Hypothesis Testing

Psychological research is primarily aimed at creating and testing theories. Most research uses the deductive approach outlined in a previous chapter to generate specific predictions from more general theories. The process of developing detailed predictions in order to examine the validity of a theory is called hypothesis testing. In a typical research program, the researchers begin with a theory, look for implications or applications of the theory to human thought or behavior, then develop research studies that can empirically measure these effects. They develop specific hypotheses for each empirical test of the theory that indicate exactly what research outcome is expected.

Altemeyer Scale (Right Wing Authoritarianism Scale)

These four items represent a fraction of the full RWA scale. (Items 2 and 4 have been modified by the author.)

Directions: Write in the number:

-4 if you very strongly disagree with the statement.
-3 if you strongly disagree with the statement.
-2 if you moderately disagree with the statement.
-1 if you slightly disagree with the statement.

1 if you slightly agree with the statement.
2 if you moderately agree with the statement.
3 if you strongly agree with the statement.
4 if you very strongly agree with the statement.

____ 1. It is best to treat dissenters with leniency and an open mind, since new ideas are the lifeblood of progressive change.

____ 2.The self-righteous "forces of law and order" threaten freedom in our country a lot more than most of the groups they claim are "godless secular humanists."

____ 3. Obedience and respect for authority are the most important virtues children should learn.

____ 4. It would be best for American society if the proper authorities censored the internet to keep pornographic material away from children.

Scoring: Calculate #3 + #4 - #1 - #2These four items alone are a poor measure of the construct, so don't read to much into your score.

For example, social psychology theorizing on the personality trait "authoritarianism" suggests that the authoritarian personality is a package ("syndrome") of traits that include black-and-white thinking, submission to powerful people, abuse of less powerful people, strong conservatism, rejection of psychological ways of thinking, racial prejudice, anti-Semitism, and concern with others' sexual activities. The theory, using Freudian principles, claims that the syndrome springs from one's experience in a family: parents' child-rearing methods and their attitudes and beliefs. Highly authoritarian people are said to be ripe for extremist ideology when historical circumstances make such ideas available, the case in point being acceptance of Nazi ideas in 1920s and 1930s Germany. In modern times, one might look to Slobodon Milosovich (ex-Prime Minister of Serbia/Yugoslavia) or, at a less extreme level, Rush Limbaugh.

From this theory, researchers have generated many interesting hypotheses. For example, the theory implies that children of parents who use harsh and inconsistent methods of discipline will become authoritarian adults. To study this relationship, "harsh and inconsistent child-rearing practices" must be operationalized as a measureable variable and authoritarianism must be adequately assessed in an operational measure such as the Right Wing Authoritarianism scale (see sidebar). The resulting hypothesis might be:

> H_A: children of parents who were harsh and inconsistent will be more authoritarian than children of parents who were not harsh and inconsistent

H_A is referred to as the alternate hypothesis or research hypothesis. Some books use H_1 instead of H_A.

Hypothesis testing is performed within a philosophical orientation to science that claims that theories can be falsified but never proven (see chapter Theory-2 for an explanation of this concept). This counter-intuitive idea gives modern hypothesis testing its counter-intuitive structure: hypotheses are phrased not as what one expects to find, but rather as what one would expect if the theory were *false*:

> H_0: children of parents who were harsh and inconsistent will be of equal authoritarianism as children of parents who were not harsh and inconsistent

The subscripted zero after the "H" indicates that this is the null hypothesis, where "null" means "no difference." It can be written this way:

$$H_0: A_h = A_{nh}$$

where A = level of authoritarianism, h=harsh family, nh=not harsh family

The alternate hypothesis, H_A:

$$H_A: A_h \neq A_{nh}$$

The task of the researcher is to perform a study that can test H_0 adequately. The researcher's goal is to reject H_0, that is, show that it cannot be true given the research data that are obtained. If H_0 is rejected, then H_A is accepted. However, if the data come out in such a way that H_0 cannot be rejected, we *cannot* accept H_0.

