**Analysis of Quantitative Data**

I. Analyzing Data
   a. Commonly, researchers report their research in the form of journal articles, dissertations, and so forth. In doing so, the researcher had to complete a series of other less dramatic steps to produce the polished report. The following steps outlined set the stage for researchers to begin analysis of previously collected data.

II. Dealing With Data
   a. Coding Data
      i. Coding data means systematically reorganizing raw data into a format that can be entered into a computer for further analysis. Researchers begin thinking about a coding procedure and codebook before they collect data (also known as precoding). To complete coding, researchers use a:
         a. Codebook
            i. A codebook is a document describing the coding procedure and the location of data for variables in a format that computers can use.
         b. Coding Procedure
            i. Is a set of rules stating that certain numbers are assigned to variable attributes.
   b. Entering Data
      i. In most circumstances, researchers will have to input the data into the program. When inputting data into the Statistical Program for the Social Sciences (SPSS), each respondent or case is on the row, while each column is representative of one variable. In other words, if there were 23 rows and five columns filled with data, there would be a total of 23 cases or respondents by five variables. Data may also be entered using optical scan sheets.
   c. Cleaning Data
      i. Accuracy is extremely important when coding or entering data. Errors made when coding or entering data into a computer threaten the validity of measures and cause misleading results. To protect against too many errors, researchers typically take measures to re-verify their data was inputted accurately.
      ii. Techniques Used to Clean Data
         1. Take a small sample of the inputted data and re-verify that the data was entered accurately.
         2. Wild Code Checking
            a. Involves checking to make sure that all the numbers inputted for a particular variable are within the acceptable range of scores. For example, if a value of three was entered for sex of respondent, a researcher would immediately know that there was an error made. The maximum value should only be two: female=1, male=2. In other words, since there are only two possible attributes for the variable sex, any value out side of 1,2, and the "missing value" must be data inputting errors. A simple way to perform wild-code-checking is to run a frequency distribution on every variable.

III. Univariate Analysis – The Analysis of a Single Variable
   a. Frequency Distributions
      i. The easiest way to describe the numerical data of one variable is with a frequency distribution. It can be used with nominal, ordinal, interval, or ratio level data and takes many forms.
   b. Measures of Central Tendency

i. Researchers often want to summarize the information about one variable into a single number. They use three measures of central tendency: mode, median, and mean, which are often called averages.
ii. Mode
    1. The modal value is simply the most common or frequently occurring value.
    2. Appropriate to Be Used With:
        a. Nominal
        b. Ordinal
        c. Interval
        d. Ratio
iii. Median
    1. The median is the point at which 50% of the distribution falls above it and below it. The median is also referred to as the score at the 50th percentile.
    2. Appropriate to Be Used With:
        a. Ordinal
        b. Interval
        c. Ratio
iv. Mean
    1. The mean is also called the arithmetic average, is the most widely used measure of central tendency. Compute the mean by adding up all the scores in a distribution, then divide by the number of scores (N).
    2. Appropriate to Be Used With:
        a. Interval
        b. Ratio

c. Distributions
    i. Normal
        1. If the frequency distribution forms a "normal" or bell-shaped curve, the three measures of central tendency are identical.
    ii. Skewed Distributions
        1. If the distribution is a skewed distribution (i.e., more cases are in the upper or lower scores), then the three measures of central tendency will be different.
        2. Types of Skewed Distributions
            a. Positively
                i. Occurs when a large proportion of the distribution is made up of scores that are low.
            b. Negatively
                i. Occurs when a large proportion of the distribution is made up of scores that are high.

d. Measures of Variation
    i. Measures of variation indicate the amount of dispersion, spread, or variability around the center. Thus, a measure of zero variation would indicate that every score was identical to the mean. Researchers measure variation in three ways: range, percentile, and standard deviation.
    ii. Range
        1. The range is the simplest. To compute the range, a researcher subtracts the largest and smallest scores in a distribution.
        2. Appropriate to Be Used With:
            a. Ordinal
            b. Interval
            c. Ratio
    iii. Percentiles
        1. Percentiles tell the score at a specific place (reported as a percentile) within a distribution. One percentile you already learned is the median, the 50th percentile.

2. Appropriate to Be Used With:
    a. Ordinal
    b. Interval
    c. Ratio

iv. Standard deviation
1. Is the most difficult to compute; it is also the most comprehensive and widely used. It is based on the mean and gives an "average distance" between all scores and the mean. It is used for comparison purposes. The standard deviation and the mean are used to create z-scores. Z-scores let a researcher compare two or more distributions or groups. The z score, also called a standardized score, expresses points or scores on a frequency distribution in terms of a number of standard deviations from the mean. Scores are in terms of their relative position within a distribution, not as absolute values.
2. Appropriate to Be Used With:
    a. Interval
    b. Ratio

