# Advanced Behavioral Statistics

Paul Meinz, Ph.D

# Contents

# Chapter 1

# Preface

## 1.1 You can learn statistics

Many textbooks in mathematics begin with a useful high level overview of the content of book. We'll get to that in a moment, but I want to start this book off a little differently. *You can learn statistics.* I believe that you can learn statistics and I'm happy you have opened this book to do it. Statistics is a wonderful field, and I am excited for you to learn it. I believe in you.

## 1.2 Why I wrote this book

First and foremost, I wrote this book because I'm excited about statistics (really excited; who writes a textbook for fun?), and I want other people to be excited about statistics too.

Second - and most importantly - the reason I wrote this textbook is because I would have appreciated a book like this as a behavioral scientist learning statistics. Over the years, I became very comfortable with the applied end of statistics, but as a graduate student I found myself approaching statistics like a recipe book, e.g., if you have this type of data, it should be analyzed in this way. Over the course of my studies and career, I picked up "explanations" for why certain analyses were appropriate, and although these explanations were accurate (in a sense) they were sometimes quite amorphous. I have grown to appreciate the concrete and exact nature of mathematics - particularly the field of mathematical statistics. And, as a young graduate student, I would have appreciated a deeper understanding of *why* in the concrete mathematical sense.

As I studied mathematics and mathematical statistics throughout my career, I came to realize that a deeper understanding is within the reach of the graduate

student in the behavioral sciences. I decided to write a book that takes a student - perhaps with an okay memory of algebra and maybe with some knowledge of calculus (not at all required) - to a more mathematical understanding of statistics.

## 1.3 What content to expect in this book

To that end, this book will review many concepts learned in an undergraduate behavioral statistics course, but with more mathematical emphasis. This book can be categorized into three parts:

1. Chapters 1, 2, and 3 give a background on probability and probability distributions. Within behavioral statistics courses, we often discuss a vaguely defined "population" of study. These three chapters define more concretely what we mean when we refer to a population or populations. Hint: Populations can be thought of as random variables and their corresponding **probability distributions**. We'll dig into this later.
2. Chapter 4 is our first foray into statistical inference. While Chapters 1, 2, and 3 tell us about populations, this chapter will begin our journey of **inferring** some basic things about these probability distributions. We will learn about the **Law of Large Numbers** and the amazing **Central Limit Theorem**.
3. Chapter 5 (and the chapters thereafter) will introduce us to hypothesis testing - the cornerstone of statistical inference in the behavioral sciences. We will cover several statistical techniques such as $z$-tests, $t$-tests, correlation, chi-square tests, and ANOVA.

Thank you for reading!

## 1.4 What you'll need to know to get the most out of this book

The most important "thing" you need to read this book is excitement. I would not describe myself personally as a mathematical prodigy in any way. I'm just excited about the subject, and that provides me with the extra boost to understand the difficult stuff. The second thing you'll need is an okay grasp of algebra. If you are rusty, I suspect enough will come back to you to understand everything in this book. I'll try to give gentle reviews where I can throughout. There is lots of content online (like this book) that would offer an awesome review of algebra, so if you are confused and I don't provide a review, you could almost certainly find something to get you up to speed.

## 1.5 Future editions of this book

This book is a working document. I plan to work in practice problems and tutorials on R sometime in the future. Stay tuned.

# Chapter 2

# Set Theory and Probability

In the course of research, behavioral scientists will often take samples of human behavior, e.g. a survey of a group of people, measurements of reaction time, etc. These samples will take some *outcome*, and most of us understand and have observed the inherent randomness of this act. That is to say, you might take the exact same sample in a carefully controlled circumstance and get different results every time.

Statisticians often refer to an analogous circumstance when talking about probability. Suppose we are able to describe all the possible outcomes of our sampling procedure. We take samples over and over again in the same circumstances, and each time the outcome changes. This is referred to as a **random experiment**. In order to familiarize you with some introductory concepts around probability, we're going to start in this context of a random experiment. First, we'll learn about sets and set theory in order to describe our **sample space**. This will also be useful in describing the form that outcomes of a random experiment may take. Then we'll more concretely define the randomness of our experiment by talking about **probability** - the quantification of randomness.

## 2.1 Sets and Basic Set Theory

Before we get into it, we aren't going to learn set theory in it's entirety. That would be overly (and hilariously) ambitious. What we are going to learn will make understanding later sections easier, and maybe it'll make it a bit easier to understand set notation in other circumstances. The first step in conducting a random experiment is clearly defining all the possible outcomes of your sampling procedure. This is referred to as the **sample space**. Suppose we have five different possible outcomes in our sample space: A, B, C, D, and E. Let's call our sample space $S$. Then we would describe our sample space like this:

$$S = \{A, B, C, D, E\}$$

That is to say we'll describe the set $S$ with curly braces, and on the inside of the curly braces, there shall be a list of *distinct* elements of our set. The sample space above is **discrete** - meaning it either contains a finite number of elements or it is *countable*. Here the sample space is finite (it has five elements). That's pretty straight forward.

