

1

Introduction

Basis of How All Quantitative Statistical Based Research

Any research study should have a solid design, properly collected data, and draw its conclusions on effectively analyzed data. All of which are nontrivial problems.

This is a book about performing quantitative data analysis. Unlike most research methods texts, which focus on creating a good design, the focus is on analyzing the data. It is not on how to design the study or collect the data; there are many good sources that cover those aspects of research. Of course, poor designs or data collections lead to poor data that means the results of the analysis are useless. Instead, this book focuses on how to analyze the data.

The stereotypical linear view of a research study is shown in Figure 1.1a. Figure 1.1b expands on what is contained within the “analyze data” element. This book only works within that expansion; it focuses on how to analyze data from a study, rather than either how to perform the study or how to perform individual statistical tests.

The last two boxes of the expansion in Figure 1.1 “Make sense of the results” and “Determine the implications” are where performing a high-quality data analysis differs from someone simply crunching numbers.

A quantitative study is run to collect data and draw a numerical-based conclusion about that data. A conclusion that must reflect both the numerical analysis and the study context. Thus, data must be analyzed to *help* draw a study’s conclusions. Unfortunately, even great data collected using a great design will be worthless unless the analysis was performed properly. The keyword in the sentence is *help* versus *give* the study’s conclusions. The results of statistical tests are not the final conclusion for research data analysis. The researcher must study the test results, apply them to the situational context, and then draw conclusions that make sense. To support that process, this book

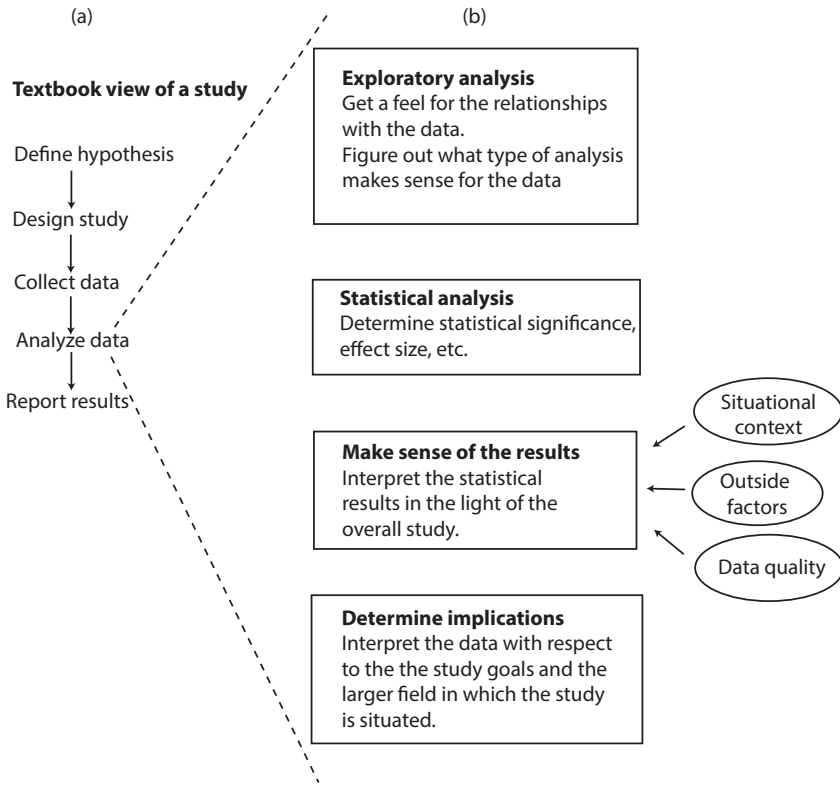


Figure 1.1 View of data analysis as situation within the overall study.

works to place the tests within the context of a problem and provide the background to connect a specific type of data with the appropriate test.

The outcome of any statistical analysis needs to be evaluated in terms of the research context and any conclusions drawn based on that context.

