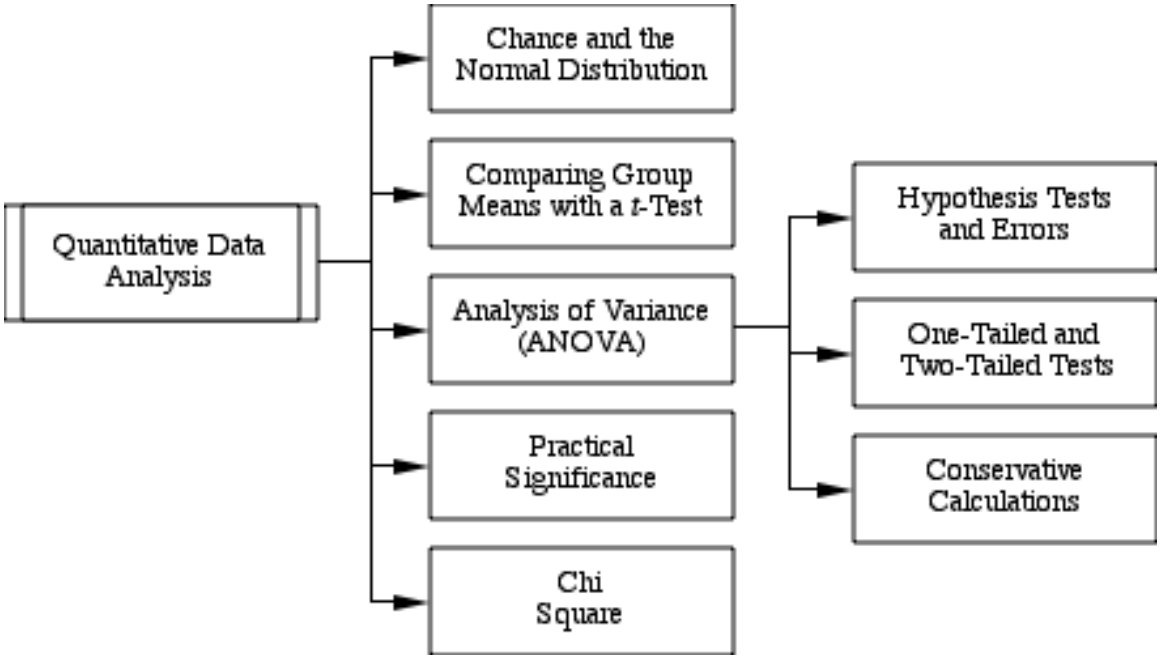
Chapter 8: Analyzing Data: Inferential Statistics



Chapter Overview

Chapter 8 is the quantitative counterpart to Chapter 6. This chapter provides instruction on statistical data analyses, focusing on those most commonly used by classroom teachers in their research. We will not include multiple regression, analysis of covariance, or factorial analysis as most beginning researchers do not design a study that requires these types of statistical analyses.

INTRODUCTION

We like to believe that the things we do have some positive impact. Sometimes this is an illusion. As an example, not long ago I wanted to see if I could get better gas mileage in my car. Because I have a long commute from home to work, I drive mostly in the left lane. I wondered if the slower pace of the right lane would make a difference in

my gas mileage. I switched to the right lane and carefully computed the gas mileage for a week. To my joy the mileage was a bit better. But, alas, over the next few weeks it turned out that right lane driving was actually worse. All the stop and go caused the vehicle to consume more gas. The first week's mpg was simply an example of normal random variation in my mileage and switching had not really made a difference. The illusion is that often we believe we have affected something in our lives but the reality is that differences may appear for other reasons.

Statistics can help us sort out the normal variation in the things we are trying to do and help us see when we really have had an important impact. The tools to do this fall in the category of **inferential statistics**.

Studies that use inferential statistics (studies that look to see if chance is a good explanation for a result appearing) are designed in a lot of different ways. We will talk about some of the designs that are likely to be most useful to you in the next chapter. Right now you need to know that the way to organize the variables in your study, regardless of its design, is to identify **dependent** and **independent variables**. A dependent variable is the thing you are measuring. If you want to know if students learned more, then test grades might be the dependent variable. If you want to know if boys were taller than girls in your class then height would be the dependent variable. If you want to know if one group of teachers had been teaching longer than another then years of teaching would be the dependent variable.

Independent variables are called the grouping variable. In inferential studies you will be comparing one group to another or comparing the same people before and after something happened to them (an intervention). So, the independent variable is the

variable that identifies which comparative group someone is in. In the case of where the dependent variable was learning the independent might be when the scores were gathered (before the intervention or afterward). In the case of comparing the heights of girls and boys the independent variable would be gender. In the case of looking at years of teaching the independent variable might be whether the teachers were elementary or secondary teachers. You get the idea. Generally, you will gather independent and dependent data for everyone in your study. Independent variable data tell you who is in what group that will be compared and dependent data gathered will be the same for everyone so that you can compare the groups.

CHANCE AND THE NORMAL DISTRIBUTION

As educational researchers we spend a lot of time trying to figure out if something happened by chance (i.e., it is likely to have occurred randomly) or if it happened because of something that affected what we were measuring. Did a curriculum really improve student learning? Did a new conflict management program really reduce student referrals? Did an exercise program for 3rd graders really increase time on task? We want to know, like the gas mileage reduction example discussed earlier, that change we see after we have done something cannot easily be explained as something that is likely to have occurred by chance. If it is unlikely to have occurred by chance then we can have confidence that what we did actually had a predictable effect.

