Page 84-86

Chapter 5 Testing Hypotheses

Learning Objectives

As you know, the previous chapters presented simple ways to organize, display, and summarize the distribution of a single variable. Sometimes this is the only type of analysis that we need, as our only intention is to accurately describe the distribution of a variable within a sample or population. This is often the case in research studies that are primarily descriptive, such as those that employ one-group posttest-only designs, or in cross-sectional case study designs.

However, we frequently want to learn more from any given data set. We may want to know, for example, if there is a relationship between two or more variables within the data set. This requires different methods of data analyses than the ones presented in the previous four chapters.

Much of the knowledge that we employ in social work practice has been obtained from the discovery of relationships between (i.e., only two) or among (three or more) variables. For example, it's helped us to understand why certain individual and social problems may occur, who is most at risk to experience them, and how best to intervene to address them. Knowledge of a non-relationship between or among variables (a so-called null finding) also can be very useful. It helps us to avoid continuing to pursue "dead ends," that is, pursuing possible relationships between variables that probably do not exist or that may be so weak that they are of little usefulness for everyday human work practice situations.

In reality, null findings can help to dispel stereotypes about people or to provide justification for eliminating or modifying human service programs that have not been proven to be very effective. Of course, any of this "knowledge" is useful only if they are valid. If the findings are not valid, however, they may result in bad practice-related decisions being made and are worse than no "knowledge" at all.

How could we mistakenly conclude that variables are related when they are not—or conclude that they are not related when they really are? As we shall see, it can happen in many ways, but several of them are directly related to a common step in the research process known as sampling.

As you know from your study of research methods, research sampling is the selection of a sample of people (or objects) that are supposed to represent a larger group (a population) from which they were drawn. Research samples are used for reasons of efficiency. It's less expensive and often more feasible to study one or more samples than to study an entire population when searching for relationships between or among variables. However, all samples have a major inherent problem—they cannot be assumed to be perfectly representative of the populations from which they were drawn.

When it comes to research sampling issues, the relationship between two variables has four possibilities. The relationship:

exists in the research sample, but does not exist within the population from which it was drawn;

does not exist in the research sample, but exists within the population from which it was drawn;

neither exists in the research sample nor exists within the population from which it was drawn; or

exists in the research sample and in the population from which it was drawn.

The fourth situation is referred to as a true relationship, that is, a relationship between (and among) variables within a population that is accurately reflected within a sample. While we can never be 100 percent certain which of the four situations might have occurred, we can use statistical analyses to arrive at a tentative answer. This entails the use of statistical inference, a concept first introduced in Chapter 1. It's used to answer the question, "How safe would I be if I were to conclude (infer) that a relationship between or among variables that occurred within my research sample was a real one; that is, does the relationship found in my sample also exist within the population from which my sample was drawn?"

Alternative Explanations for Relationships within Samples

Suppose the distribution of attributes for two variables within a sample of cases drawn from a population indicates that they are related—the pattern is readily apparent within our sample of cases. We can only be 100 percent certain that they are related within our sample—and only in our sample. However, in order to make inferences about the relationship between the same two variables within the population from which the sample was drawn, we have to be reasonably confident that the relationship cannot be attributed to other reasons that might explain the relationship—known as alternative explanations. The three alternative explanations are (1) rival hypotheses, (2) weaknesses within the research design, and (3) sampling error.

Page 94-96

Rival Hypotheses

Rival hypotheses refer to the other variables that may have caused our two variables of interest to be related. Some other variables, for example, may have caused both of our variables of interest to co-vary; e.g., they may have intervened between the occurrence of our two variables, or they may have

produced the relationship between them in some other, more complicated way. (Other variables also can cause two variables that really are related to appear unrelated in a research sample; that is, they can obscure a true relationship between variables.)