Consider this example of how this book approaches data analysis.

You are interested in which books are being checked out of a library. So, you gather data using many titles that fit within study-defined categories. For example, topical nonfiction or categories for fiction of a particular genre (historical, romance, etc).

At the end of the study's data collection, the analysis looks at the following:

- Graphs of checkouts by month of the various categories. Do the types of categories vary by day/week through the month? How do the numbers compare? Do the trends of checkouts for each category look the same or different?
- Run statistics on the daily/month checkouts of the book categories versus demographics of the people who checked them out (age, gender, frequency of

library use, etc.). Does age or gender matter for who checks out a romance versus a thriller. From this we can find whether there is a statistically significant difference (e.g., that older readers read more romance than younger readers).

Data Analysis, Not Statistical Analysis

Too many people believe if they can figure out how to run statistical software, then they know how to perform a quantitative data analysis. No! Statistics is only a single tool among many that are required for a data analysis. Likewise, the software is only a tool that provides an easy way to perform a statistical test. Knowing how to perform a *t*-test or an ANOVA is similar to knowing how to use styles and page layout in Word. Just because you know how to use styles does not make you a writer. It will not make you a good layout person if you do not know when and why to apply those styles. Neither the software nor the specific tests themselves are sufficient; necessary, yes, but sufficient, no! Run the wrong test, and the results are wrong. Fail to think through what the statistical test means to the situation and the overall study fails to have relevance.

It is important to understand that statistics is not data analysis. Learning how to use a software package to perform a *t*-test is relatively easy and quick. But good data analysis requires knowing when and why to perform a *t*-test; a much more different, and complex task. Especially for researchers in the social sciences, the goal is not to be a statistical expert, but to know how to analyze data. The goal is to be able to use statistical tests as part of the input required to interpret the study's data and draw valid conclusions from it. There is a wide range of statistical tests relevant to data analysis; some that every researcher should be able to perform and some that require the advice/help of a statistical expert. Good quantitative data analysis does not require a comprehensive knowledge of statistics, but, rather, knowing enough to know when it is time to ask for help and what questions to ask. Many times throughout the book, the comment to consult a statistician appears.

Figure 1.1 shows data analysis as one of five parts of a study; a part that deserves and often requires 20% of the full study's time. I recently had to review a set of undergraduate honors research project proposals; they consistently had several weeks scheduled for data collection, a couple of weeks for data clean up, and data analysis was done on Tuesday's. This type of time allocation is not uncommon for young researchers, probably based on a view that the analysis is just running a few *t*-tests and/or ANOVAs on the data and copying the test output into the study report. Unfortunately, with that sort of analysis, the researchers will never reach more than a superficial level of understanding of the data or be able to draw more than superficial conclusions from it.

The purpose of a quantitative research study is to gain an understanding of the research situation. Thus, the data analysis is the study; the study results come directly out of the analysis. It is not the collection and not the reporting; without the data analysis there is no reason to collect data and there is nothing of value to report.

Use dedicated statistical software

There are many dedicated statistical software programs (JMP, SPSS, R, Minitab) and many others. When you are doing data analysis, it is important to take the time to learn how to use one of these packages. All of them can perform all the standard statistical tests and the nonstandard tests, while important in their niche case, are not needed for most data analysis.

The one statistical source missing from the list is Microsoft Excel. This book uses Excel output on many examples, but it lacks the horsepower to really support data analysis. It is great for data entry of the collected data and for creating the graphs of the exploratory analysis. But, then, move on to a higher powered statistical analysis program.

What Statistics Does and Does Not Tells You

In statistics the word “significance” is often used to mean “statistical significance,” which is the likelihood that the difference between the two groups is just an accident of sampling. A study’s data analysis works to determine if the data points for two different groups are from the same population (a finding of not statistically significance) or if they are from different populations (a finding of statistically significance).