In Chapter 7 we figured out how to compute percentile ranks. Percentile ranks were the percentage of scores that appeared lower than a given score in a normal distribution. A percentile rank of 90 meant that 90 percent of the comparison group scored lower than

the score that represents the 90th percentile. Conversely that would mean that 10 percent scored at that point or higher. If we were interested in any given student before they took the test we could ask how likely is it that that student would score in the 90th percentile or higher. Since 10 percent of the students score at those levels you could say the student had a one in ten chance (10 percent) of scoring that high even if we knew nothing else about that student.

Instead of looking an individual student's score, how could we determine the probability that a mean score for a group would appear? We said earlier that statistics is almost never about looking at individuals but rather about understanding groups. So, we need to understand how to look at what would seem to be the percentile rank of a group. The procedure for doing this looks similar to what we did to determine percentile rank for an individual, but it is a bit more complicated. The 20 students in Mrs. Johnson's math class take a standardized math exam. The mean score for the class is 81.47. How does Mrs. Johnson's class compare to other classes taking this test? Trying to figure out the percentile rank of 81.47 has a problem. The normal distribution for the test is based on individual scores, and our mean score is based on averaging the scores from a group of 20. Those two are not the same. To overcome this problem a new normal distribution is built based on the mean scores of comparable sized groups; in this case, a group of 20. If the mean score of every possible combination of 20 students was determined and then plotted to show a new normal distribution it would look very similar to the original distribution based on individual scores. If random groups of 20 students were selected, some averages for these groups would be higher than the mean for the whole group and some would be less. After enough randomly selected groups of 20 the mean scores will

begin to cluster around the whole group mean with fewer considerably higher and lower than the whole group mean. What happens is the new distribution of means of groups of 20 ends up having the same mean as the whole group, but because it is based on means of 20 students, fewer means appear as you move farther above and below the mean than would be the case with the distribution based on individual scores. This new distribution has a name—**sampling distribution of the mean**—and the new standard deviation that appears also has a name—**standard error**. In other words, the same parameters we examined for individual scores have a counterpart when we examine group or population scores: the mean becomes the mean of the sampling distribution and the standard deviation is replaced by standard error (see Figure 8-1).

[Insert Figure 8-1 here]

With this new distribution we could figure out how likely it was that Mrs. Johnson's class mean score would have appeared by chance. This might be nice to know because if it was really unlikely to appear by chance, we might begin to believe that there was something unique about the students in her class or her teaching.

Unfortunately, when we are doing our own research it is rare that our study collects scores on standardized assessments. Most of the time we are looking at some assessment to which there is no larger group to compare; that is we use criterion referenced vs. norm referenced assessments. If Mr. Smith teaches a unit on dinosaurs to his second graders, it would be common for him to use an assessment he designed himself. So how could we know if Mr. Smith's students' scores on the test could be explained as likely to occur by

chance or, as we would hope, more likely to have occurred because of Mr. Smith's instruction on dinosaurs? Since there is no population to which to compare Mr. Smith's class, some other group needs to stand in as the comparison group. In this case the group that took the pretest on the dinosaur curriculum will be the comparison group; that is, his class at the start of the unit. Is the group that took the pretest somehow now a different group after having experienced the curriculum?

The last bit of statistical maneuvering that is needed to accomplish this task is called **estimating parameters**. If we do not know the mean of the sampling distribution or the standard error, then they have to be estimated. From Mr. Smith's students' pre-test scores we could easily compute the mean and standard deviation and we know the group size. We need the group size to estimate the sampling distribution of the mean, the standard deviation to estimate the standard error and the mean of the group to estimate the mean of the sampling distribution of the mean. At this point you may be beginning to glaze over. Think back to percentile ranks. What we want to know (in theory) is the percentile rank of Mr. Smith's class in some theoretical distribution of a number of other classes of 20 students. If Mr. Smith's post-test scores were unlikely to occur by chance in this distribution that would be an indication that Mr. Smith probably taught the unit well (or poorly if the scores went down). Fortunately, you will never have to do all of these theoretical calculations. The computer will do them for you.

Take a look at Figure 8-2. The idea is to find the position of the post-test score in relationship to the theoretical distribution generated by estimating the parameters from the pre-test score. This works whenever you are comparing two group means even if they

are not pre-test/post-test; for instance when you would compare two classes in which you used different instructional strategies.

[Insert Figure 8-2 here]

COMPARING GROUP MEANS WITH A *t*-TEST

In the case of percentile rank, an individual score is being compared to the larger group. From the last chapter, you will recall that the calculation to do that is called a z-score (the distance a score is from the mean in units of standard deviation). When we are looking at a group mean in comparison to another group the idea is the same—how far is a group's score from the mean of a distribution of scores of groups of the same size in units of standard error? The calculation to do this is similar to the z-score, but because it involves estimating parameters, it is not the same. It is called a *t*-score or more familiarly, after all the calculations are done, a ***t*-test**. Just as a reminder, because we are dealing with mean scores, we are talking about analyzing interval data.

Here is an example. Mrs. Greene gives her students a pre-test on earthquake facts. She then teaches a unit on earthquakes followed by the test again. She puts all of the scores into a spreadsheet and then tells the computer to compare the two sets of scores, the pre-test and the post-test, using a *t* -test. What the computer returns is a **probability** (a *p* value) that the two sets of scores were likely to have occurred from the same group by chance. If the probability is high that they could have appeared by chance, there is no evidence that Mrs. Greene's teaching had any impact. If the probability was low that the group means could have occurred randomly, then we would assume that the second set of

scores came from a different group or, in our case, from a group that had been changed by something. The most likely reason for the difference is the teaching. Good thing.