In some research studies, such as those that use the one-group cross-sectional survey design, researchers simply want to learn if there is a relationship between two variables: for example, job performance and job satisfaction among social workers in a human service agency. (They would not use the labels “independent” and “dependent” variable.”). Job satisfaction may be related to job performance, but the relationship may not be a cause–effect one or even a true one. That is because both variables may also be related to variables such as age, skill level, supervisory structure, reward systems, and educational level, among a host of other variables that can be considered rival hypotheses.

Testing the Null Hypothesis

We never refer to a research hypothesis as proven or not proven; we state only that we found support for it or that we did not. This conclusion is based on whether, using our knowledge of research design and statistical analyses, we feel that all other explanations for a relationship between variables in a research sample (besides a true one) are highly unlikely.

We also never test a one-tailed or two-tailed research hypothesis directly (see Chapter 1).Instead, we test its null form, what’s referred to as simply the null hypothesis. The null hypothesis should not be confused with the null research hypothesis (the prediction that variables will be found to be unrelated), which we discussed in Chapter 1 as one of three forms of research hypotheses.

The null hypothesis is the contention that any apparent relationship between or among variables within a sample (or samples) is really the work of sampling error. (It does not address the issues of measurement bias or sampling bias.) The null hypothesis is thus a kind of skeptic that must be effectively dismissed or discredited before we can claim support for a one-tailed or two-tailed research hypothesis.

What if we wanted to demonstrate support for a null research hypothesis—a research hypothesis that predicts that variables will be found to be unrelated? Let’s assume for the moment that our research study has adequately controlled for rival hypotheses and research design flaws. There might still be a weak relationship between the two variables in the sample. Does that mean our null research hypothesis was incorrect and that the variables really are related in the population from which the sample was drawn? Not necessarily.

We could still gain support for our null research hypothesis through statistical analyses. Using statistical analyses, we could demonstrate that the relationship between the two variables in the sample is really so small or weak that it's fairly likely to have been produced by sampling error. This would be the case if a statistical analysis produced a p value greater than the predetermined rejection level (usually .05), or even much greater such as .50, or even .75. In other words, we would hope to show that there is not enough strong evidence to reject sampling error as the explanation for the apparent relationship between the two variables. If we could do this, we would then have demonstrated support for our null research hypothesis by being unable to "reject the null hypothesis."

In the previous example, you decided that you had justification for stating the following one-tailed hypothesis:

Within the hospital system, female medical social workers will have higher levels of job satisfaction than male medical social workers.

Restated in its null form:

Within the hospital system, there is no relationship between the gender of medical social workers and their levels of job satisfaction. (Any apparent relationship was too likely to be the work of sampling error.)

As you analyzed the data set, the scores of males and females might seem to provide a reason to reject (or not to reject) the null hypothesis, particularly because the number of social workers is small, and it would be easy to identify any differences in the distribution of job-satisfaction scores between the two genders. If, for example, the female group and the male group both had a mean job-satisfaction level of 60, you would not reject the null hypothesis because, even within the sample, there is no difference between the mean job-satisfaction levels for males and females. No additional statistical analyses would be required.

On the other hand, if the females were found to have a mean job-satisfaction level of 90 and the males had a mean job-satisfaction level of 10, you would feel that you had very strong support for the rejection of the null hypothesis. Additional statistical analyses might again be unnecessary. However, such a clear pattern of a non-relationship or relationship between variables is rare. If it exists, you probably already know about it. But what if (as more often occurs) there is a difference in the mean job-satisfaction levels of the two groups, but it's not so dramatic?

What if the mean job-satisfaction level was, say, 60 for females and 50 for males? How likely is it that the relationship within your small sample is a true one that represents a relationship between the two variables within the larger population—medical social workers within your hospital system? Could you infer, based on these central tendency data alone, that as a group, female medical social workers (including those not in our sample) possess a higher mean level of job satisfaction than males?