On the other hand, the definition of countable can be tricky to grasp at first, and if you don't understand it here, my hope is you can gain a small intiution for it. Specifically, a discrete sample space can also contain an infinite number of elements - so long as those elements can be lined up one-to-one with the natural numbers (e.g. 1, 2, 3, 4, ...). This is probably not the most intuitive definition for people who don't think about mathematics every day (it certainly wasn't for me). In another, more intuitive sense, a discrete sample space - finite or infinite - has elements with nothing in between them. In $S$ above, there is nothing in between A and B, for example. If your discrete set had an infinite number of elements, you could also find nothing between the first and second elements.

With that side point out of the way, we may also find ourselves sampling from a **continuous** sample space, e.g. all the values between and including 0 and 1 (e.g $0 \leq x \leq 1$). Since you can't enumerate all the values of a continuous sample space as we did $S$ (people have "tried"; we'll talk about one such example later), we use a different form of notation. Suppose we have a continuous sample space $R$, then we would specify it as

$$T = \{x \,|\, 0 \leq x \leq 1\}$$

The inclusion criteria for the set is written after the | in the curly braces. The | can be read loosely as meaning "where", so reading the whole thing in the curly braces left to right you get x *where* $0 \leq x \leq 1$. Therefore, our set is all the $x$'s from 0 to 1.

Now that we have discussed some more concrete ways of describing our sample space, let's talk about sampling from it. Suppose we are randomly drawing a single item from our sample space $S = \{A, B, C, D, E\}$. We shall call a particular outcome an *event*. Events are described by a *subset* of the sample space. For example, we might say the event is $C = \{A, B, C\}$, and if our sample lands in the that subset (as either A, B, or C), then for the sake of brevity I shall say that event was a success (congratulations all around!). You can indicate a set is a subset of another set by:

$$C \subset S$$

We might also, for example say that our event is a single element $C_1 = \{A\}$ or the entire sample space $C_2 = \{A, B, C, D, E\}$. We can be as creative as

we would like to be, so long as the event is in our sample space. Speaking of creativity, suppose we have two events $C_1 = \{A, B, C\}$ and $C_2 = \{B, C, D\}$, and we are interested in whether or not our randomly sampled element lands in $C_1$ or $C_2$. Well, it's helpful here to consider the circumstances where this event occurs. Namely, If our realized element was an A, B, C, or D, then we would be successful (success!). That is to say, we've taken all the unique elements of either set and combined them into a new set. If our sample lands in that new space, we can be confident that $C_1$ or $C_2$ occurred. There's a handy symbol that indicates this operation:

$$C_1 \cup C_1 = \{A, B, C\} \cup \{B, C, D\} = \{A, B, C, D\}$$

The *union operator* takes the unique elements of both sets and puts them into a set of their own. In this circumstance it is used synonymous with *or*. If the event of interest is our sample landing in $C_1$ OR $C_2$, then we would be successful if the event landed in the union of the two sets, e.g. all the unique elements from both sets. Here's another example, suppose we are interested in if our sample falls into $ C_1 = {A} $ or $ C_2 = {B} $. Then our event would occur if the element was A or B:

$$\{A\} \cup \{B\} = \{A, B\}$$

Since we're being creative, let's consider another scenario (and a new operator!). Consider our two previously described events - $C_1 = \{A, B, C\}$ and $C_2 = \{B, C, D\}$. Suppose we are interested in whether or both events occur when we sample our element. That is, we shall be successful if our element falls in both $C_1$ and $C_2$. We are therefore interested in the *shared* elements of our two events. If we draw a B or C, then we would be successful, but an A or D would be bad news for us because each is unique to a set! The *intersect operator* will help us signify this process:

$$C_1 \cap C_2 = \{A, B, C\} \cap \{B, C, D\} = \{B, C\}$$

Unlike $\cup$, the intersect operator represents **and**. That is to say, if we are interested in the event that our single draw lands in one set AND another, then the intersect is used. This operator can create some interesting conundrums, so it's worthwhile to do another example. This time let's define our sets to be $C_1 = A, B$ and $C_2 = C, D$. Now we're interested in whether or not our sample falls in $C_1$ and $C_2$.

$$C_1 \cap C_2 = \{A, B\} \cap \{C, D\} = \{\}$$

Wow, wait a second, what happened there? The answer is our two sets had no overlap. You couldn't possible draw a single element that landed in *both* the sets. As a result, we get a very special set - the *null set* - $\{\}$

The continuous case is pretty much the same although it takes some drawing of number lines if you are rusty with inequalities (no problem). Suppose we are sampling from a sample space $S = \{x \,|\, 0 \leq x \leq 10\}$ and we are interested in if the event lands in $C_1 = \{x \,|\, 0 \leq x \leq 2\}$ OR $C_2 = \{x \,|\, 2 \leq x \leq 5\}$. A successful sample here would occur if our sample landed anywhere from zero to five. In other words:

$$\{0 \leq x \leq 2\} \cup \{2 \leq x \leq 5\} = \{0 \leq x \leq 5\}$$