This sounds easy but we have a subjective judgment to make here. At what point could we say that a difference in mean scores is so unlikely to have occurred by chance that it did not occur by chance; that is, a **significant difference** occurred? Researchers make different decisions on when groups are significantly different based on the importance of being wrong about that decision. In other words, how sure do we have to be that the difference is the result of something we did and not have occurred purely by chance? If a researcher is testing a drug to see if it has critical side effects, that researcher is going to want to be really sure that the results did not occur by chance. In that case, she might set the significant difference level at one in ten thousand. She would want to be as sure as possible that the drug did no harm.

With social science research we are seldom this exacting. Generally, we set our probability level at .05 or 5%. When studying the behavior and opinions of people, we assume that if there is less than 5% chance (1 in 20 chance) that group differences could have appeared randomly, then they did not appear randomly. They appeared because there was an actual difference between the two groups. In statistics, probabilities are always expressed as a decimal rather than a percent so a probability of random occurrence 5 percent of the time would be .05. To say this slightly differently, when comparing groups, we want to know that the probability that the group mean differences could have occurred by chance is less than .05. If we can show that the group mean differences are likely to occur by chance less than 5 percent of the time ($p < .05$) then we can say that the group differences are **statistically significant**. (As a habit when you are

writing reports on quantitative research you should not use the word "significant" except when referring to statistical significance.)

Think back to Mrs. Greene and her earthquake test. Here is what is conceptually going on when she tests for group mean differences (does a *t*-test). She gives a pretest. Using estimated parameters the computer generates a theoretical distribution of group mean scores from groups the same size as her class. Then when she gives the posttest the computer figures out how likely it is that the second group mean score could have appeared randomly in the theoretical distribution of group mean scores on the pretest. If it is highly probable (more than a 5 percent chance) that the second score could have appeared randomly, she would conclude that there is no difference in the two groups. In other words, the children had not changed in a statistically significant way in their ability to answer questions on the test after instruction. If the probability that the group mean difference could have appeared by chance is really low (less than 5 percent), then she can conclude that her students are not the same as the group that took the pretest. There is a statistically significant difference, presumably due to her instruction. She knows this instantly after the computer does the *t*-test because the computer generates a *p* value. If *p* is less than 5 percent (written $p < .05$), there is a statistically significant difference.

When you are doing research in which you are comparing group mean scores what you care about is whether the group mean differences are significant (unlikely to have occurred by chance). When the computer calculates the *p* value it will be a precise value, like .035. What is important is whether the *p* value is smaller than the level you as a researcher have set to test for significance. How much smaller the number is doesn't really make that much difference. Even so, in the 6th edition of the APA manual APA is

suggesting that the actual values generated by the calculation are put into the tables in the results portion of your study report. You and your research mentor need to agree on what style of presentation makes the most sense.

Hypothesis Testing and Errors

When you read quantitative research in which group comparisons are reported, often the reports will be constructed by reporting whether group differences are significant, much the way that it was described above. We feel quite confident making conclusions based on statistically significant differences, but those conclusions may not be correct 100 percent of the time. The truth is that we can never know things absolutely. What we find is the best explanation for the moment but leaving open the possibility that a disconfirming fact may still be discovered. So, researchers end up doing something that always sounds way more complicated than it needs to be. They do studies in which they try to reject the explanation that what they want to find is not true.

Here is what that looks like: I think that implementing a new discipline strategy wherein students are taught conflict management will reduce referrals in my school. My *hypothesis* is that if I compare a group that was taught the new strategy to a group that was not, I will see significantly fewer referrals in the group that received the instruction. In order to show my hypothesis is probably true I am going to start by stating the hypothesis in a form that says it is not true: the new discipline strategy will have no effect on student referral numbers. This is called a **null hypothesis** (null as in none or no difference). After I have gathered the data and run the *t*-test, I find that the probability that there is no difference in the groups (my null-hypothesis) is really small ($p < .05$). I

can now reject the null hypothesis that there is no difference because it is so unlikely there is no difference. This means there probably is a difference. And, the best explanation that I have (although it is possible that there are others) is that the difference in the two groups came from implementing the new discipline strategy.

Trying to keep all of the negatives straight when you are reasoning out hypothesis testing is a daunting task, which is why research reports often omit the statement of the null hypothesis. Procedurally, it makes little difference whether you include the statement of the null hypothesis in your report because readers assume that it exists whether or not you state it. Technically, the null hypothesis has to exist if we are doing good research because we always have to be cautious about the existence of alternative explanations to what we have found. We can never prove that something is true in all cases because we would have to test every possible case--hence, the approach of the null-hypothesis. Practically, you will read many research reports where the null hypothesis is never mentioned. You and your research mentor will have to make decisions about the report style that is best for your work.

Thinking in terms of hypotheses does help us pay attention to an important problem in social science research. Earlier we said that establishing significance levels in research, called **alpha levels,** is based on the importance of not being wrong. If the alpha level is .05 that would mean that you would reject the null hypothesis and believe that the alternative (what you are interested in finding) was likely to be true if the probability that the null hypothesis was true was less than 5 percent. Put another way you could be wrong to reject the null hypothesis 5 percent of the time. We just assume that 5 percent is so small that it is worth the risk of being wrong.