Every population has a mean and standard deviation. However, those values are typically not known by the researcher. Part of the study’s goals is to determine them. If a study randomly selected members from the population in Table 1.1 any of those four groups could be picked.

If you take multiple samples from the same population, there will always be a difference between them. Table 1.1 shows the results of Excel calculating six random numbers that fit a normal distribution with a mean = 10 and standard deviation = 2. The numbers were generated randomly, but they could reflect the data from any number of studies: time to perform a task, interactions during an action, or, generally, anything that can be measured that gives a normal distribution. The important point here is that although they all come from the same population, each sample’s mean and standard distribution is different.

Now put those numbers into a study. We have a study with two groups that are looking at the time to perform a task (faster is good). We have old way

Table 1.1 Random numbers generated with a normal distribution of mean = 10 and SD = 2.

	Trial 1	Trial 2	Trial 3	Trial 4
Data points	8.043	7.726	10.585	7.679
	7.284	7.374	9.743	12.432
	11.584	11.510	9.287	13.695
	9.735	11.842	9.102	8.922
	8.319	9.651	4.238	6.525
	8.326	11.849	11.193	11.959
Mean	8.882	9.992	9.025	10.202
SD	1.544	2.063	2.475	2.891

Because of the nature of random numbers and small sample sizes, each trial has a different mean and standard deviation, although they all come from the same population.

(performed by people that resulted in the times of trial 1: mean = 8.882, SD = 1.544) and a new way (performed by people that resulted in the times of trial 2: mean = 9.992, SD = 2.063). Because Table 1.1 presents data without context, a simple *t*-test (or ANOVA) to determine significance is all we can perform. If we simply looked at the mean and standard deviation numbers, it would seem trial 2 was worst, assuming we wanted fast task times. Yet, they are both random sets of numbers from mean = 10 and SD = 2. A proper statistical analysis would return a result of not statistically significance differences.

Within the context of a study, the data analysis requires moving beyond calculating a statistical value (whether a *p*-value or some of the results available with more complex statistical methods) and interpreting that statistical result with respect to the overall context. The research also has to determine if the difference is enough to have practical significance. From a practical viewpoint, there would be no reason to spend money to use the new way, since it is not any faster than the old way. If they had been significantly different, then the conclusions have to consider other factors (such as time and expense of the change) to determine if it is worthwhile. Quantitative data analysis is more than just finding statistical significance; it is connecting the results of the statistical analysis with the study's context and drawing practical conclusions.

Statistical significance is usually calculated as a "*p*-value," the probability that a difference of at least the same size would have arisen by chance, even if there really were no difference between the two populations. By social science convention, if $p < 0.05$ (i.e., below 5%), the difference is taken to be large enough to be "significant"; if not, then it is "not significant." In other words, if $p < 0.05$, then the two sets of data are declared to be from different populations. In the hard sciences, the *p*-value must be smaller.

Data is significant or not significant. There is no maybe. Maybe . . .

Research reports often contain sentences such as “The p -value of 0.073 shows the results are trending toward significance.” Researchers have long running debate about this type of wording. The basis of the argument is that a result is yes/no: it is either significant or it is not significant.

However, the definition of significance as $p = 0.05$ is itself arbitrary and is only a long-standing convention in the social sciences.

More importantly, thinking in terms of a statistical yes/no ignores the study context. The effect size (loosely defined as the practical significance) is more important in the final result than the statistical significance. If a study fails to show statistical significance—in this case, $p = 0.073$ —but has a large effect size, the results are much strong than one with the same p -value (or even $p = 0.04$ that is significant) but a small effect size.

Both p -value and the effect size should be reported in a study’s report.