However, the hypotheses as stated so far are still not quite correct. Note that H_A states that harsh and non-harsh parents will produce *different* kinds of children ($\neq$), but the implication of the theory is that harsh parents' kids will be *more* authoritarian. $H_A: A_h \neq A_{nh}$ is a nondirectional hypothesis, it just says that harsh parents' kids will be different than non-harsh parents' kids, but not which direction this difference will take. In statistics, nondirectional hypotheses are analyzed using two-tailed tests.The theory clearly states the direction, however, so we need to include this in the hypotheses. The directional alternate hypothesis would be written as:

$$H_A: A_h > A_{nh}$$

And the corresponding null hypothesis as:

$$H_0: A_h \leq A_{nh}$$

Note that combining H_A and H_0 yields all possible outcomes: equal, less than, and greater than. Rejecting H_0 means that A_h is neither equal to nor less than A_{nh}, which leaves the researcher clearly with H_A. In statistics, directional hypotheses are analyzed using one-tailed tests.

The research on authoritarianism supports this part of the theory. Children of harsh, inconsistent parents are found to be more authoritarian than children of non-harsh, consistent parents. Analysis of such research data would, minimally, calculate a mean authoritarianism score for each group and compare them statistically. If authoritarianism were measured on a 100-point scale, the research might find something like this:

Mean $A_h = 65$

Mean $A_{nh} = 45$

Depending on the size of the research sample and how consistent the data were with each other, a statistical analysis would usually find that 65 is indeed larger than 45, so the null hypothesis $A_h \leq A_{nh}$ would be rejected.

Why can't we accept the null hypothesis if it is not rejected in our study? In empirical research, we almost always use samples of subjects or research participants, not the entire population. A study of child rearing and authoritarianism might collect a sample of 50 harsh and 50 non-harsh families. The world has more than 100 families, so if we find that in *this* sample of 100 families there is no difference between harsh and non-harsh families, we can't be certain that some other sample would not find a difference. How many 100-family samples might be taken from 6,000,000,000-plus people?

This logic is not quite the same when we reject the null hypothesis. Finding just one sample in which the null hypothesis is rejected is sufficient to make the claim that the two groups are not the same. Still makes no sense? Read the paragraphs again?

You may respond to this logic by thinking that it is some sort of picky, formalistic nonsense: when research fails to find a difference, maybe there really isn't one. If you feel this way, you are in line with actual practice in science. Although we can technically never accept the null hypothesis, when a lot of good research fails to reject it, we will eventually come to accept it. In reporting these patterns of findings in scientific publications we might say "research has failed to find a difference between the willingness of males and females to deliver pain in social psychology research" even when we have come to believe that "research has found that males and females are equally likely to deliver pain." Note that the first of these two statements is more cautious.

Type 1 and Type 2 Errors

Hypothesis testing is only as valid or correct as the research procedures upon which it relies. Psychological research is inherently messy, and mistakes are made. Methodologists have developed a terminology to describe some of the ways that hypothesis testing can go wrong. In quantitative research, samples are collected, data are obtained from the people in the sample, and statistical analyses are performed to detect differences between groups and relationships among variables. Just exactly how this is done is covered in a statistics class.

Digression into Statistics

In order to explain hypothesis testing errors, a short digression into data analysis is necessary. (Sorry.) The basic issue in statistics is that we are working with probabilities, not absolutes (in contrast to some of the natural sciences). When data are obtained from people or animals, there is always variability in the data due to the many unaccountable factors influencing their responses. Furthermore, any sample collected for a study will, by chance, be composed of a slightly different mix of people than any other sample. Comparisons between groups of people are always made against a backdrop of two problems: (1) variability within each sample due to differences among the people in the sample, and (2) a consideration of the fact that every sample will be slightly different from every other sample that could have been randomly selected. The first of these problems is called, among other things "within-group variance," "error variance," "noise," and "ignorance" (on the part of the researcher). The second one is an issue of sample size: it gets more serious when samples get smaller.

In statistics, calculations are performed that compare the observed mean difference between the groups to the amount of variability (error variance) within the groups. For example, in a t-test:

$$t = \frac{\text{Mean difference}}{\text{Standard Error}}$$

You will learn more about t-tests later in this book. For now, suffice it to say that bigger *t* values are better in rejecting the null hypothesis. Standard error in this formula is a function of both of the problems described in the previous paragraph: the within-group variability and the sample size. As the mean difference becomes larger (e.g., the difference between A_h and A_{nh} in the example), *t* will increase. As the standard error get larger, *t* will decrease. (Error variance is bad.)

As *t* in this formula increases, the strength of the difference between groups (e.g., harsh vs. non-harsh families) gets stronger. At some point, we would say that the difference is strong enough to conclude that the null hypothesis should be rejected. How strong is strong enough?