When the null hypothesis is rejected but should not have been (group differences did just appear randomly) that is called a **Type I error** or a false positive. That is, you thought the two groups were significantly different, but they were not. The opposite can be true also. It is possible that the null hypothesis was not rejected but that there really were differences in the groups. If the *t*-test says that the probability that the groups were the same (null hypothesis) was .25 we would say that was a pretty high probability and we would not reject the null. But, it is possible that there really were differences and we just lacked the evidence to be confident that we had seen them. Not rejecting a null hypothesis when there really were differences is called a **Type II error** or a false negative. That is, you thought there was no difference between the two groups when there really was.

Generally, for social scientists, the best way to reduce both Type I and Type II errors (that is, to be truly sure your results are not due to random error) is to have larger groups from which you gather data and to randomly select participants from the populations to which you wish to generalize. Often in our work neither of these are possibilities. Educational researchers are often using intact groups of students or other school personnel in which group sizes are limited and random selection is not a possibility. This is the point at which to remind yourself to be very humble about your findings. We need to work hard in designing research (especially quantitative research) to reduce alternative explanations to our findings, but alternative explanations are almost always possible. Reflecting on what those alternative explanations might be is part of what you will need to write in the conclusions section of your report.

One- Tailed and Two-Tailed Tests

When Mrs. Greene teaches about earthquakes, if the assessment of her students' learning is aligned with the curriculum, it is hard to imagine that the group mean score from pretest to posttest would do anything but go up. In most other cases of doing social science research we cannot be so confident. For instance, using our example of teaching a conflict management strategy, it could be that our efforts had exactly the opposite effect that we had hoped or some other mitigating factor intervened and referrals actually go up in the treatment group. When comparing group mean scores it makes an important difference if we know ahead of time which set of scores will likely be higher than the other. This is easiest to visualize by looking at a normal curve (see Figure 8-3).

[**Insert Figure 8-3 here]**

As researchers, we have set the alpha level for our study at .05. What does this really mean? In most instances, when doing *t*-tests we want to know if the second group mean is so unlikely to occur that it appears in the area representing 5 percent of the mean scores at the far end of the sampling distribution of the mean. If the score does reside in this area then we say it is so unlikely to occur by chance that it did not appear by chance; rather, it occurred for some other reason, which we hope is our intervention. That 5 percent area could be at either end of the distribution. In the case of the dinosaur curriculum we hope to see the class mean on the posttest at the right end of the distribution because we expect student test scores to increase after instruction. In the case of the conflict management strategy we would hope to see the group mean for referrals at

the left end of the distribution (significantly fewer referrals), as teaching conflict management should reduce behavioral referrals.

What if we do not know ahead of time which direction the second group mean will be from the first? What if we are unsure whether the treatment will cause an increase or decrease in the second group mean? Are you absolutely sure using cooperative learning as an instructional strategy will help students learn math facts better or is it possible they may actually do worse? We still have to abide by having the second score appear in the distribution less than 5 percent of the time. If we do not know which end of the curve to look toward, then the 5 percent has to be split between both ends of the curve. The score for the second group mean would have to appear in the area that represents 2.5 percent of curve on either end. If you did not split the 5 percent area and allowed the possibility of 5 percent at each end of the curve, then you would actually have 10 percent of the area available to demonstrate significance—double what you had set for your study.

When you know ahead of time what the direction of change is likely to be for your study, you can establish that the 5 percent area to show significance is in fact all at one end of the distribution. This is called a **one-tailed test**. When you can not anticipate the direction of change, you must split the 5 percent of the area to show significance to both ends of the distribution. This is called a **two-tailed test**. This makes a big difference in your study. Practically it becomes twice as difficult to show significance if you are using a two-tailed test.

As a researcher you have to decide which test to use. Establishing that you can anticipate the direction of change is not always that easy. In cases where it may not be obvious, you should support your claim for using a one-tailed test with other research

related to yours. As you read research papers you will find many examples where researchers inappropriately used a one-tailed test. It is surprisingly difficult in almost all social science research to be absolutely confident of the effect of an intervention. We do research because we are not sure of the answers. Therefore, almost by definition we should be cautious about anticipating the results of our work. When in doubt, use a two-tailed test.

Most statistical programs will give both the one and two-tailed results automatically so you, the researcher, must choose the correct value to report. If the statistical program you are using only provides the one-tailed result, it is very easy to compute the two-tailed result. The *p* value for a two-tailed test is exactly double that of the value generated for a one-tailed.

Conservative Calculations

Part of the art of quantitative research is to design studies that eliminate all possible explanations of changes that are observed except those changes that occurred because of the intervention that you are studying. We want to know that what we did (or what we watched happen) made a difference. The problem is that studying people is really messy. And to make it worse we are usually studying them in natural environments where we have little control over all of the other things that might be affecting behavior other than our intervention. For example, we generally do not test on Mondays or during spirit week or during other times when students' minds might not be totally focused on their work. As another example, we have little control over whether a student studies at home, or how much parental help or encouragement students receive.

To a certain degree, statisticians compensate for this lack of control by making the ability to show significance mathematically more difficult in situations where variation in the groups is more pronounced (one of the things that is hard to control). The calculations are designed to reduce Type I errors (where we have shown an important difference in groups and probably shouldn't have). If we feel the two groups we are comparing were different from the start (e.g., college bound and non-college bound seniors), we choose to do our statistical analyses using **unequal variance**. If we feel the two groups we are comparing are similar (e.g., a control and treatment group for two different seventh grade science curriculum on the same topic), we choose to do our statistical analyses using **equal variance**. A special kind of analysis is done if the two groups being compared are composed of exactly the same individuals (e.g., a pre/post test design); in that instance we can use a **paired t-test**.

