The Science of Statistics

Nathaniel E. Helwig

University of Minnesota

1 What is Statistics?

The field of "statistics" is a branch of mathematics concerned with testing hypotheses, modeling and predicting data, and quantifying the uncertainty of data based conclusions. Unlike the (recently popular) field of "data science", which is primarily concerned with obtaining insights from collected data, the field of Statistics is primarily concerned with understanding and quantifying the confidence that we can have in conclusions drawn from collected data. In Statistics, we are concerned with clearly defining the population of interest and what assumptions we want to make about that population, which enables us to quantify the uncertainty of inferences that are drawn from a sample of data collected from the population. Thus, the science of Statistics can be thought of as the *science of uncertainty quantification*.

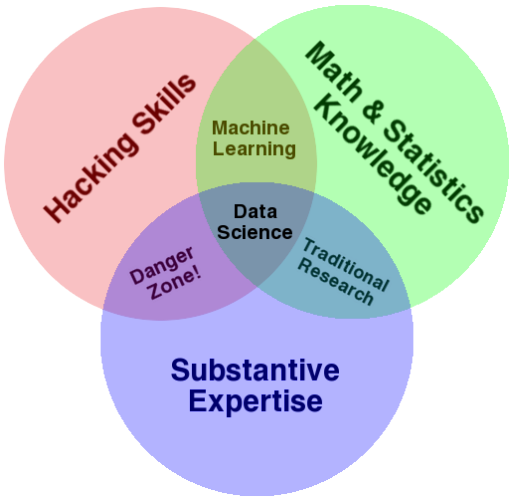

Figure 1: Data science Venn diagram by Drew Conway. Note that Data Science is the intersection of computer (hacking) skills, Statistics knowledge, and domain knowledge. Without Statistics knowledge, you are in the "Danger Zone", which is a terrible place to be! http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram.

2 Populations and Samples

Any statistical problem begins by defining the population of interest, as well as the sampling procedure that will be used to sample (i.e., collect) observations from the population.

Definition. In Statistics, the population refers to the set of objects from which data will be collected, and the sample refers to the set of objects that are collected from the population.

At any given point in time, the population of interest is fixed, whereas the sample may (and likely will) vary across different studies. The population of interest will depend on a given study's goals, and will affect the generalizability of the study's results. You can think of the population as the (possibly infinite) group of objects that you want to infer things about, and the sample as the (finite) group of objects that you observe to draw those inferences. Thus, it is important that the sample of data is collected in a way such that it is representative of the population of interest. If the sample is obtained in an unrepresentative way, then the results of the study may not generalize well to the population of interest.

***Example* 1.** Suppose that you want to study the mathematics aptitude of high school students in Minnesota. In this case, the population is all high school students in Minnesota. The sample is the collection of students that are observed in your study, which should be representative of the population. If you only sample students from a single school district, this may not be a representative sample—so your results may not generalize well to all high school students in Minnesota.

***Example* 2.** Suppose that you want to study functional connectivity in the brains of schizophrenics. In this case, the population is all individuals with schizophrenia. The sample is the collection of schizophrenics that are observed in your study, which should be representative of the population. If you only sample schizophrenics from a single location (e.g., only those seeing a particular doctor), this may not be a representative sample—so your results may not generalize well to all individuals with schizophrenia.

Thus, if you want your study's results to be generalizable to the intended population, it is important to carefully devise a sampling technique that will ensure that your sample is reasonably representative of your population of interest. You cannot—and should not—expect your results will generalize to populations that are not represented in your sample.

3 Sampling Techniques

Definition. In Statistics, a sampling technique refers to the methodology that is used to obtain a sample of objects from the population of interest.

Some popular sampling techniques:

1. *Simple random sampling*: all objects in the population have an equal chance of being included in the sample.
2. *Stratified sampling*: the population is divided into groups (i.e., strata) based on some defining characteristic(s), and simple random sampling is applied within each strata.
3. *Clustered sampling*: the population is divided into groups (i.e., clusters) based on some defining characteristic(s), and simple random sampling is used to select clusters.
4. *Convenience sampling*: any individuals who are willing and able to participate in the study are included in the sample.
5. *Subjective sampling*: the researcher decides who should participate in the study, and selects individuals who are (subjectively) thought to be ideal participants.

With any sampling method, a researcher needs to be worried about potential bias, i.e., lack of representativeness of the sample with respect to the population. Sampling techniques 1-3 are probability based sampling methods, which have the potential to produce reasonably unbiased samples—but may be difficult to implement in practice. In contrast, sampling techniques 4-5 are much easier to implement in practice, but have substantially more potential to result in unrepresentative samples. Unfortunately, many psychological studies seem to rely on convenience sampling, which can make it difficult (or impossible) to understand which population is being studied.

***Example* 3.** Returning to the high school mathematics aptitude example, we would likely want to use either stratified or clustered sampling, where the school districts (or individual high schools) are used to group the students.

***Example* 4.** Returning to the schizophrenic functional connectivity example, we many want to consider stratified sampling where strata are formed using various symptomatology and treatment variables, such as severity of symptoms, medication history, etc.

4 Descriptive versus Inferential Statistics

Colloquially usages of the word "statistics" (such as in sports) are typically referring to descriptive statistics, whereas scientific usages of the word "statistics" (such as in research articles) are typically referring to inferential statistics.

Definition. Descriptive statistics involve the calculation of quantities that summarize collected samples of data, whereas inferential statistics involve using sample quantities to infer things about the population of interest.

Note that descriptive statistics are useful for summarizing our sample of data, but, on their own, descriptive statistics tell us nothing about the population from which the data were sampled. In research studies, we are typically interested in using our collected sample of data to make inferences (i.e., draw conclusions) about some population of interest—so most of this course will be focused on inferential statistics. But descriptive statistics are important as well. As we shall see throughout the course, many descriptive statistics (such as the mean and standard deviation) can be viewed as estimates of population quantities if certain assumptions are made about the collected sample of data. Thus, descriptive statistics can be viewed as the building blocks that are used to construct inferential statistics—in other words, to infer things about a population, we first need to be able to describe our sample of data collected from the population.

5 Bayesian versus Frequentist Statistics

Another distinction that you will frequently hear in statistics is that of Bayesian versus Frequentist statistics. In a nutshell, Bayesians treat population parameters as unobserved random variables, whereas Frequentists treat population parameters as unknown fixed constants.[1] Historically, there has been a lot of fuss about which framework is correct for analyzing data. Nowadays, most (level-headed) statisticians would agree that there are benefits and drawbacks to both statistical philosophies. This course will be primarily focused on Frequentist statistics, given that (i) Frequentist statistical methods are more popular in traditional psychological research, and (ii) a thorough understanding of Frequentist statistics is required for understanding Bayesian statistics.

[1] We haven't yet defined a "parameter" or "random variable", so it is okay if this distinction seems opaque.