Chapter 0:

Preface

I sat there, as I was wont to do, in the Spillane Reading Room, drinking coffee in the early morning hour, trying to find my wakefulness. The air was peaceful, with two freshmen discussing and sharing insights about international politics and world events with each other and with me. It was life as I expected it to be in academia.

Invariably, especially as the term wore on, Simon would rush into the room and explode, "I don't know what this *thing* is telling me!" Then he would throw down five pages of printout from a well-known statistical package and throw up his hands, as if beseeching the gods of statistics to send down an answer to him.

After allowing the situation to calm, and the freshmen to start breathing, I would ask the question "What did you do to get the printout?" For some reason, I always expected the answer to be different from the times previous; I expected him to tell me what commands he performed, what information those commands were supposed to give him, and why he needed that information.

"I did some menu things and this came out." As expected.

Then he and I would sit down and go over the five pages of printout, examining what each of the tables and numbers meant in relation to his research. Eventually, after dealing with several "Why is it giving me *that* information?!" questions, Simon would be vaguely satisfied with the printout and could select several statistics from the printout that would provide the information he wanted.

However, there were many more questions I wanted to ask him. Most centered on questions about the validity of the tests performed. I knew, however, that such a line of questioning would be moot with the statistical package he (and his class) used. Either the company that owned the package had the test available, or it did not. There was no (easy) way to add tests and procedures. Thus, Bartlett's test of equal variances was not an option, even when the data was such that it should be analyzed using it. Furthermore, the number of available tests was quite limited.

In addition to the extensibility issue, there is the issue of clicking your way to an analysis. If Simon needed to repeat the exact analysis, except for a tweak or two, he would have to start from scratch and repeat it all, hoping that he did not make a mistake along the way. This repeat analysis happens quite frequently in life.

❧❧❧

One of the many pleasures I have as a statistician is my exposure to many different disciplines. I was consulting for a woman doing research in science education. Her specific problem was to determine if a specific Science Teaching Unit made the students like science more. She performed her experiment on a class of fifth and sixth graders in rural North Dakota. She gave them a 50-question pre-test, taught the Unit, then gave the same students a 50-question post-test.

She then contacted me and sent the data.

On the data she sent to me, I spent about three hours analyzing the data, coming to some interesting (and counter-intuitive) conclusions. In my experience, researchers have a sufficient feel for their discipline that surprising results are usually a function of analysis error. Thus, I checked my analysis for errors.

I was actually able to check the analysis because I wrote a script — a series of commands — detailing every bit of the analysis I performed. Mouse-clicking my way through the analysis would make it all but impossible to check my work (something my math teachers in grade school always emphasized). Thus, I was confident that the analysis I returned to her were correct . . .

. . . conditional on the data being correct.

The next day, she emailed me and let me know that she found several serious errors in her data. Relying on mouse-clicks would have made those original three hours a waste of time. However, once she sent the corrected version of the data, the analysis took 90 seconds. Clicking my way to re-analysis would have taken the same amount of time as the first analysis — time we did not have (we were facing a deadline). Re-running the script only took processing time.

While I *am* a fan of mouse-clicks under many circumstances, I am not a fan when it comes to statistical analysis. Scripting provides three definite advantages over mouse clicks: You get what you request, you can check your work easily, and you can re-run the analysis with little effort.

❧❧❧

On the academic job market, it is customary for a candidate to make a presentation about their current research. June 2007 found me standing in front of a formidable group in a large city in Denmark giving such a presentation. After I had discussed the current literature on the causes of terrorism, my statistical model, the results, and my conclusions, I opened it up for questions. As the position was for a quantitative methodologist (sub-discipline open), the first several questions were about the statistical theory undergirding my analysis. After about five minutes of this, one professor asked how my analysis would change were I to add an offset to the model.

After a brief panic, I decided to actually run the regression with the offset in front of the audience. I apologized for not knowing the answer to his question, but I would be happy to hypothesize the effect of the variable offset while running the analysis. The professor smiled as I tried to open my analysis script to modify and run it. It turned out that `R` was not installed on the presentation computer. No worries. I opened my `R` folder, doubleclicked on the R program, and proceeded with the altered analysis — all the while discussing the theoretical effects of using such an offset under these circumstances. Before I could finish hypothesizing, R gave me the answer (which, thankfully, agreed with my hypotheses).

Now that I had my model laid bare before all, many more questions arose about different alterations I could (or should) make to the model. All of which I was able to perform in front of the now hyper-interested crowd.

It turned out that the professor was smiling because he knew there were no statistical programs on that computer and he wanted to see how I would handle myself under those circumstances.

These three vignettes illustrate many strengths of a statistical environment like R. First, it encourages one to write out their analysis and "show their work." This makes it easier to see the entire scope and sequence of the analysis. It also makes it easier to check for errors. Second, it is extensible. If there is a cutting-edge test or procedure you wish to run, there is probably a package that contains it. If not, you are quite free to write it yourself. Finally, one can carry R around with them on a USB drive, allowing them to perform analyses whenever they have a computer, like in Denmark.