Here are more detailed examples of these comparisons. Let's say we are interested in knowing if participation in after school sports affects GPA in high school students. A common way to do this would be to look at the students in a given high school and split them into two groups: those that participate in after school sports during the year and those that do not. Gather all of the GPAs for a given semester and run a *t*-test to see if the groups are significantly different. The problem with this study is that there are so many possible things that could be affecting these students' choices about participating in sports that our "test" may not be measuring the relationship between GPA and sports at all. The two groups are probably enormously different on a number of characteristics to begin with: after school responsibilities, interests, hobbies, etc. We would say that the two groups have unequal variance to mean that the groups are not very much alike.

Caution would be warranted in this case to not get too excited about the results of the *t*-test before beginning to eliminate some of the other possible influencing factors. Since we cannot control for all these variables between the groups, we rely on mathematics to help us. With your statistical program you would automatically make a more conservative estimate of possible significance by choosing a *t*-test of groups with unequal variance.

On the other end of the scale is a case like Mrs. Greene and the earthquake test. The students who took the pretest are exactly the same as the students who took the posttest. In cases where exactly the same people are in the two compared groups we would use a paired t-test. For this situation the differences in the two groups are dramatically reduced and we do not have to worry as much about other differences in the groups accounting for what we thought we saw as a significant difference based on our intervention. The calculation of significance can be less conservative because the groups are so similar.

Depending on the statistical software you are using there is a third choice. Imagine you are comparing two 7^{th} grade classes in the same school. You have every reason to believe that these children are similar in most ways and (this is important) the size of the two classes is essentially the same. In this case we still need to be worried about unaccounted for differences in the two groups, but less so than in the case of GPA and sports participation. This third category is when you would use a *t*-test for groups with equal variance.

Here again the researcher needs to make some subjective judgment. If the two comparison groups have absolutely the same members in both groups (you can not do this if one student was gone on the day of the posttest) then use a paired *t*-test. If you

believe the groups to be similar in most important ways and the two groups are almost the same size then use a *t*-test for groups with equal variance. All other cases, especially when comparing groups of different sizes, use a *t*-test for groups with unequal variance.

ANALYSIS OF VARIANCE (ANOVA)

It is very common that researchers design studies in which more than two groups need to be compared at once. For instance, in a study of strategies for improving reading comprehension you might want to have Group A read the stories only, Group B read the stories and draw pictures to illustrate what they read, and Group C complete a worksheet after reading the stories. After these interventions all of the students would receive the same comprehension test. To find out whether group mean differences were significant, it would seem that three comparisons would need to be made: A with B, A with C, and B with C. Doing three *t*-tests to make the comparisons would violate statistical logic. Statisticians would worry about all of the comparisons being independent of each other. The correct procedure is to use an **Analysis of Variance (ANOVA)** whenever more than two group means are being compared at the same time. An example of how this might look in a results section is shown in Chapter 9 (refer to Figure 9-4).

An ANOVA generates a *p* value interpreted the same way it would be for a *t*-test. It tells whether group mean differences are likely to have occurred by chance. Somewhere in the comparisons of group means one or more of the comparisons is statistically significant if $p < .05$. An ANOVA does the multiple comparisons (in our example, A and B, B and C, and A and C), but just tells you whether a significant difference exists—not which groups are statistically different. To determine which of the comparisons are

significantly different, a further test must be done. These are called **post hoc tests**. Although there are a number of procedures available the two most common are Tukey's and Scheffe's. Generally, Tukey's is used when group sizes are equivalent and Scheffe's when they are not. (Scheffe's is a more conservative calculation under the same reasoning as more conservative *t*-tests above.) Most statistical software will allow you to choose which post hoc analysis to use. Unfortunately, as of the writing of this text, Microsoft Excel does not provide that calculation although there are some third party "add-ins" which do.

When the post hoc analysis is run, the results will give another *p* value for the specific two group comparisons which are subsumed under the larger ANOVA analysis. As a caution, it is possible for a post hoc analysis to show a significant comparison even when the *p* value for the ANOVA is not significant. Always look at the results of the ANOVA before determining if post hoc analysis is warranted. Reporting on significance can be done from the post hoc *p* values, but the results of the ANOVA should also be shown.

Although different statistical programs display results differently, the example shown in Figure 8-4 is typical. Each group mean and standard deviation is calculated. Then the calculations for the ANOVA are presented in an ANOVA Table. In this table, one block is marked as "*p*". This the result of the significance test for the overall ANOVA. In this case it is .002. Then, depending on which post hoc analysis is chosen, a table will display the specific group comparisons. In this case there is no significant difference between groups 1 and 2 ($p = .382$). There are significant differences between groups 1 and 3 ($p = .001$) and 2 and 3 ($p = .025$). In most of the papers, you will read the

significant difference between groups 1 and 3 would be reported as $p < .01$ (you would have to assume that the .001 was rounded from a larger number) and between groups 2 and 3 as $p < .05$. Again, the newer APA guidelines suggest putting the actual p values into the table. With ANOVA tables this presents more challenges than with *t*-test results tables. We provide an example of how this may be done but you need to discuss the style for your report with your research mentor.

