) but we acknowledge that there are many other sources of variance not accounted for. In the example of a fairly strong correlation of *r* = 0.5, *r2* = 0.25 meaning a quarter of the variance is accounted for and three quarters (75%) is still unknown. So even in the case of a strong relationship, there’s still a lot we have not captured in the statistical model. For milder correlations in the range of 0.3, we would be happy to account for 10% of the variance. The *r2* statistic is generally reported when using any GLM (regression) analysis. Many ANOVA programs are using a GLM behind the scenes (ANOVA is also a special case of the broader regression approach) and may report the *r2* statistic as well. If so, you can use this heuristic to get a sense of how systematic your data are by how much of the variance is accounted for in your model.

## Assessing Relationships Among Multiple Variables

Most complex correlational research involves measuring several variables—either binary or continuous—and then assessing the statistical relationships among them. For example, Radcliffe & Klein (2002) studied a sample of middle-aged adults to see how their level of optimism (measured by using a short questionnaire called the Life Orientation Test) relates to several other variables related to having a heart attack. These included their health, their knowledge of heart attack risk factors, and their beliefs about their own risk of having a heart attack. They found that more optimistic participants were healthier (e.g., they exercised more and had lower blood pressure), knew about heart attack risk factors, and correctly believed their own risk to be lower than that of their peers.

In another example, Jouriles, Garrido, Rosenfield & McDonald (2009). measured adolescents’ experiences of physical and psychological relationship aggression and their psychological distress. Because measures of physical aggression (such as the Conflict in Adolescent Dating Relationships Inventory and the Relationship Violence Interview) often tend to result in highly skewed distributions, the researchers transformed their measures of physical aggression into a dichotomous (i.e., binary) measure (0 = did not occur, 1 = did occur). They did the same with their measures of psychological aggression and then measured the correlations among these variables, finding that adolescents who experienced physical aggression were moderately likely to also have experienced psychological aggression and that experiencing psychological aggression was related to symptoms of psychological distress.

## Multiple Dependent Variables

In the development of novel survey instruments, it is common to collect a set of dependent variables and look for relationships among these. This can be used to calculate both convergent and discriminant validity of the scale. For example, when Cacioppo & Petty (1982) first reported the Need for Cognition Scale—a measure of the extent to which people like to think and value thinking—they used it to measure the need for cognition for a large sample of college students, along with three other variables: intelligence, socially desirable responding (the tendency to give what one thinks is the “appropriate” response), and dogmatism. The results of this study are summarized in a correlation matrix (below) showing the correlation (Pearson’s r) between every possible pair of variables in the study. For example, the correlation between the need for cognition and intelligence was +.39, the correlation between intelligence and socially desirable responding was +.02, and so on. (Only half the matrix is filled in because the other half would contain exactly the same information. Also, because the correlation between a variable and itself is always +1.00, these values are replaced with dashes throughout the matrix.) In this case, the overall pattern of correlations was consistent with the researchers’ ideas about how scores on the need for cognition should be related to these other constructs.

 Correlation Matrix Showing Correlations Among the Need for Cognition and Three Other Variables Based on Research by Cacioppo and Petty (1982)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Need for cognition | Intelligence | Social desirability | Dogmatism |
| Need for cognition | — |  |  |  |
| Intelligence | +.39 | — |  |  |
| Social desirability | +.08 | +.02 | — |  |
| Dogmatism | −.27 | −.23 | +.03 | — |

## Factor Analysis

When researchers study relationships among a large number of conceptually similar variables, they often use a complex statistical technique called factor analysis. In essence, factor analysis attempts to organize the observed data as arising from a smaller number of predictor variables than were originally used. As an example, researchers Rentfrow & Gosling (2008) asked more than 1,700 university students to rate how much they liked 14 different popular genres of music. They then submitted these 14 variables to a factor analysis, which identified four distinct underlying factors. The researchers called them Reflective and Complex (blues, jazz, classical, and folk), Intense and Rebellious (rock, alternative, and heavy metal), Upbeat and Conventional (country, soundtrack, religious, pop), and Energetic and Rhythmic (rap/hip-hop, soul/funk, and electronica). The underlying idea is that the rating of the blues, jazz, classical and folk music tended to be similar to each other, so these are reduced to one underlying factor. Note that this analysis does not tell you anything about what the factor means, which is often considered a weakness of this approach, leaving it up to the authors to decide to describe this cluster as “Reflected and Complex.” The analysis provides a table of “factor loadings” (below) which indicate how well each of the observed measures relates to the inferred cluster (factor).