Some researchers try to skip the statistical analysis and observe and draw conclusions directly from the data. Looking at the data in Table 1.1 without a statistical analysis might conclude that trials 1 and 4 were from different populations, but this is not the case. If this was a study to determine if a current practice should change, lots of money could be spent on a change that will have no real affect. Without a proper analysis, the researcher can fall prey to confirmation bias. A confirmation bias occurs when the data to support a desired claim is specifically looked for and data that refutes the claim is ignored. A simple example is politicians from opposing parties using the same report to claim their agenda is right and the opposing agenda is wrong. Yet, both are cherry picking data from the report to make that claim.

Although statistics can tell you the data are from different populations (there is a statistical significant difference), it does not tell you the *why*.” Yet, the purpose of a research study is to uncover the why, not just the proof of a difference. Thus, a statistical analysis itself is not the answer to a hypothesis. A research study needs to move beyond the statistics and answer questions of why and what really happened to give these results, and then move past those questions to figure out what the answers mean in the study’s context. That is the outcome of a good data analysis.

Data Analysis Focuses on Testing a Hypothesis

Early in a study’s design, a set of hypothesis is created. Then the data needed to test those hypotheses are defined and eventually collected. At the same time the

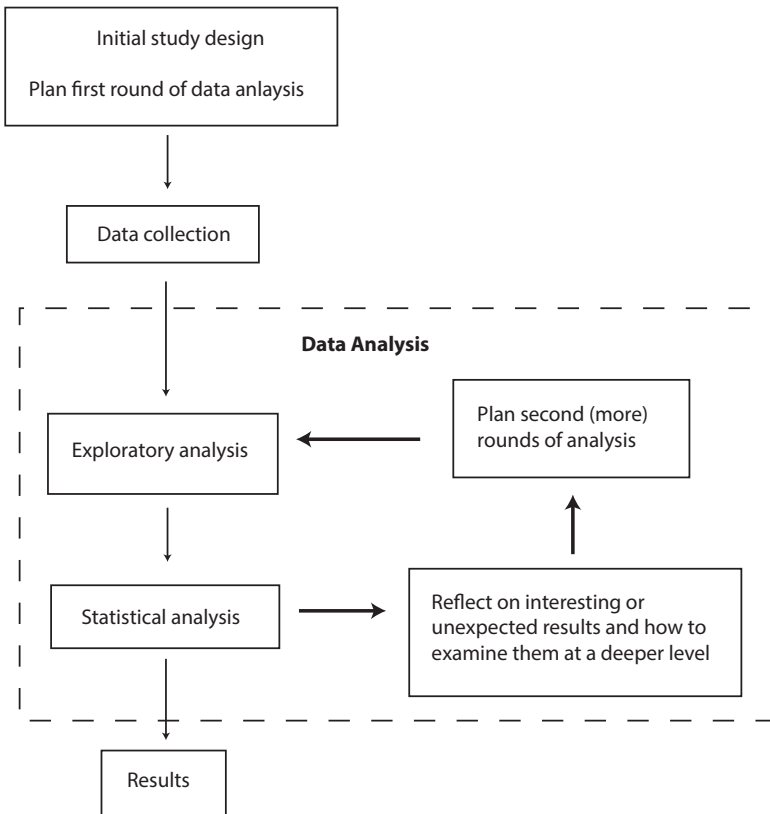


Figure 1.2 Cyclical nature of data analysis.

needed data is defined, the analysis tests to be performed on the data should be defined.

Good data analysis knows from the start how the first round of analysis will be performed; that was defined early in the study design. The second (and other) rounds of analysis that each drill deeper and explore interesting relationships found in the previous rounds. Obviously these cannot be defined until you are engaged in the analysis, but understanding how to pursue them distinguishes a poor from a good researcher (Figure 1.2).

Poor data analysis often just collects data and then runs all possible combinations of data, most of which make no sense to even test against each other. The goal of the data analysis has been shifted from understanding the data within the study to “finding significance . . . any significance.” Unfortunately, with significance defined as $p < 0.05$ that means 5% of the combinations may show significance when it does not exist.

