The Logic of Data Analysis Using Statistical Techniques
M. E. Swisher, 2016

This course does not cover how to perform statistical tests on SPSS or any other computer program. There are several courses on campus that focus almost explicitly on how to run various analyses using different statistical programs, SPSS being the most popular in the social sciences at this time. It is not about the many considerations that you have to take into account in the process of statistical analysis. An example would be everything you have to consider to make sure that your data meet the assumptions for a specific test. Most of the "how to run SPSS" type courses will **not** teach you that, but some do. Many (perhaps most) graduate statistical methods classes do teach that. If you don't know those things, you are apt to perform procedures that are simply invalid for your data. Just because SPSS will let you do it does not mean it's a valid use of the test. This course is, to some degree, about how to interpret the results of a statistical test, but to be quite honest the treatment of this topic is very superficial in this class. A good statistical methods class will also teach you this skill because this is inherently tied to the validity of using a specific test.

So what is this section of this course about? It is about the ***logic*** of data analysis from a research design perspective. ***Planning*** your data analysis is part of the design process. Decisions that you make in design will constrain the data analyses you can perform – some severely so. Conversely, the ***planned generalization*** and ***planned contributions to the body of knowledge*** will drive what kinds of data analyses you need to perform.

Have the following materials available for reference as you read this

1. Cheat sheet Comparative Characteristics of Design Groups
2. The "big fun table" as Frey calls it on the pages immediately after the Table of Contents in his book
3. My Statistics Cheat Sheet linked at the course home page and the home page for this week

Descriptive Designs

As Gorard points out and as I have pointed out, descriptive studies especially those using a design that have

- Only one group (sample)
- No intervention or "treatment"
- Data taken at only one point in time
- No theoretical basis
- No formal hypotheses

… are very weak in explanatory power. They are simply descriptions of a few traits of a specific set of trees or ants or people at a single point in time. Here is what that design looks like in the research design notation.

N O

They don't permit the researcher to arrive at warranted conclusions about causality – and that's a huge part of what science does. Science explains and causality (not necessarily direct cause and effect) is central to explanation. If you remember from the Gorard text (Ch. 5), the four components you need to establish a causal claim are:

- A correlation between proposed cause and proposed effect

- The effect appears ***after*** the cause in time
- Changes in the strength of the cause produce changes in the effect
- A plausible explanation of the process by which cause and effect are related (a warrant)

As you can see, the "N O" design permits very few conclusions about causality and therefore, few explanations of how and why something occurs. This design does permit a few analyses. Look at the big fun table under the heading Frequency Analyses (the first page of the table). Now look under "Number of Groups." There are four analyses that are "OK" to use with this common single group design, but as Frey points out, the only one of these tests that will let you examine the relationship between cause and effect (predictor and outcome variables) is the two-way chi-squared test. Explaining inherently requires explaining the relationship between at least two things – so the explanatory power of the single group test is very limited. The limitations are even greater when the variables are not theoretically based and/or when there are no hypotheses or propositions. So much for the single group, single point in time designs. We honestly will not discuss them in this class much, although you will see TONS of them in the literature.

Group Comparisons with Single Observations

The descriptive single group design discussed above *is* a cross-sectional design, but an extremely weak one in explanatory power. The simple addition of a ***meaningful*** comparison group makes the design much stronger in this regard. By meaningful I mean that you have some reason (previous data, better yet theoretical) to believe that two or more existing groups (college graduates, high school graduates, and people who did not complete high school) will differ with regard to the outcome or dependent variable. These are naturally occurring groups. There is no intervention on your part. Sometimes, however, there has been some event in the past that you think has affected their lives. An example is health outcomes for people exposed to the air in Manhattan (NYC) after the September 2001 attack on the World Trade Center. This was not a "treatment," but it certainly is an event that could have repercussions on health later – even years later. Here is a design for examining the effects of this event.

N0 O
N1 O

Look at page 2 of the big fun table. We have two levels of the independent variable. This sounds confusing until you realize that Level 1 is "people not near ground zero" and Level 2 is "people who breathed the air at ground zero." That is, number of levels simply refers to the number of grouping variables. The number of levels will ***always be equal or greater than the number of groups – no matter what the design,*** e.g., there cannot be more comparison groups than there are number of levels of the independent variable. There are two tests we can use to see if there is difference between no treatment and treatment. The Mann-Whitney Test is a non-parametric test and you can use it if the level of measurement for the dependent variable is ordinal. The independent t-test is a parametric test and you must have interval data for the dependent variable.

Let's assume that we think that educational level will have a big effect on income at age 40. Here is a simple multi-group cross-sectional design to test the hypothesis that "educational level is positively associated with income at age 40."

N0 O (did not finish high school)
N1 O (high school graduates)
N2 O (college graduates)

Now we have three levels and three groups. The table in Frey shows that are now several more tests we can use, including the Kruskal-Wallace Test and Analysis of Variance. Kruskal-Wallace is non-parametric so you could have income as <40,000/year; 40,000-80,000/year, and >80,000/year. That is ordinal data. If you want to use analysis of variance, you need the actual dollar amount that each individual earned (interval data) because this is a parametric test.