Such a conclusion would seem a little premature, even if you had carefully designed and implemented the research study so that you were fairly certain that rival hypotheses and research design flaws could be ruled out as alternative explanations for the relationship within your sample. But what about sampling error? After all, 20 cases is a pretty small sample. Any relationship between the two variables could still be the work of sampling error (the null hypothesis). Fortunately, inferential statistical analyses in the form of statistical tests could tell you the mathematical probability that sampling error might have produced this large a difference in the mean job-satisfaction levels between the female and male social workers.

In attempting to gain support for your research hypothesis, you can never totally eliminate sampling error as the explanation for an apparent relationship between your variables. After all, in your small-scale study, even if females did have a mean job-satisfaction level of 90 and males had a mean job-satisfaction level of 10, you theoretically could have drawn a highly unusual sample of 20 in which females might differ by that much from males, even if the two variables really are unrelated within the population from which your sample was drawn.

Although you'll never be 100 percent certain that what you found was not the work of sampling error, you generally can be "reasonably certain" that what you observed was not a fluke occurrence caused by sampling error. In this example, should the conventional $p < .05$ be used as the rejection level where you would reject the null hypothesis? Or should you use a higher or lower p value as the cutoff point?

This decision requires you to rely on your ethical principles and plain old common sense. You do not want to report a relationship between variables that appears to be real when it's not. At the same time, you do not want to be so rigid or so unreasonable that you won't claim support for a relationship between variables just because there is a very remote possibility that sampling error may have produced it. If you did that, few, if any, research findings would ever see the light of day.

Page 104-107

The Hypothesis Testing Process

At this point it's useful to summarize what has been covered so far. The process of hypothesis testing is an orderly one—it follows certain logical steps:

State the research hypothesis. Where appropriate, specify the independent and dependent variables. Describe how your variables were conceptualized and operationalized.

Remember: The research hypothesis guides the entire conceptualization and operationalization process. In addition, it helps determine the best research design to use in an effort to rule out alternative explanations that could explain an apparent relationship between variables. Research hypotheses, if appropriate, should be supported by theoretical assumptions derived from a literature review.

A research hypothesis can be stated as one-tailed or two-tailed (and, less frequently, as a null research hypothesis). A one-tailed or two-tailed research hypothesis is not statistically tested; it's either indirectly supported—or not supported—by testing its corresponding null hypothesis. Another important point to remember is that not all research studies contain an independent or dependent variable. Some may simply wish to find out more information on a single variable such as is done within many simple needs assessments. A needs assessment may wish, for example, to ascertain the level of gang activity within a specific neighborhood. If this was the case, descriptive statistics are all that is needed to describe the variable, gang activity.

State the null form of the research hypothesis. Conceptualize what the data would look like if the null hypothesis were, in fact, correct.

Remember: The null hypothesis states that any relationship between the variables in your sample is likely to be the work of sampling error. It covers all outcome possibilities that are not explicit in the research hypothesis. Thus, if a directional research hypothesis states that the level of job satisfaction for medical social workers has increased significantly over the last 4 years, the null hypothesis includes the possibilities that job satisfaction has not changed and also that it has decreased.

If the null hypothesis can be rejected, support is present for the one-tailed or two-tailed research hypothesis. Hypothesis testing is a process of indirect proof. We never directly prove that the research hypothesis is correct; rather, if the null hypothesis can be rejected, the one-tailed or two-tailed research hypothesis (the alternative hypothesis) is supported only indirectly. In those rare instances where there is a null research hypothesis (predicting no relationship between variables), support for the research hypothesis is achieved if the statistical analysis does not allow the researcher to reject the null hypothesis.

Specify the statistical rejection level to be used. If any level other than .05 is to be used, specify the justification for its use. Which type of error, Type I or Type II, are you more interested in seeking to avoid, and why?

Remember: This step entails determining the consequences of erroneously rejecting—or not rejecting—the null hypothesis. The possibility of Type I errors can be minimized (among other ways) by setting a lower rejection level (e.g., .01, .025) rather than a higher one (e.g., .10, .15). This, however, would increase the likelihood of a Type II error. A statistical power analysis can help you estimate the likelihood of the latter.

