Week 1:
Introduction,
Concepts, &
Step 1: Study Map

The Steps of Data Analysis

6 Week Webinar Class

Presented By

William Bannon, PhD

StatsWhisperer

New York, New York

INTRODUCTION

The core purpose of the *Steps of Data Analysis* webinar is to present the methods needed to produce a legitimate and effective quantitative study. Specifically, this program presents the *essential facts, methods,* and *guidelines of data analysis*, as well as the stepwise order in which each must be applied toward completing such a project. The webinar class is designed to lay a solid foundation for learning and implementing a data analysis, while attempting to avoid a level of detail so great that it might overwhelm the reader.

The weekly materials are presented in the order that reflects the steps necessary to sequentially conduct a data analysis study. Instruction is included on how to conduct each procedure in the SPSS statistical software, as well as how to effectively interpret the statistical output.

Furthermore, *The 7 Steps of Data Analysis* textbook will be used as the central text for the course. The weekly emailed PDF materials will provide a synopsis of the learning for that week. Each major point presented within the weekly PDF materials will cite the page number within the text where the topic is presented in greater detail. The weekly PDF is meant to serve as an effective *overview* of the learning for that week. Essentially, the PDF is meant to serve as a *crib* sheet for quick reference.

Week one of our six week *Steps of Data Analysis* webinar presents our course **Introduction** and covers various bits of **Foundational Knowledge** that no quantitative researcher should be without. Within *The 7 Steps of Data Analysis* text, these materials are covered in detail in Part One (Introduction), Part Two (A Model for Conducting Statistical Research), and Part Three (The Data Analysis Concepts You Need To Know). However, below you will find an abbreviated version of the written materials within the text, as well as the information presented in the webinar videos.

The Definition & Components of Data Analysis

Data analysis may be defined as applying a series of statistical tests and procedures in a specific stepwise progression in order to examine a dataset. You might also say that data analysis involves taking a set of tools (statistical tests/procedures) that when applied in a certain order reveal the message(s), lesson(s), and answer(s) the data have to tell us.

Data analysis could be considered as consisting of three components. First, data analysis must incorporate an *outline of the steps to follow* toward conducting a data analysis study. Second, a data analysis project requires the proper *materials* (i.e. a dataset) necessary to produce a quantitative study. Lastly, data analysis requires the proper *tools* needed to transform and analyze the data, which are typically statistical tests and procedures. Please see pages 5-8 of the text for greater detail, as well as the presentation of how data analysis is similar to the process of making a cake.

A Model For Conducting Statistical Research

The *Steps of Data Analysis* are meant to serve as a model for statistical analysis that meets three basic criteria, which are being *fundamentally sound*, *conceptually clear and understandable*, and provides *a clear presentation of the ordered steps* that must be followed to successfully achieve the desired outcome. Please see pages 19-32 for an elaboration upon these points, as well as a comparison of how the steps of data analysis are similar to the steps of courting a suitor (i.e., dating analysis).

Concepts

The Data Analysis Concepts You Need To Know

There are several core concepts that you need to know before you approach a data analysis study. These concepts are few in number and often relatively simple, but knowledge of each is absolutely critical.

Parametric and Non-Parametric Statistics

Statistical tests are often categorized as *parametric* and *non-parametric* (see pages 35-36). Parametric statistics are a group of tests that make certain *assumptions* (e.g., a normal distribution of scores) about the data used in analysis and include procedures presented in this program such as the t-test, correlation, ANOVA, and linear (multiple) regression. If assumptions are met, these tests can produce accurate and precise estimates, but if these assumptions are not met, results are likely to be misleading.

Non-Parametric statistics are a group of statistical tests that examine (typically through rank) the relationship between variables without making as many assumptions about the parameters and qualities of the data (such as the Chi-square test included in this program). Non-parametric tests are of use in many instances, such as when the dependent variable is an ordinal variable or a rank, as well as when the assumptions

(e.g., no undue effect of outlier scores) of parametric testing are not met. Within this program, we will mention the non-parametric equivalent of many tests, when such tests are available.