Table of Factor Loadings of the 14 Music Genres on Four Varimax-Rotated Principal Components. Based on Research by Rentfrow and Gosling (2003)

|  |  |
| --- | --- |
|  | Music-preference dimension |
| Genre | Reflective and Complex | Intense and Rebellious | Upbeat and Conventional | Energetic and Rhythmic |
| Blues | .85 | .01 | -.09 | .12 |
| Jazz | .83 | .04 | .07 | .15 |
| Classical | .66 | .14 | .02 | -.13 |
| Folk | .64 | .09 | .15 | -.16 |
| Rock | .17 | .85 | -.04 | -.07 |
| Alternative | .02 | .80 | .13 | .04 |
| Heavy metal | .07 | .75 | -.11 | .04 |
| Country | -.06 | .05 | .72 | -.03 |
| Sound tracks | .01 | .04 | .70 | .17 |
| Religious | .23 | -.21 | .64 | -.01 |
| Pop | -.20 | .06 | .59 | .45 |
| Rap/hip-hop | -.19 | -.12 | .17 | .79 |
| Soul/funk | .39 | -.11 | .11 | .69 |
| Electronica/dance | -.02 | .15 | -.01 | .60 |
|  |

## Exploring Causal Relationships

Another important use of complex correlational research is to explore possible causal relationships among variables. This might seem surprising given the oft-quoted saying that “correlation does not imply causation.” It is true that correlational research cannot unambiguously establish that one variable causes another. Complex correlational research, however, can often be used to rule out other plausible interpretations. The primary way of doing this is through the statistical control of potential third variables. Instead of controlling these variables through random assignment or by holding them constant as in an experiment, the researcher instead measures them and includes them in the statistical analysis called partial correlation. Using this technique, researchers can examine the relationship between two variables, while statistically controlling for one or more potential third variables. You will typically see this described as “controlling for,” as in X appears to cause a change in Y controlling for Z in an analysis where the effect of Z are attempted to be controlled for to try to give an independent view of how X affects Y.

For example, assume a researcher was interested in the relationship between watching violent television shows and aggressive behavior but she was concerned that socioeconomic status (SES) might represent a third variable that is driving this relationship. In this case, she could conduct a study in which she measures the amount of violent television that participants watch in their everyday life, the number of acts of aggression that they have engaged in, and their SES. She could first examine the correlation between violent television viewing and aggression. Let’s say she found a correlation of +.35, which would be considered a moderate sized positive correlation. Next, she could use partial correlation to reexamine this relationship after statistically controlling for SES. This technique would allow her to examine the relationship between the part of violent television viewing that is independent of SES and the part of aggressive behavior that is independent of SES. If she found that the partial correlation between violent television viewing and aggression while controlling for SES was +.34, that would suggest that the relationship between violent television viewing and aggression is largely independent of SES (i.e., SES is not a third variable driving this relationship). On the other hand, if she found that after statistically controlling for SES the correlation between violent television viewing and aggression dropped to +.03, then that would suggest that SES is indeed a third variable that is driving the relationship. If, however, she found that statistically controlling for SES reduced the magnitude of the correlation from +.35 to +.20, then this would suggest that SES accounts for some, but not all, of the relationship between television violence and aggression. It is important to note that while partial correlation provides an important tool for researchers to statistically control for third variables, researchers using this technique are still limited in their ability to arrive at causal conclusions because this technique does not take care of the directionality problem and there may be other third variables driving the relationship that the researcher did not consider and statistically control.

## General Linear Models

Calculation a partial correlation is one of many useful things that can be done with a regression model, which is an example of the more general class of General Linear Models. These can be used for statistical inference about variables, accounting for variance in the population and also predictions about new data that is to be collected. The specific set of statistical tools we have focused on for simple experimental design are all derived from a GLM approach, although it is generally useful to call a t-test a t-test and an ANOVA an ANOVA.

If you continue to study more complex research methods beyond this class, you will encounter regression analysis, ANCOVA (analysis of covariance), MANOVA (multiple variable ANOVA), logistic regression, and even more complex tools such as structural equation modeling. Many of these are used in non-experimental research studies to try to increase the confidence in drawing causal conclusions from datasets where the independent variables cannot be manipulated (e.g., epidemiology, economics). It may be useful to you then to know that these are all founded on the GLM core and developing an understanding of this multiple regression approach will help you grasp a wide range of more specialized tools used for those areas of research.