In most cases when you are using statistical software, calculations will be reported that are more sophisticated than what you need. If you look at Figure 8-4, you will see many numbers returned in the read-out that you do not need to deal with at this point in time. We want you to be cautious, informed users of statistical analysis, but descriptions of all of the terms reported are beyond the design of this book and the number of terms displayed may vary depending on which statistical program you are using. On the other hand, we hope that this introduction will make you curious enough to investigate further on your own.

[Insert Figure 8-4 here]

PRACTICAL SIGNIFICANCE

In a study of the differences of test scores between boys and girls on a standardized mathematics test in a school district, it might be discovered that boys scored significantly higher than girls. On closer examination of this study we would see that 423 boys and 412 girls took the test and that the mean score for the boys was 67.7 and the mean score for the girls was 66.2. Statistically the differences were unlikely to have happened by

chance ($p < .05$) so we would conclude that boys on average scored significantly better on the test than girls. But, is a 1.5 point difference really meaningful? When group sizes grow very large, statistical significance appears with smaller and smaller numerical differences between the groups. In our example the significant difference is real, but it represents a very small difference in the actual mean scores of the girls and boys. From this study it would be difficult to rationalize changing the mathematics curriculum to assist girls because practically they are already doing as well as the boys. A difference of 1.5 points out of 100 is not of true concern. So, while the difference is statistically significant, it is really of little **practical significance**. Practical significance refers to how useful a statistically significant finding is in real life.

How can you tell how much weight to give to statistical findings? First it is important to discover if statistical differences appear. The probability that differences we see could happen by chance needs to be very low to start with. Once that has been established then we need to figure out how to determine if these differences are big enough to justify further action. Some tool needs to be used that can look at any study to determine the size of the impact of an intervention. This standardized computation of the amount of a difference between groups is called **effect size**.

Determining effect size looks similar to determining z scores. The idea is to figure out the number of standard deviations between the two group mean scores (instead of between the mean and a given score as we do with z scores) in terms of standard deviations. Start by determining the mean of both groups and then subtracting one from the other to compute a "distance" between the two mean scores. Then we need to figure out this distance in terms of standard deviations. The question is which standard

deviation should be used for the division. (Recall, there are two groups, so two means and two standard deviations.) Generally, averaging the two standard deviations and dividing that number into the mean differences will solve the problem. Statisticians do something a bit more complicated even though it seldom gives a solution much different from the suggestion above. They use something called a "pooled" standard deviation. This is computed by averaging the square of both standard deviations and then finding the square root (see Table 8-1).

[Insert Table 8-1 here]

Once you have the effect size of a group mean difference, you can report practical significance in your study. In general if you have an effect size around 0.2 (that is, 0.2 standard deviations difference) the effect size is considered small. Effect sizes around 0.5 are labeled medium and around 0.8 or larger are considered large. This is another of those subjective judgments we make as researchers. When reporting effect sizes, try to provide a rationale for why you believe in your case they represent small, medium or large effects in addition to the computed number.

Here is an example. In a section of a biology class (section A) you teach a unit on mammals and administer a unit test at the end. In another section (section B) of the class you teach the same unit but this time you allow the students to examine some web sites that reinforce what you have been teaching. That group gets the same unit test at the end. Group A has a mean of 45.2 on the test and a standard deviation of 14.3. Group B has a mean of 52.7 and a standard deviation of 12.4. You run a *t*-test and the group mean

differences turn out to be significant. Group B scored, on average 7.2 points higher on the test. The average of the standard deviations is 13.35. The mean score difference divided by the average standard deviation equals 0.56. Since this is a distance between the two mean scores it makes no difference if the computation comes out positive or negative. Just use the absolute value of the result. You can report a moderate practical significance in this case.

Effect Size = 52.7-45.2/ ((14.3 + 12.4) / 2) = .56

CHI SQUARE

So far we have been talking about looking for statistically significant differences in groups by comparing group means. Sometimes you will have gathered data from groups that are not interval data and consequently no mean will be available. When data are in the form of nominal (using categories, e.g., gender, ethnicity) or ordinal variables (things can be ordered, but not uniformly measured, such as scales like once a month, once a week, once a day), it is still possible to use chance as way to determine if differences are likely to have appeared randomly or if something else important is going on. The statistical procedure for doing this is called **Chi Square**. As you will see, the richness of interpretation from this procedure is considerably less than comparisons of group means. Nonetheless, Chi Square can be a useful tool for group analysis. Unfortunately, Chi Square can also take more work to calculate than comparisons of group means.

For Chi Square, the number of responses in each response category is listed in a table when two variables are compared (one on the x axis and one on the y axis of the table). The resulting table is called a **contingency table** or a **cross tabs**.

Imagine that a reading teacher wants to know if students have preferences for the kind of stories they read. She asks each student if they would rather read stories about people, animals or travel. But, she is worried that boys might have a different preference than girls so she keeps track of whether each response comes from a girl or a boy. The question is, can she just ask for reading preference or will gender make a difference in the responses. She wants to know if the variables of reading preference and gender are *independent* or if she must know one to understand the other. The Chi Square test in this case is called a **test of independence**.

With a test of independence the observed data are from two variables at the same time. To calculate whether boys and girls have statistically different reading preferences, start by placing the observed frequencies into a contingency table. In this case, we would have gender as the rows and types of stories as the columns. Refer to Figure 8-5, which compares reading topic preference for boys and girls.

[Insert Figure 8-5 here]

Generally, the logic of Chi Square is to look at the distribution of the responses in the cells of the contingency table and to use a calculation to compare those responses to what you would have expected to find in the cells. The calculation produces a probability that the differences between the observed responses and the expected responses could be

explained by random variation. If the probability is really low ($p < .05$) then random variation is not a good explanation for the differences and something else is. The two variables are not independent. Onto to the calculations!