Descriptive and Inferential Statistics

Statistical procedures are often categorized as descriptive or inferential (see pages 36-37). *Inferential Statistics* are the statistical procedures used to make predictions or inferences about a population based upon the observations made through the sample. *Descriptive statistics* (also described as *univariate statistics* in this text) are a group of statistical procedures used to describe the characteristics of a sample or population.

Level of Variable Measurement

There are generally four levels of measurement regarding a study variable, which are *Nominal*, *Ordinal*, *Interval*, and *Ratio*. In quantitative research there variables are often referred to as categorical (*Nominal* or *Ordinal*) or continuous (*Interval* or *Ratio*) variables. Many critical decisions in analysis (e.g., selecting the proper statistical test) are based upon the level of variable measurement.

A **nominal variable** has two or more categories, *but cannot be ranked into an intrinsic order*, such as the variables gender (male/female) and racial identity (Caucasian/White, African American/Black, etc.).

Ordinal variable. An ordinal variable has two or more categories that can be ranked into a meaningful order (e.g., categories such as *poor*, *good*, *very good*, and *excellent*). However, there is *not an equal distance between categorical measurement points*.

Dichotomous variable. The much used term in research *dichotomous variable* refers to a categorical variable (either nominal or ordinal level) that has a two category response, such as Yes or No, or good or bad.

In real world application, *Nominal* and *Ordinal* level variables are referred to as *categorical variables*. Subsequently, we will refer to these variables using this term throughout this program.

Interval variable. An interval variable measures the presence of a construct along a continuum with equal distances between measurement points, *but does not incorporate a meaningful true zero*. For example, an *IQ test* is an interval level measure of *general*

intelligence, but a score of zero is not meaningful as it does not reflect a complete absence of general intelligence.

Ratio variable. A ratio variable measures the presence of a construct along a continuum with equal distances between measurement points and does *incorporate a meaningful true zero*. For example, measuring *weight* in pounds would be considered a ratio level measurement as an increase in value reflects greater weight and a weight of zero is meaningful as it reflects an absence of weight.

In real world application, *Interval* and *Ratio* level variables are referred to as *continuous variables*. Subsequently, we will refer to these variables using this term throughout this program.

Single Versus Composite Item Scores

Another important issue in quantitative research concerns if study variables are measured using a single item or multiple items that may be combined into an overall composite score. For example, the variable Happiness might be measured using a *single* item, such as *Do you feel Happy?* Alternately, the variable *Happiness* might be measured through combining several items (e.g., *Do you feel happy?*, *Do you feel sad?*, etc.) the measure different dimensions of *Happiness*, in which case the multiple combined items would reflect a composite score. Please see pages 39-40 for more detail.

Study Variable Type

In any quantitative study it is essential to identify the study variable type, such as the independent, dependent, and covariate variables.

The **independent variable** (at times referred to as a **predictor variable** in our materials) is the *variable of interest that explains the dependent variable* within the data analysis study.

The **dependent variable** (also referred to as the **outcome variable** in our materials) is the *variable being explained through the data analysis study*.

The **covariate variable** (at times referred to as a **predictor variable** or **control variables** in our materials) is a *variable in a dataset that may not be the focus of the study, but must be included in analysis because the factor has a significant impact on the dependent or independent variables*.

Please see pages 40-46 for greater detail, as well as how the film *It's a Wonderful Life* illustrates the function of the covariate variable.

The Clarity of the Study Hypothesis

A hypothesis is a predictive statement regarding the relationship between study variables. The alternative hypothesis (also called the research hypothesis) and null hypothesis are the two rival hypotheses, one of which will be supported through statistical testing. The **alternative hypothesis** (**research hypothesis**) is the relationship between variables that the researcher believes exists. The **null hypothesis** is the default position that states that there is no relationship between the variables under examination.

