Unit Descriptions

I: Descriptive Statistics

The first instinct of the scientist should be to organize carefully a question of interest, and to collect some data about this question. How to collect good data is a real and important issue, but one we discuss later. Let us instead assume for the moment that we have some data, good or bad, and first consider what to do with them[1]. In particular, we want to describe them, both graphically and with numbers that summarize some of their features.

We will start by making some basic definitions of terminology – words like **individual**, **population**, **variable**, **mean**, **median**, *etc.* – which it will be important for the student to understand carefully and completely. So let's briefly discuss what a definition *is*, in mathematics.

Mathematical definitions should be perfectly precise because they do not *describe* something which is observed out there in the world, since such descriptive definitions might have fuzzy edges. In biology, for example, whether a virus is considered "alive" could be subject to some debate: viruses have some of the characteristics of life, but not others. This makes a mathematician nervous.

When we look at math, however, we should always know exactly which objects satisfy some definition and which do not. For example, an *even number* is a whole number which is two times some other whole number. We can always tell whether some number n is even, then, by simply checking if there is some other number k for which the arithmetic statement $n = 2k$ is true: if so, n is even, if not, n is not even. If you claim a number n is even, you need just state what is the corresponding k; if claim it is not even, you have to somehow give a convincing, detailed explanation (dare we call it a "proof") that such a k simply does not exist.

So it is important to learn mathematical definitions carefully, to know what the criteria are for a definition, to know examples that satisfy some definition and other examples which do not.

Note, finally, that in statistics, since we are using mathematics in the real world, there will be some terms (like **individual** and **population**) which will not be exclusively in the mathematical realm and will therefore have less perfectly mathematical definitions. Nevertheless, students should try to be as clear and precise as possible.

The material in this Part is naturally broken into two cases, depending upon whether we measure a single thing about a collection of individuals or we make several measurements. The first case is called **one-variable statistics**, and will be our first major topic. The second case could potentially go as far as **multi-variable statistics**, but we will mostly talk about situations where we make *two* measurements, our second major topic. In this case of **bivariate statistics**, we will not only describe each variable separately (both graphically and numerically), but we will also describe their relationship, graphically and numerically as well.

II: Good Data

It is something of an aphorism among statisticians that

> *The plural of anecdote is not data.*[7]

The distinction being emphasized here is between the information we might get from a personal experience or a friend's funny story – an anecdote – and the cold, hard, objective information on which we want to base our scientific investigations of the world – data.

In this Part, our goal is to discuss aspects of getting good data. It may seem counter-intuitive, but the first step in that direction is to develop some of the foundations of *probability theory*, the mathematical study of systems which are non-deterministic – random – but in a consistent way. The reason for this is that the easiest and most reliable way to ensure objectivity in data, to suppress personal choices which may result in biased information from which we cannot draw universal, scientific conclusions, is to collect your data *randomly*. Randomness is a tool which the scientist introduces intentionally and carefully, as barrier against bias, in the collection of high quality data. But this strategy only works if we can understand how to extract precise information even in the presence of randomness – hence the importance of studying probability theory.

After a chapter on probability, we move on to a discussion of some fundamentals of *experimental design* – starting, not surprisingly, with *randomization*, but finishing with the gold standard for experiments (on humans, at least): *randomized, placebo-*

controlled, double-blind experiments [RCTs]. Experiments whose subjects are not humans share some, but not all, of these design goals

It turns out that, historically, a number of experiments with human subjects have had very questionable moral foundations, so it is very important to stop, as we do in the last chapter of this Part, to build a outline of *experimental ethics*.

III: Inferential Statistics

We are now ready to make (some) inferences about the real world based on data – this subject is called **inferential statistics**. We have seen how to display and interpret 1- and 2-variable data. We have seen how to design experiments, particularly experiments whose results might tell us something about cause and effect in the real world. We even have some principles to help us do such experimentation ethically, should our subjects be human beings. Our experimental design principles use randomness (to avoid bias), and we have even studied the basics of probability theory, which will allow us to draw the best possible conclusions in the presence of randomness.

What remains to do in this part is to start putting the pieces together. In particular, we shall be interested in drawing the best possible conclusions about some population parameter of interest, based on data from a sample. Since we know always to seek simple random samples (again, to avoid bias), our inferences will be never be completely sure, instead they will be built on (a little bit of) probability theory.

The basic tools we describe for this inferential statistics are the *confidence interval* and the *hypothesis test* (also called *test of significance*). In the first chapter of this Part, we start with the easiest cases of these tools, when they are applied to inferences about the population mean of a quantitative RV. Before we do that, we have to discuss the *Central Limit Theorem [CLT]*, which is both crucial to those tools and one of the most powerful and subtle theorems of statistics.