State all assumptions about the data and how they were collected. What levels of measurement are assumed to exist within your data set? Which variables are assumed to be normally distributed? What specific sampling methods did you use and why? How large is your sample?

Remember: You must know the level of measurement for each of the variables to be analyzed. The method by which your sample is drawn, sample size, and the way the variables are distributed in their population also affect the choice of which statistical test to use.

Describe the most relevant characteristics of the sample and/or population. Select and compute one or more inferential statistical test(s) to test the research hypothesis. Is each test used appropriately for the conditions described in Step 4? Is a statistical consultant to be used in the selection of the tests? Is computation of the test to be computer generated? If so, what statistical software package will you use? Conceptually, how will each test generate a p value?

Remember: The statistical test used does not determine the degree to which rival hypotheses or research design flaws (e.g., measurement error, sampling bias) may have helped create an apparent relationship between variables in your research sample. It only determines the likelihood that sampling error may have played a role in your study's findings.

Determine whether the relationship between variables is statistically significant. Is the probability (p value) smaller than the predetermined rejection level? If it is, and if a one-tailed research hypothesis was used, is the direction of the relationship that was found the same as stated in your hypothesis? If so, are you reasonably certain that other factors (e.g., rival hypotheses, research design flaws) did not produce the relationship between the variables within your sample? If so, it's safe to reject the null hypothesis and to conclude that a true relationship between the variables probably exists within the population.

Remember: You can never be 100 percent certain that sampling error did not produce an apparent relationship between variables. When you reject the null hypothesis, you're only saying that you are reasonably certain that the apparent relationship is a true one, but you might still be wrong—you may have drawn a very atypical sample because of sampling error.

Determine whether each statistically significant relationship is meaningful. To what degree did your sample size contribute to statistical significance? How strong is the absolute relationship between the variables (effect size)? How surprising are the study's findings? How useful are they for social work practitioners, administrators, educators, policy makers, or researchers? To what extent would you feel safe in generalizing the findings beyond your study's sample to the population from which it was drawn?

Remember: These decisions require a thoughtful combination of ethics, common sense, convention, and practice expertise. They must be addressed before your study's findings could be implemented in social work practice situations.

Concluding Thoughts

This chapter briefly examined the underlying logic of statistical testing for relationships between and among variables. Statistical analyses help us determine the mathematical probability that an apparent relationship between variables within a research sample may have been produced by sampling error—the natural tendency of a sample to differ from the population from which it was drawn. Technically, a statistical analysis tests the null hypothesis (the theoretical assertion that variables really are unrelated and only appear to be related because of sampling error) and determines the probability that it may be correct. Thus, it indirectly tests a one- or two-tailed research hypothesis. Only if the null hypothesis can be rejected (there is only a small possibility that sampling error produced a relationship between variables in a research sample) can we claim support for either a one-tailed or a two-tailed research hypothesis.

We have briefly emphasized what statistical significance means—that sampling error is unlikely to have produced a relationship between variables found in a research sample. We also emphasized some of the more common misunderstandings that exist about statistical significance.

When, after evaluating the likelihood that something else may have produced a relationship between variables within a research sample, we reject the null hypothesis and conclude that the relationship between variables is statistically significant, we run the risk of a Type I error. If we fail to reject the null hypothesis, the relationship may in fact be a true one, and we will have committed a Type II error. In deciding whether to reject the null form of a research hypothesis, we must consider the ethical implications of making either a Type I or a Type II error. If one were to occur, which would be the more acceptable one?

If we correctly reject the null hypothesis and conclude that the two variables probably are related in the population from which a sample was drawn, we also still must address the issue of the practical value, or meaningfulness of the finding. Even if we have statistical support for the likelihood of a true relationship between variables (in the form of statistical significance), the relationship may still be very weak, well known, or otherwise meaningless. We examined some ways to begin to evaluate these possibilities. Decisions about the value of a relationship between variables should always be made with reference to their potential to benefit or to harm those served, our clients.