Logic Rule #1: If there are two or more groups in your design, no matter whether they are formed by you for an experiment or exist naturally, the first logical step in analysis is usually to test for differences among the groups. Inherently, by identifying comparison groups, you are hypothesizing that they differ.

Group Comparisons with Multiple Observations

Designs with an intervention are normally true or quasi-experiments. Both groups of designs require an intervention (see the cheat sheet on characteristics of designs). The difference between the two is that in true experiments individual participants (usually meaning a human being for us, but can be something else) has to do with random assignment. We may not be able to ***randomly*** assign people to treatment versus control because they come to us "already in groups" like classrooms of students. We then use the quasi-experiments. True experiments require random assignment to comparison groups of the ***individual people in the experiment.*** Quasi-experiments randomly assign ***some functional unit*** like schools or classrooms to the comparison groups. Usually we do this because it would be impossible or very disruptive to try to assign individuals to the groups in the research setting. They both use random assignment – the difference is that in quasi-experiments assignment to comparison groups is ***not random at the individual level.*** Whichever is the case, the logic of the experimental designs is that we start with people who are very much alike with regard to traits or characteristics that could affect the outcome of the study, randomly assignment them to different comparison groups, implement the intervention and then measure the dependent or outcome variable score to see if the intervention made a difference. The tricky part here is that almost always we mean a ***change*** in the outcome score. This adds a layer of complexity to data analysis.

In the design below there is a true control (X0) meaning that we did absolutely nothing with this group other than measure the outcome score and there is one treatment (X1). In reality, true controls are rare and we usually mean "the current or traditional or most common" intervention.
In this design we have to measure the change in score (O) and see if the treatment (X1) caused it to change more than the change that happened with no treatment (X0). You may say "Well, why would I anticipate any change if there was not treatment?" The answer is that some things just get better (or worse) on their own. In some cases, this is called the "placebo effect." I have high blood pressure. I volunteer to participate in an experiment testing a new drug. It makes me feel hopeful, I have people helping me – this makes a true difference in my blood pressure. I'm less worried, less stressed, and feel more cared for so it goes down (down being an improvement in this case). On average, this "placebo effect" could conceivably be just as big as the effect of the "real" treatment. In other cases, everybody goes down – gets worse – ***but*** it could be that the group with our treatment "went downhill less." This is sometimes an outcome with social interventions. Whatever the case, we now collect data at two or more points in time. These designs require the use of statistical tests that will account for the change in score. That's because people with a pretty good score at the start of the study innately cannot improve as much as people who had a very bad score at the start of the study. These differences in potential improvement in score require that the effect of the first score (or all previous scores when there are more than two scores) be included in the calculations of both individual scores and mean group scores. The repeated measures tests do this. Note that in statistics and on my cheat sheet about statistics these are called ***dependent samples or paired samples*** test – e.g., paired-samples t-test. These are dependent or paired samples because the mathematics involved use the

score at time one for an individual to calculate the change to take into account "how much" each respondent "could have changed." Page 3 of the big fun table lists these tests.

N1 O X0 O
N2 O X1 O

Here's an example. This experiment deals with the effects of supervised exercise on weight loss. X0 means no supervision. People just do what they want with regard to exercise. X1 is "moderate" exercise – let's say 3 days per week. X3 is "vigorous" exercise meaning 5 days per week. Now we have three levels of the independent variable and three groups. There are several tests we can use. The repeated measures analysis of variance is probably the most commonly used parametric test and the Friedman test is in my experience the most commonly used non-parametric test. Remember, non-parametric tests permit you to have ordinal data for the variable of interest (outcome variable). If you use repeated measures analysis of variance, the outcome score must be interval (0.36 pounds, 2.18 pounds, etc.). For the Friedman test, you could just have categories -- lost less than 1 pound in a month, lost 1-3 pounds in a month, lost more than 3 pounds in a month.

N1 O X0 O
N2 O X1 O
N3 O X2 O

The tests on page 3 of the big fun table are all paired or dependent simple tests. There are those that are suitable for only two comparison groups like the dependent samples t-test or the Wilcoxon signed ranks test. Other are good for more than two groups like the Friedman test or the repeated measures analysis of variance. The two-way repeated measures analysis of variance is very commonly used, particularly for experimental designs.

All ***longitudinal designs*** require the use of paired or dependent samples, whether or not the groups are created through an intervention or not. Often, the comparison groups in longitudinal designs are cohorts, which may be based on age, time elapsed since some initial point, or other factors. The event or conditions that created the comparison groups commonly occurred prior to the start of the research. The example of the effects of breathing the air in the immediate vicinity of the World Trade Center collapsing buildings is an example. There was no way to plan that study prior to the event. However, we can track the long-term effects of breathing the air on people, comparing people in the immediate vicinity (on site), to those living and working in lower Manhattan at the time, to a similar group of urban dwellers not living or working in Manhattan. That design would look like this and would require complex statistical tests employing multiple adjustments to accommodate the repeated effects of multiple "predecessor" scores on each current score. This gets very complicated very quickly. Get statistical help.