Once you have the observed frequencies in the table, most statistical software will be able to calculate what the expected frequencies ought to be in each cell. If you are using Excel, however, you will need to determine the expected frequencies by hand. Calculate the number of responses in each column of the table. Then calculate the total number of responses in each row of the table. Select one cell in the contingency table. Multiply the row total for that cell by the column total for that cell and divide by the total number of responses. The resulting number will be the expected value for that cell. Refer again to Figure 8-5. In the case of the cell that lists the number of times boys picked people stories (upper left cell), multiply 26 (the row total) by 24 (the column total) and divide that by 61 (the total number of responses in the whole table). The expected frequency is 10.230. See the expected frequencies in Figure 8-6 calculated using this method.

[Insert Figure 8-6 about here]

If you have gotten this far, even Excel can do the rest of the Chi Square calculation for you. As noted above, if you are using a more sophisticated statistical program, calculating expected frequencies will not be necessary because the software will do it for you. In either case, you will want to have the expected frequencies when you report your findings. Whatever statistical program you use, it will provide you with a number appropriately called the Chi Square (in this case it turns out to be 7.68). You will also be

given a *p* value (in this case it turns out that it is .021) that is interpreted like any other *p* value. In Excel you have to use a formula called a chitest, but in most programs when you build the contingency table if you want it simply lists the Chi Square and *p* value below the table.

Since the probability is low ($p < .05$) that the differences between the observed and expected frequencies could have occurred by chance, the teacher would conclude that the differences did not occur by chance. Preference for reading topic is dependent on gender. You have to know the gender to help predict reading preference. The two variables are not independent.

Unfortunately, the Chi Square does not tell you much more about what the differences in the preferences for the two groups are. It just tells you that the differences between the observed and expected differences were unlikely to occur by chance. It does not tell you which differences were unlikely to occur by chance. It is similar to the ANOVA in this aspect; however, there is no post-hoc test for the Chi Square. As a researcher you now have to make a case for what the Chi Square has told you. Often this is a subjective process. Go back through the table showing the observed and expected frequencies and look for inordinately large differences. In our example you might suggest boys seem to select books about people less often than was expected, where girls selected people more often than expected. Remember that this is not statistical proof. You only know that it is really unlikely that gender and reading preference are independent, and you are making informed suggestions of where that dependence has appeared.

A common use of the Chi Square is with questionnaire responses. Very often you will see response category sets like: never, seldom, sometimes, and frequently. We have

no trouble seeing these as ordinal data. The categories are clearly ordered but it is a struggle to believe that the intervals between these responses are equal. Ask yourself the question: is the distance between never and seldom the same as the distance between sometimes and frequently? Just saying that out loud makes it clear that that is a question that makes no sense. So, these are ordinal categories and not interval. These data should not be coded with a cardinal number (never = 1 and so on) and analyzed with means, standard deviations, or group mean comparisons. The more appropriate analysis with ordinal data is Chi Square.

As an aside, while there are many statistical software programs available on most college campuses that students have access to, we have found that nearly all of our students have Microsoft Office on their home or school computers and are comfortable using Excel spreadsheets. As mentioned periodically throughout this and Chapter 7, Excel can be used to perform basic statistical analyses. Because Excel is a software program that most of our students will have ready access to both now and in the future, we have included a set of basic instructions to perform common statistical analyses using Excel in Appendix E.

NEXT STEPS

In this chapter we have looked at ways to show statistically that important differences are appearing between groups. When you design a study with the intent of testing the impact of some intervention, usually this is what you would want to find. Statistical analysis is very good at showing if differences in groups are due to random chance or more likely to be because of something we did. Unfortunately statistical analysis is not

very good at revealing if the design of your study may have allowed something else to impact your results beyond the intended intervention. We need to take a careful look at how to make as sure as possible that the impact you think you see from your intervention really is responsible for the results you observed. Before we do that, however, let's look at some specific designs often used in quantitative research.

CHAPTER SELF-CHECK

Having completed this chapter, you should be comfortable discussing the following:

- normal distribution, including standard error and estimating parameters
- tests of significance: t-test, paired t-tests, ANOVA, Chi Square
- null hypothesis
- statistical vs. practical significance
- Type I and Type II errors
- one and two tailed tests
- setting significance levels

CHAPTER REVIEW QUESTIONS

1. Why is it important to see if mean score changes are statistically significant when you can clearly see that scores have improved?

2. Why is it so difficult to *prove* something in quantitative research?

3. How do you convince a reader of your research that it was appropriate to do a one-tailed test?

4. Under what circumstances can you use a paired *t*-test?

5. What are *post hoc* tests?

6. Why is practical significance important?

7. When do you need to use a Chi Square test?

The following references provide additional explanations of many of the concepts in this chapter:

REFERENCES

Creswell, J. (2009). *Research design: Quantitative and qualitative approaches.* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Gall, M.D., Gall, J. P., & Borg, W. R. (2007). *Educational research: An introduction.* (8th ed.). Boston, MA: Allyn & Bacon.

Gay, L. R., & Airasian, P. W. (2009). *Educational research: Competencies for analysis and application*. (9th ed.). Upper Saddle River, NJ: Prentice Hall.

Figure 8-1: Comparison of normal distribution formed with individuals and a distribution formed from group sizes of 20 (sampling distribution of the mean).