The best, most well-designed study is worthless if the data analysis is inadequate. Focus the analysis on a small number of well-conceived hypotheses rather than blindly performing different statistical tests on all variable pairs and ending up with 5% of your results being significant at the 0.05 level.

Quantitative Versus Qualitative Research

Research methodology textbooks call for qualitative research to get a view of the big picture and then to use quantitative studies to examine the details. Both quantitative and qualitative have their strong and weak points, and a good research agenda requires both. That implies that a researcher needs to understand both. A problem occurs—similar to the “when you have a hammer, everything looks like a nail” problem—when researchers lack training in one, typically quantitative, and try to use a single approach for all research problems.

Quantitative research is a methodical process. Too many people with qualitative research experience—or who lack quantitative research experience—look at a situation, point out all of the interacting factors, and despair (or claim an impossibility) of figuring out the situation.

The social sciences work with large complex systems with numerous variables interacting in subtle ways. The reality is that it requires many studies with each looking at the situation in slightly different ways and with each study exposing new questions and new relationships. Qualitative research can define the variables of interest and provide preliminary insight into which variable interact and help define the hypothesis, but it requires quantitative research to clearly understand how those variables interact.

It is one thing to declare confidently that causal chains exist in the world out there. However, it is quite another thing to find out what they are. Causal processes are not obvious. They hide in situations of complexity, in which effects may have been produced by several different causes acting together. When investigated, they will reluctantly shed one layer of explanation at a time, but only to reveal another deeper level of complexity beneath.

(Marsh and Elliott, 2008, p. 239)

As its strongest point, quantitative data analysis gets at the deep structure of the data. High-quality quantitative data analysis exposes a deep structure and researchers should never be content with the superficial structure that appears at first glance. And certainly should not be content with poor/inadequate data analysis where the analysis process is seen as running a few statistics tests, reporting the *p*-value, and calling that a data analysis.

It is true that no social science study will be able to obtain the clear results with the hard numbers obtained in the physical sciences. The cause and effect relationships of the laws in the physical sciences do not exist. Instead, when people enter into the research equations, there are at best probabilistic relationships. Nothing can be clearly predicted. But to simply refuse to undertake a study because of too many interactions is poor research, as is deciding to only undertake qualitative research. Qualitative research can build up the big picture and show the existence of relationships. As a result, it reveals the areas where we can best apply quantitative approaches. It is with quantitative approaches that we can fully get at the underlying relationships within the data and, in the end, it is those relationships that contain the deep understanding of the overall complexities of the situation (Albers, 2010). Developing that deep understanding is the fundamental goal of a research agenda.

What the Book Covers and What It Does Not Cover

Every quantitative research study collects some type of data that gets reduced to numbers and must be analyzed to help draw the study's conclusions. A great study design is useless unless the data are properly analyzed. What I have found is that most textbooks fall into one of these categories.

- Research method textbooks explain how to create and execute a study, but typically are very light on how to analyze the data. They are excellent on explaining methods of setting up the study and collecting the data, but not on the methods to analyze data after it has been collected.

The data analysis chapters of many research textbooks are little more than a brief explanation of various statistical tests. As a result, reader can come away thinking the important questions to ask are "How do I run a chi-square?" "What is the best procedure, a Kruskal–Wallis test or a standard ANOVA?" and "Let me tell you about my data, and you can tell me what procedure to run" (Rogers, 2010, p. 8). These are the wrong questions to be asking at the beginning of the data analysis. Rather, data analysis needs to be addressed along the lines of "what relationships do I need to understand?" and "what do these results tell me about the research context?"

- Statistics textbooks explain how to perform statistical tests. The tests are explained in an acontextual manner and in rigorous statistical terms. They explain *how* to perform a test, but, from a research standpoint, the equally important question of *when* and *why* to perform a test gets short shrift. Likewise, statistics textbooks do not explain the need to connect the statistical results to the research context.