IV. Bivariate Analysis - Analyzing Two Variables

a. A Bivariate Relationship Explained
    i. Bivariate statistical analysis shows a statistical analysis between two variables. Statistical relationships are based on two ideas:
        1. Covariation
            a. Means that things go together or are associated. To covary means to vary together. When stating a hypothesis with covariation in mind, the researcher is generating a research hypothesis.
        2. Independence
            a. Independence is the opposite of covariation. It means there is no relationship between variables. If two variables are independent, values of one variable have no effect on the other variable. When stating a hypothesis with independence in mind, the researcher is generating a null hypothesis.

b. Techniques Used to Help Researchers Decide Whether a Relationship Exists Between Two Variables
    i. Scattergram
        1. A scattergram is a graph on which a researcher plots each case or observation, where each axis represents the value of one variable.
        2. Appropriate to Be Used With:
            a. Ordinal (rarely used for ordinal level data)
            b. Interval
            c. Ratio
        3. What can you learn from a scattergram?
            a. A researcher can see three aspects of a bivariate relationship in a scattergram: form, direction, and precision.
                i. Form
                    1. Relationships can take three forms:
                        a. Independence
                            i. Independence or no relationship is easiest to see. It looks like a random scatter with no pattern, or a straight line that is exactly parallel to the horizontal or vertical axis.
                        b. Linear
                            i. A linear relationship means that a straight line can be

visualized in the middle of a maze of cases running from one corner to another.

c. Curvilinear.

i. A curvilinear relationship means that the center of a maze of cases would form a U-curve that is either right side up or upside down, or an S-curve.

ii. Direction

1. Linear relationships can have a positive or negative direction.

a. Positive

i. The plot of a positive relationship looks like a diagonal line from the lower left corner to the upper right.

b. Negative

i. A negative relationship looks like a line from the upper left to the lower right.

iii. Precision

1. Bivariate relationships differ in their degree of precision. Precision is the amount of spread in the points on the graph.

a. High

i. A high level of precision occurs when the points hug the line that summarizes the relationship.

b. Low

i. A low level occurs when the points are widely spread around the line.

ii. Bivariate Tables

1. The bivariate contingency table is widely used. It presents the same information as a scattergram in a more advanced form. The table is based on cross-tabulations; that is, the cases are organized in the table on the basis of two variables at the same time.
2. Appropriate to Be Used With:
   a. Nominal
   b. Ordinal
   c. Interval
   d. Ratio
3. Reading a Percentage Table
   a. The To Do List:
      i. Look at the title.
      ii. Read the variable labels.
      iii. Examine any background information.
      iv. Determine the direction in which percentages have been computed (i.e., rows or columns).
         1. Researchers read percentage tables to make comparisons. Comparisons are made in the

opposite direction in which the percentages are computed. A rule of thumb is to compare across rows if the table is percentaged down (i.e., by columns) and to compare up and down in columns if the table is percentaged across (i.e., by rows). If there is no relationship in a table, the cell percentages look approximately equal across all rows or columns.

iii. Measures of Association
    1. A measure of association is a single number that expresses the strength, and often the direction, of a relationship. It condenses information about a bivariate relationship in to a single number. There are many measures of association. The correct one depends on the level of measurement. Many measures of association used by social researchers are called PRE's.
        a. Proportionate reduction in error (PRE)
            i. Most of the elementary measures discussed here follow PRE logic. The logic asks: How much does knowledge of one variable reduce the errors that are made when guessing the values of the other variable? Independence means that knowledge of one variable does not reduce the chance of errors on the other variable. Measures of association equal zero if the variables are independent. Most measures of association range from –1 to +1(The only exception in this discussion is Lambda. Lambda only ranges from 0 to +1.), with negative numbers indicating a negative relationship and positive numbers a positive relationship. The strength of the relationship is indicated by the value of the score. The closer to zero, the weaker the strength, the closer to 1.00 the greater the strength (A value of –1 is exactly the same as +1 in regards to the strength of the relationship. The difference is –1 is representative of a negative association while +1 is representative of a positive relationship.).
                1. Approximate strengths of relationships based on numerical values (Remember, the + or – is not used to interpret the strength of a relationship.).
                    a. .00 = no relationship (independence)
                    b. .20 = weak
                    c. .40 = moderate
                    d. .60 = strong
                    e. .80 = very strong
                    f. 1.00 =perfect
        b. Types of Measures of Association
            i. Lambda
                1. It is based on a reduction in errors based on the mode and ranges between 0 (independence) and 1.0 (perfect prediction or the strongest possible relationship). There exists no negative value for Lambda. This is a result of the variable only including data using categorical differences.
                    a. Appropriate to Be Used With:
                        i. Nominal
            ii. Gamma

1. It is based on comparing pairs of variable categories and seeing whether a case has the same rank on each. Gamma ranges from –1.0 to +1.0.
    a. Appropriate to Be Used With:
        i. Ordinal

iii. Kendall's Tau b
1. It is based on a different approach than gamma and takes care of a few problems that can occur with gamma. The discussion is beyond the scope of this outline.
    a. Appropriate to Be Used With:
        i. Ordinal