Oh yeah, that R is free is also a nice feature, especially as statistical packages can run from $600 to $6000 and up, *and* can have licenses requiring annual payments. As budgets get tighter, an ability to work successfully with a free (and powerful) statistical environment is invaluable.

The Prerequisites: For any book (or course), there is a necessary assumption made about the background of the reader (or student). For the material in this book, I assume that you know your high school algebra to the point that you do not have to reason through the steps of equation solving. You just do it.

Symbol manipulation is an important skill in many areas of life. It shows that you have abstracted procedures and understand the most important aspects of the procedure.

abstraction

Finally, the ability to abstract problems to additional applications is important. Bloom's Taxonomy (Figure 1) places this as the third level of his cognitive domain. It requires examining an exemplar and determining which

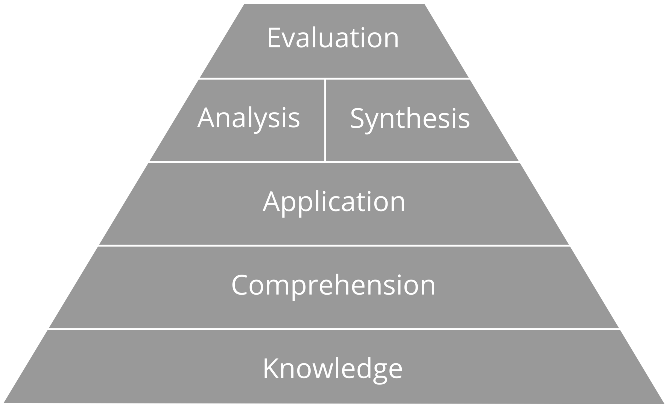

Figure 1: *A representation of Bloom's Taxonomy in the cognitive domain. Note that while the knowledge level is at the lowest level, it supports all other levels.*

parts of that exemplar are most important in selecting correct methods to use.

As you progress through statistics, you will become proficient at more and more methods. You will also discover that there is no perfect method; all make assumptions that exist only in the ideal. You will have to determine which methods are 'good enough' for your application.

perfect method

A Note on Notation: Not surprising, notation varies across the discipline. This is a result of the history of statistics: Many of the methods came from disciplines that used statistics, rather than from Statistics itself. Different disciplines use different notation for the same idea. Thus, any discussion of Survival Analysis needs to include Event History Analysis and Reliability Analysis, as they all study the same phenomena but from different disciplines (medicine, social sciences, and engineering, respectively).

Even within a discipline, there is often a variety of notation used to indicate the same ideas. For instance, probability density functions are often parametrized in different ways. The parameter of the Exponential distribution can be the rate λ or the expected value θ; the second parameter of the Normal distribution (Gaussian distribution, Gauss-Laplace distribution) may be the variance σ^2 or the precision τ^2. In this volume, I will keep consistent with notation, and I will explain the notation before I use it.

To that end, population parameters will be signified using Greek minuscules (a table of which is included below in Table 1). All random variables

minuscules

Minuscule Letter	Majuscule Letter	Name
α	A	alpha
β	B	beta
γ	Γ	gamma
δ	Δ	delta
ϵ, ε	E	epsilon
ζ	Z	zeta
η	H	eta
θ	Θ	theta
ι	I	iota
κ	K	kappa
λ	L	lambda
μ	M	mu
ν	N	nu
ξ	Ξ	xi
o	O	omicron
π	Π	pi
ρ	R	rho
σ	Σ	sigma
τ	T	tau
υ	Υ	upsilon
ϕ	Φ	phi
χ	X	chi
ψ	Ψ	psi
ω	Ω	omega

Table 1: *The entire Greek alphabet in the canonical order. Being familiar with the letters will make it easier to recognize the implied meaning behind the letter.*

majuscules

data

are Roman majuscules. All realized random variables (also known as data) are Roman minuscules. Violations of these rules will exist, but should be kept to a minimum.

Thus, if we theorize that our measurements come from a population that is Normally distributed, with mean 15 and standard deviation 10, I would write this as

$$X \sim \mathcal{N}(\mu = 15, \sigma = 10),$$

where the mean of the population is denoted by μ, the standard deviation by σ, and the Normal distribution by $\mathcal{N}$.

Now, once we take those measurements, we would call the variable x. The difference between random variables and realizations of those random variables is that the random variable has a probability distribution associated with it; the realized data are just numbers.

A list of the important probability distributions is provided in Appendices A and B. Those appendices provide the distribution's name, probability mass (or density) function (pmf or pdf), cumulative distribution function (CDF), symbol, mean, variance, and median, along with notes about the distribution and some related distributions. The appendices also provide some example and practice that allows you to better understand the distributions.

And so, with all of this said, turn the page and begin your trek through statistics. The first part daels with collecting and summarizing data. It is an excellent place to start, as you cannot analyze data without data.