Also conceptually relevant to the hypothesis is the concept of **type I error** and **type II error**. **Type I error** involves a rejection of a true null hypothesis. Specifically, type I error leads one to conclude that a significant relationship between variables exists, when in reality it does not. A **type II error** is the failure to reject a false null hypothesis.

The hypothesis is also described as being **directional** or **non-directional**. A directional hypothesis posits how variables will be related. A non-directional hypothesis posits a relationship exists, but does not specify how the variables will be related. When formulating and presenting a hypothesis, it is crucial to be sure your hypothesis is ***succinct*** and possesses considerable ***clarity***. Please see pages 46-50 for additional information.

Three Dimensions of a Relationship Between Variables

The goal of most data analysis studies is to examine the relationship between study variables. However, before such a relationship can be examined, a researcher must be aware of the three dimensions of any variable relationship that must be considered, which are *significance, directionality*, and *magnitude*.

In testing the relationship between variables, *significance*, or more specifically, statistical significance is expressed through the level of probability (a.k.a., *p*-value), which reflects how likely the relationship presented is a result of chance. Although the level of probability may vary, the typical scientific standard is that probability must be less than 5 times in 100 (i.e., $p<.05$) to achieve statistical significance.

Directionality indicates if a relationship between variables is negative or positive relationship between variables. A **positive relationship** between two variables in

statistics could mean, as the first variable increases in score, the second variable increases in score. A **negative relationship** between two variables in statistics could mean as the first variable increases in score, the second variable decreases in score. Typically, in inferential statistical testing, a negative relationship is indicated by the presence of a minus (-) sign in front of the statistical coefficient. A positive relationship is indicated if there is not a minus sign (no -) in front of the statistical coefficient.

The *magnitude* reflects how impactful one variable is upon another, thereby estimating the magnitude of a relationship between variables. Magnitude is indicated in several ways, including the calculation of an effect size. The effect size is categorized as being small, medium, or large. Thus, this value indicates if one variable has a small, medium, or large impact upon another.

Please see pages 50-58 for a more detailed presentation, as well as an illustration of how these same three dimensions of a variable relationship are present in a human relationship.

The Russian Doll Effect in Analysis

As you perform each necessary step and procedure within a data analysis study, you will notice that each task incorporates many subtasks, which also incorporate many smaller tasks. I like to call this tendency the "Russian Doll" effect of data analysis. Specifically, just as each (or at least most) of the dolls have smaller dolls within, each statistical procedure and task often incorporates a series of smaller tasks. Please see pages 58-59 for greater detail.

Haste or Paste: Save the Syntax

When using the SPSS software, each statistical procedure offers the choice of running the test immediately through clicking the **OK** button or clicking the button marked **Paste**, in which case the syntax may be pasted into a syntax file and saved. For many reasons, we recommend taking the extra step and saving the syntax (i.e., Paste). The **OK** button is what we term as haste, as it is a faster method, but leaves no record of the various tests employed in the data analysis. Please see page 59 for more instruction on saving your syntax.

Step 1: Study Map

This week also includes mention of the first step of data analysis, **Step 1: Study Map**, which is the stage of analysis where the study variables and hypothesized variable relationship(s) are listed in both text and diagram form. The study map for sample study 1, in both tables and text, can be viewed on page 62 of TSSODA textbook. Likewise, the study map for sample study 2, in both tables and text, can be viewed on page 243 of TSSODA textbook.

In Summary

Thus far we have reviewed a brief introduction, as well as the needed pieces of foundational knowledge that are essential for data analysis, as well as Step 1: Study Map. As we progress into our data analysis study, it will become apparent how fundamental each of these *must know* facts are to performing quantitative analysis properly. However, prior to approaching any statistical procedures, week two of this program will address the critical process of entering (and then cleaning) data from hard copies of a study survey into the SPSS database.