V. Analyzing More Than Two Variables
a. Statistical Control
i. Showing an association or relationship between two variables is not sufficient to say that an independent variable causes a dependent variable. In addition to temporal order and association, a researcher must eliminate alternative explanations- explanations that can make the hypothesized relationship spurious. In order to eliminate alternative explanations, a researcher must introduce control variables. A researcher controls for a third variable by seeing whether the bivariate relationship persists within categories of the control variable. If the bivariate relationship weakens or disappears after the control variable is considered, it indicates that the initial bivariate relationship is spurious and suggests that the third variable is the true cause of the change in the dependent variable. Statistical control is a key idea in advanced statistical techniques. A measure of association like gamma only suggests a relationship. Until a researcher talks about the net effect of an independent variable- the effect of the independent variable "net of," or in spite of, the control variable. There are two ways to introduce control variables: trivariate percentage tables and multiple regression (multiple regression will not be discussed!).

b. The Elaboration Model of Percentage Tables
i. Constructing Trivariate Tables - The Analysis of Three Variables
1. A trivariate table has a bivariate table (called the original table) and one additional table (called a partial table) for each attribute of the control variable. In other words, if you have a control variable with three attributes, you will have a total of four tables- one original (contains the original bivariate table) and three partial tables (one for each attribute of the control variable). A major limitation of trivariate tables is they are difficult to interpret if the control variable has more than four attributes.

ii. Elaboration Paradigm
1. The elaboration paradigm is a system for reading percentaged trivariate tables. It describes the pattern that emerges when a control variable is introduced. Five terms describe how the partial tables compare to the initial bivariate table, or how the original bivariate relationship changes after the control variable is considered.
2. Patterns of Explanation
    a. Replication
        i. It is when the partials replicate or reproduce the same relationship that existed in the bivariate table before considering the control variable.
            1. Pattern Seen When Comparing Partials to the Original Bivariate Table
                a. Same relationship in both partials as in bivariate table.
                    i. Effect of the Control Variable

ii. Had no effect.

b. Specification
    i. It occurs when one partial replicates the initial bivariate relationship but other partials do not.
        1. Pattern Seen When Comparing Partials to the Original Bivariate Table.
            a. Bivariate relationship is only seen in one of the partial tables.
                i. Effect of the Control Variable
                ii. Here, a researcher can specify the attribute of the control variable in which the original relationship persists.

c. Interpretation
    i. Describes the situation in which the control variable intervenes between the original independent variable and dependent variables.
        1. Pattern Seen When Comparing Partials to the Original Bivariate Table
            a. Bivariate relationship weakens greatly or disappears in the partial tables (control variable is intervening).
                i. Effect of the Control Variable.
                ii. The control variable helps interpret the meaning of the complete relationship.

d. Explanation
    i. Looks the same as interpretation. The difference is the temporal order of the control variable. In this pattern, a control variable (antecedent) comes before the independent variable in the initial bivariate relationship.
        1. Pattern Seen When Comparing Partials to the Original Bivariate Table
            a. Bivariate relationship weakens greatly or disappears in the partial tables (control variable is antecedent)
                i. Effect of the Control Variable
                ii. The control variable helps interpret the meaning of the complete relationship.

e. Suppressor Variable
    i. The suppressor variable pattern occurs when the bivariate tables suggest independence but a relationship appears in one or both of the partials.
        1. Pattern Seen When Comparing Partials to the Original Bivariate Table
            a. Relationship appears in partial tables only.
                i. Effect of the Control Variable
                ii. No bivariate relationship exists. When the control

variable is introduced, the true relationship becomes evident.

VI. Inferential Statistics
- a. The Purpose of Inferential Statistics
  - i. Inferential statistics use probability theory to test hypotheses formally, permit inferences from a sample to a population, and test whether descriptive results are likely to be due to random factors or to a real relationship. Inferential statistics rely on principles from probability sampling, where a researcher uses a random process to select cases from the entire population.
- b. Statistical Significance
  - i. Statistical significance means that results are not likely to be due to chance factors. It indicates the probability of finding a relationship in the sample when there is none in the population.
- c. Tests of Significance
  - i. Types of Tests
    1. Chi-Square
       - a. The chi-square test of independence is a test of whether there is a relationship between subjects' attributes on one variable and their attributes on another. As its name suggests, the test is based on the chi square distribution. If a relationship is found to be significant using chi-square, a researcher rejects the null hypothesis and states that a statistical relationship exists. To determine the direction and strength a researcher must interpret the values of the measure of association used. If a relationship is not found to be significant, a researcher accepts the null hypothesis. To determine whether a relationship is significant, a researcher must first determine the level of significance.
       - b. Levels of Significance
         - i. Researchers usually express statistical significance in terms of levels. The level of statistical significance (usually .05, .01, .001) is a way of talking about the likelihood that the results are due to chance factors- that is, that a relationship appears in the sample when there is none in the population.
       - c. Type I and II errors
         - i. Type I error
           1. A type I error occurs when the researcher says that a relationship exists when in fact none exists. In other words, it means falsely rejecting a null hypothesis.
         - ii. Type II error
           1. Occurs when a researcher says that a relationship does not exist, when in fact it does. It means falsely accepting a null hypothesis.