Statistics is about probabilities. Psychology and social science have come to the arbitrary conclusion that the answer to the "strong enough?" question is: when there is only a 5% chance that we're wrong. In other words, when the data analysis tells us that there is a .05 probability (on a 0 to 1 scale) or less that we would be wrong in rejecting the null hypothesis, we're allowed to reject it. This value is termed the "alpha level" in statistics. You will learn how we determine when the probability reaches .05 in a statistics course. When an alpha level of .05 is used to determine when it's safe to reject the null hypothesis, of course we will be wrong 5% of the time (Type 1 Error; see below). Does this mean that one in every 20 published studies is wrong? Not really...researchers think about their data in more complex ways than this.

Hypothesis Testing (continued)

In testing our null hypothesis, we will conclude either that it should be rejected or it should not be rejected. Assuming that there is a real world (the Physics Model), our conclusion will be correct or not correct. A correct conclusion would be either that we should reject the null, and in fact (in the real world) there is a difference between the groups; or that we should not reject the null, and in fact the groups are equal in the real world. (We can say "equal" in this context because we are talking about the *real* world, not the world that we perceive through research.) There are of course two incorrect conclusions: we reject the null, but the groups are actually equal; we don't reject the null, but they are in fact different. Out of this comes a 2 X 2 scheme:

		The Real World	
		Groups are Equal	Groups are Different (H_0 is false)
Your Conclusion	Reject Null Hypothesis	***Type I Error***	***You are Correct*** *(You Win)*
	Don't Reject Null Hypothesis	***You are Correct***	***Type 2 Error***

Type I Error: Your statistical analysis tells you that there is only a .05 probability that the groups are not different, so you reject the null hypothesis. But this is the one time in 20 where in fact the groups *are* equal. The probability of committing a Type I Error is the alpha level that you used (usually .05).

Type 2 Error: Your statistical analysis tells you that there is greater than a .05 probability that the groups are not different, so you decide to not reject the null hypothesis. But this is the time when in fact your data have led you astray and the groups really are different. The probability of committing a Type 2 error is difficult to calculate.

You Win: The upper/right cell, "reject H0 and you are correct," represents the goal of most studies: predicting a difference, then finding it.

You Sort of Lose: The lower/left cell, "don't reject H_0 and you are correct," is an outcome in which your study has failed in the sense that you didn't find any differences, but at least you didn't blunder into claiming any differences that weren't really present. This situation is common in science, and begs the question: was it the theory, or the research methods? A faulty theory might lead a scientist to formulate a hypothesis that won't be supported, and the hypothesis itself may be an inappropriate deduction from the theory. On the other hand, the theory could be OK and the hypothesis deduction sound, but the methods used to test the hypothesis might have been at fault. Which is it? The answer is: do additional research. Most serious research in psychology is conducted in research programs, not one-shot studies, so the researchers have developed methods that seem to work, and they can evaluate their findings in light of other studies on similar topics using similar methods.

Relationships, Not Differences

So far in this chapter I have used examples and concepts based on research that seeks differences between groups, for example, differences in children's authoritarianism between a group of harsh parents and a group of non-harsh parents. Psychologists also perform research designed to look for relationships between variables. For example, the authoritarianism research described previously could have been done by measuring amount of harshness/inconsistent discipline in a sample of families, then looking for a relationship between this measure and a measure of children's authoritarianism.

In this type of research, the null hypothesis would be that there is no relationship between harshness and authoritarianism, and the alternate hypothesis would be that there is a positive relationship (i.e., as harshness increases, authoritarianism increases).

H_0: Harshness is not related to authoritarianism (or is negatively related)

H_A: Harshness is related positively to authoritarianism

Much social science research looks at relationships rather than differences. Research on authoritarianism has found that it is related to a variety of social attitudes: political orientation, stands on social issues such as abortion, tolerance of free speech, social class, and so on. However, at a more basic level, differences and relationships are the same thing. The frequent finding that anti-war protestors are lower on authoritarianism than others (a difference) actually shows that there is a relationship between the variables "willingness to protest" and (non-) authoritarianism.

Conclusion

This chapter has presented hypothesis testing as if it were a simple, straightforward procedure based on clear logical rules. In a statistics class, you must approach your data analyses within this formal approach. However, in on-the-ground research, scientists think about their theories and hypotheses in a more complex way, viewing the concepts, operationalizations, and research results in a comprehensive "global" manner that tells them how the theory is progressing—or not.