This book differs from these two categories because it is focused on explaining how to analyze data from a study, rather than how to perform the study or

how to perform individual statistical tests. Early in the Chapter 1 said “data that must then be analyzed to *help* draw the study’s conclusions.” The keyword in the sentence is *help* versus *give* the study’s conclusions. The results of statistical tests are not the final conclusion for research data analysis. The researcher must study the test results, apply them to the situational context, and then draw conclusion that make sense.

The importance of data analysis is clearly summed up within this quote by Phillips.

In moving from data collection to data analysis, we confront a variety of complex mathematical procedures. Thus it is essential not to lose sight of the theoretical aspects of the research process as we become involved in those procedures. The scientific method may be seen as a continuing chain. If our theoretical definition of the problem becomes a series of weak links, the entire chain is weakened. The reverse holds true as well: We do not want our procedures for analyzing data to become weak links in the scientific method. This requires that we learn to make use of the best methods available. Further, since we wish to strengthen the entire chain, we must also learn the limitations of existing analytic tools and be open to ways of solving the problems involved.

(Phillips, 1985, p. 386)

The bulk of the book is examples that walk through a data analysis. They explain the exploratory analysis of getting a good feel for the data and then explain why various statistical tests are performed on the data and how to interpret them.

The examples in this book will not address whether the variables from the data collection make sense or even how the data are collected. These are issues that must be addressed early in the experimental design and clearly affect the overall results, but they are not a data analysis concern. Granted, if the categories make no sense, then the results make no sense. But that is a fault of a poorly designed study, not a fault of poorly performed data analysis.

It is, unfortunately, very easy to have a well-designed study with good data that suffers from poor data analysis or a poorly designed study with bad data that has great data analysis (Table 1.2). In either case, results are useless and may even be misleading. All research studies should strive to be in the lower right quadrant.

Book Structure

This book is striving to explain the when, why, and what for, rather than the button pushing how to.

Table 1.2 Four quadrants of study design and data analysis.

		Data analysis	
		Poor	Good
Study design	Poor	Poorly designed study, poorly performed data analysis. Bad data that is poorly analyzed. The study is a complete failure.	Poorly designed study, well-performed data analysis. Bad data that is properly analyzed. Unfortunately, since it started with bad data, the results are useless.
	Good	Well-designed study, poorly performed data analysis. Good data, but with questionable results. With proper analysis, a good study can be salvaged.	Well-designed study, well-performed data analysis. A study with valid results.

To support the process of how-to-do data analysis, this book works to place the tests within the context of a problem and provide the background to connect a specific type of data with the appropriate test. The work is placed within long examples and the entire process of data analysis is covered in a contextualized manner. It looks at the data analysis from different viewpoints and using different tests to explain how and when to apply different analysis methods.

The ideas and concepts of data analysis form a highly interconnected web. Not understanding a concept makes understanding its application difficult. On the other hand, explaining a concept repeatedly makes a text difficult to read for people who do understand it. To help with the problem, the book uses extensive cross-reference concepts to the page where they are discussed in detail.

The goal of this book is to emphasize that data analysis is not just crunching numbers, but a process of revealing the underlying patterns and trends that allow a researcher to gain an understanding of the data and its connection to the research situation.

References

- Albers, M.J. (2010) Usability and information relationships: considering content relationships when testing complex information. In: M.J. Albers and B. Still (Eds) *Usability of Complex Information Systems: Evaluation of User Interaction*, Boca Raton, FL: CRC Press.

Marsh, C. and Elliott, J. (2008) *Exploring Data: An Introduction to Data Analysis for Social Scientists*, 2nd edn, Cambridge: Polity.

Phillips, B. (1985) *Sociological Research Methods: An Introduction*. Homewood, IL: Dorsey.

Rogers, J.L. (2010) The epistemology of mathematical and statistical modeling: a quiet methodological revolution. *American Psychologist*, **65** (1), 1–12.