N1 O O O O O O
N2 O O O O O O
N3 O O O O O O

Logic Rule #2: When you collect the same information from the same people on two or more occasions, no matter how many groups there are or whether the design was an experiment or not, you must use a paired or dependent samples statistical test in order to take into account the influence of previous scores on subsequent scores.

Taking Independent Variables into Account

Rarely are we interested only in the effect of the grouping variable or treatment on the outcome or dependent variable. Even in experiments, we normally have several other non-experimental variables. Non-experimental simply means that we did not manipulate them as we did the treatment. Usually these are traits or characteristics of an individual that we think could affect the outcome (dependent) variable. For example in the exercise and weight loss example, we would have people who started the experiment at different levels of overweight. We would for sure want to know if "initial overweight" affected the efficacy of the exercise program. We would have people of different ages. We would also probably have people with extensive previous experience with exercise and some with little experience or people who had have been overweight for many years and people who recently became overweight. We can control for these non-experimental factors in two ways. One – preferred in true experiments – is to ***screen*** for them. Limit the age range, limit participation to people who are "more than 10% overweight for age and height," take only people who have been trying to lose weight for at least three years, etc. But even then, there will still be variance in all these factors. In many cases, the conditions under which we conduct a study will prevent us from screening as much as we would like. Cross-sectional and case study designs prohibit any experimental control, as do many longitudinal studies. In these cases, we use correlational analyses to assess the effect of non-experimental variables on the dependent or outcome variable. In the case of true and quasi-experiments, we refer to this as "statistical control of non-experimental sources of variance." Page 4 of the big fun table provides a long list of correlational analyses.

Be cautious and precise in distinguishing between regression analyses and other types of correlational analyses. The two are ***not*** the same thing.

Correlation refers to the strength of a relationship between variables. There is no assumption that one *predicts* the other and there is no assumption that one is the independent (predictor) and the other the dependent (outcome) variable. At the most basic level, think of it this way – a correlation does not produce a line through two points that would let you know how much a unit of change in one variable (the independent) produces in the other variable (the dependent). It just tells you the strength of the relationship between the two "in general" for a set of data. You can use nominal, ordinal or interval data. Probably the most common is the Spearman Rank Correlation which uses ordinal data.

In regression, there are two or more (usually more) independent (predictor) variables and a dependent (outcome) variable. This is important because you must have a logical reason, typically a theoretically justified reason, for thinking that A, B, C, D and E are predictors of O (the outcome variable). The regression model is predictive. In a simple case of A and O, the value of A will vary and you will know how much the value of O changes for each increment in the value of A. Further, multiple regression analyses let you know the relative contribution of each independent variable to the outcome variable. This means, for example, that you could identify which of several factors has the greatest effect on an outcome – which can help you design more effective interventions. Regression, with the exception of logistic regression, requires interval data for both the independent and dependent variables.

A very common use of correlational analyses is to follow the test for differences between groups with a correlational analysis, usually regression. For example, regression analysis would let you know whether the same independent variables are important for two groups that do differ with regard to the

all studies using statistical analyses use some form of correlational analysis. Regression is preferred because it provides much more information that does simple correlation.

Logic Rule #3: Follow comparisons for differences with group with additional analyses, probably correlational in nature, that will help you understand how experimental variables (planned, of interest, part of your research) and non-experimental variables (things you cannot control and are not of interest to you, but could affect your study) affected the results and identify which of them had the greatest effects.

Other Considerations

The level of measurement is critical. Nominal measurement (except for grouping variables) permits almost no statistical analysis and ordinal measurement greatly limits the options open to you. ***Collect data at the highest level of measurement possible if you plan to conduct any statistical analyses.***

All of the tests listed in the "big fun table" have assumptions that you must meet to use them. For example, non-parametric tests like the t-test require that the data for both comparison groups be normally distributed and that the two groups have equal variance. Frey lists these under Other Considerations. They are stated on the Swisher cheat sheet for statistics as well. ***You cannot use a statistical test if your data do not meet the assumptions for the test. Therefore, before you ever run something like a t-test, you have to run tests for normality and tests for homogeneity of variance.*** These are all described in the Swisher cheat sheet.

You may also decide to run some unplanned post-hoc tests after you collect your data. For example, you may see something that leads you to suspect that there is a difference between male and female respondents. You can run statistical tests to see if these differences are significant, ***but you cannot generalize the results beyond your sample because you did not draw independent samples of males and females.***

In general, one-tailed hypotheses are better than two-tailed hypotheses for three reasons. (1) A hypothesis that simply says "there will be a difference" is weak. If you've looked at previous research results and have a sound theoretical basis for your study, you should be able to make a one-tailed hypothesis that says "Scores will be ***higher*** for Group A than Group B" or something like that. (2) Your ability to detect a difference is higher in a one-tailed than a two-tailed test. The same sample that would give you a confidence interval of 0.10 for a two-tailed test gives you a confidence interval of 0.05 for a one-tailed test. (3) Logically, based on (2), you can often use smaller samples to achieve a given confidence interval if you formulate one-tailed hypotheses..