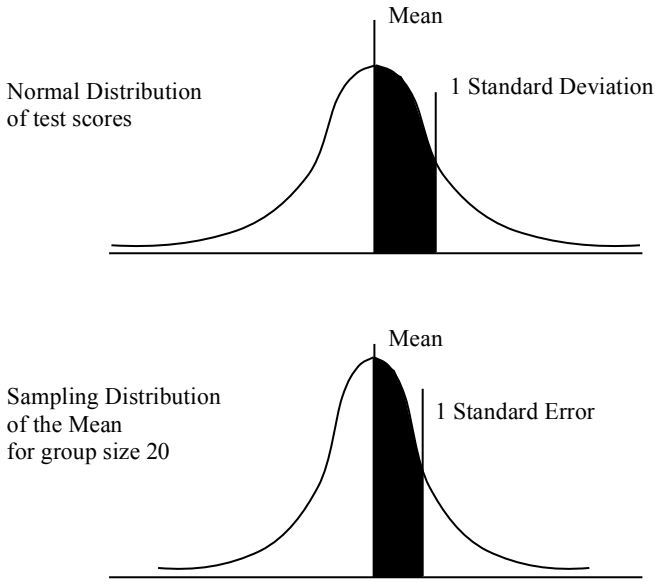


Figure 8-2: Mr. Smith's Curriculum (Estimating Parameters)

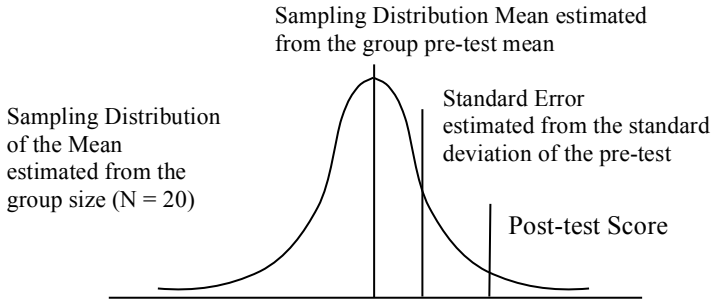
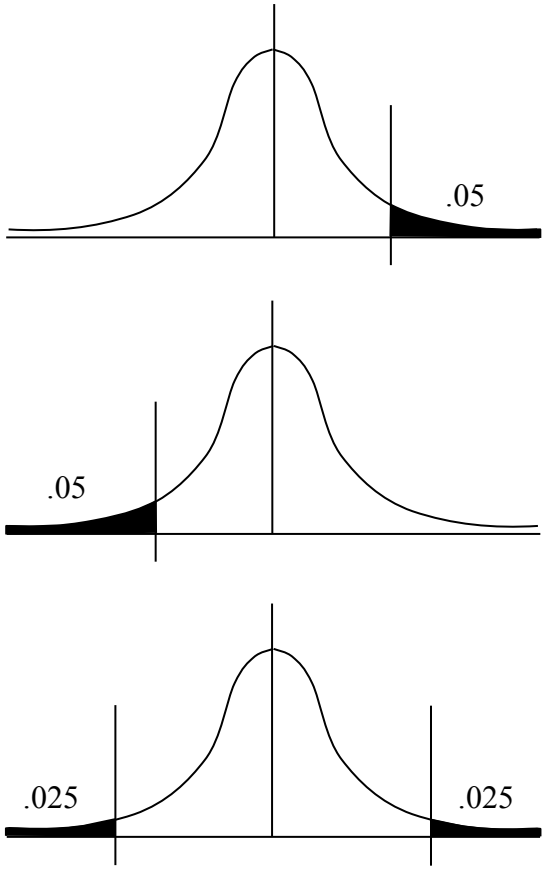


Figure 8-3: Differences between one-tailed and two-tailed tests with p set at .05



One-Tailed Test
Confident mean differences will show positive shift.

One-Tailed Test
Confident mean differences will show negative shift.

Two-Tailed Test
Not sure if mean differences will show positive or negative shift.

Figure 8-4: Example ANOVA Results

Grand Mean	111.60
N	93

Group(group)	**N**	**Group Mean**	**Std Deviation**
1	31	95.35	29.62
2	30	104.43	48.88
3	32	134.06	52.57

ANOVA Table

Source of Variance	**SS**	**DF**	**MS**	**F**
Between Groups	25867.941	2.000	12933.971	6.422
Within Groups	181262.338	90.000	2014.026	
Total	207130.280			

P	**.002**
Eta Squared	.125

Post Hoc tests	**Comparison**	**Mean Difference**	**T-Value**	**P - Unadjusted**
Group_1				
	1 and 2	9.078	.881	.382
	1 and 3	38.708	3.585	**.001**
Group_2				
	2 and 3	29.629	2.294	**.025**

Table 8-1: Computing Effect Size

$$\text{Effect Size} = \text{mean}^1 - \text{mean}^2 \;/\; ((\text{std dev}^1 + \text{std dev}^2)/2)$$

Figure 8-5: Cross Tabs Example

	Reading Preference		
	People	**Animals**	**Travel**
Boys	5	8	13
Girls	19	6	10

Figure 8-6: Observed and Expected Frequencies in a Chi Square

	Reading Preference			
Gender	**People**	**Animals**	**Travel**	**Row Total**
Boys	5	8	13	**26**
Expected	*10.230*	*5.967*	*9.803*	
Girls	19	6	10	**35**
Expected	*13.770*	*8.033*	*13.197*	
Columns Total	**24**	**14**	**23**	**61**