The Science of Statistics

Nathaniel E. Helwig

Associate Professor of Psychology and Statistics
University of Minnesota

August 26, 2020

Table of Contents

Table of Contents

Statistics: The Original Data Science

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.

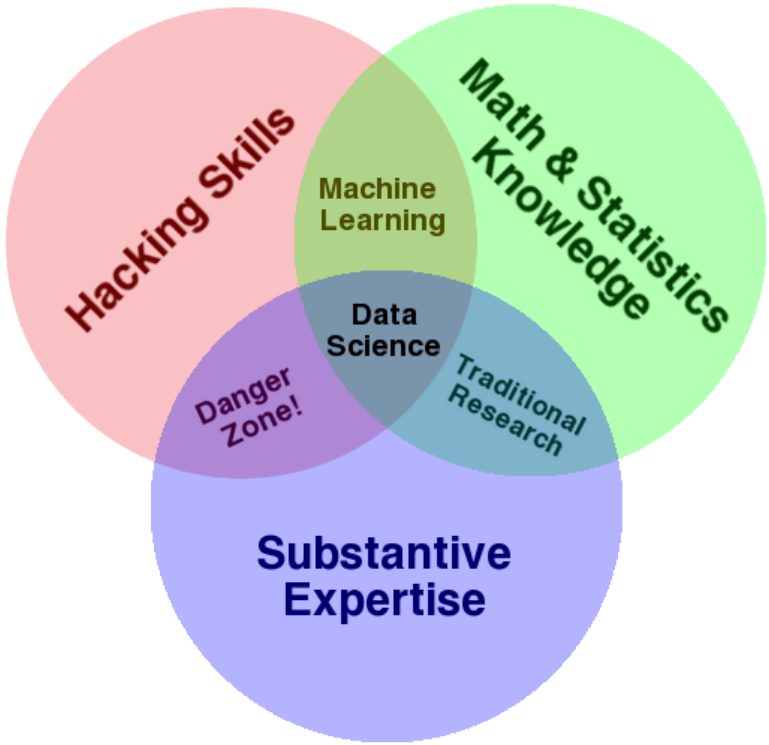

Figure 1: Data science Venn diagram by Drew Conway. Without Statistics knowledge, you are in the "Danger Zone", which is a terrible place to be!

http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Table of Contents

Population (Theory World) and Sample (Real World)

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.

You cannot—and should not—expect your results will generalize to populations that are not represented in your sample.

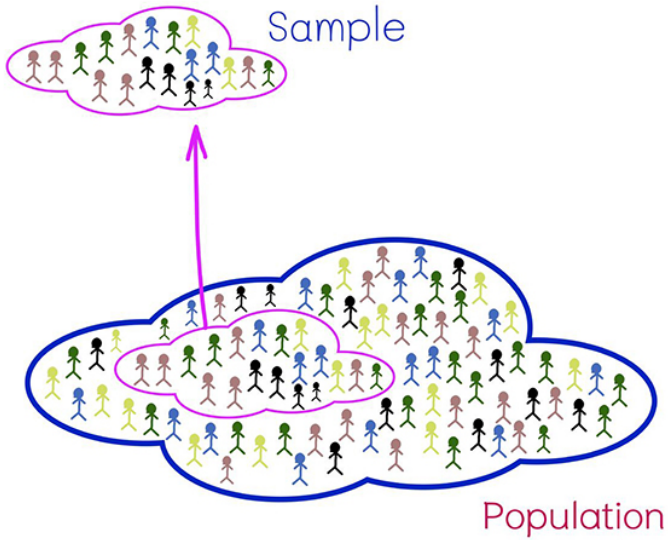

Figure 2: Depiction of a population and a sample.

http://analytics-magazine.org/survey-sampling/

Examples of Populations and Samples

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.

- Sampling a single school district will not be representative

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.

- Sampling a single doctor's patients will not be representative

Table of Contents

Common Sampling Techniques

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

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.

Sampling Bias and Unrepresentative Samples

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.

Sampling techniques 4-5 are much easier to implement in practice, but have substantially more potential to result in unrepresentative samples.

Many psychological studies seem to rely on convenience sampling.

Table of Contents

Descriptive versus Inferential Statistics

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

Many descriptive statistics (such as the mean and standard deviation) can be viewed as estimates of population parameters if certain assumptions are made about the collected sample of data.

Descriptive statistics can be viewed as the building blocks that are used to construct inferential statistics

Parameters and Statistics

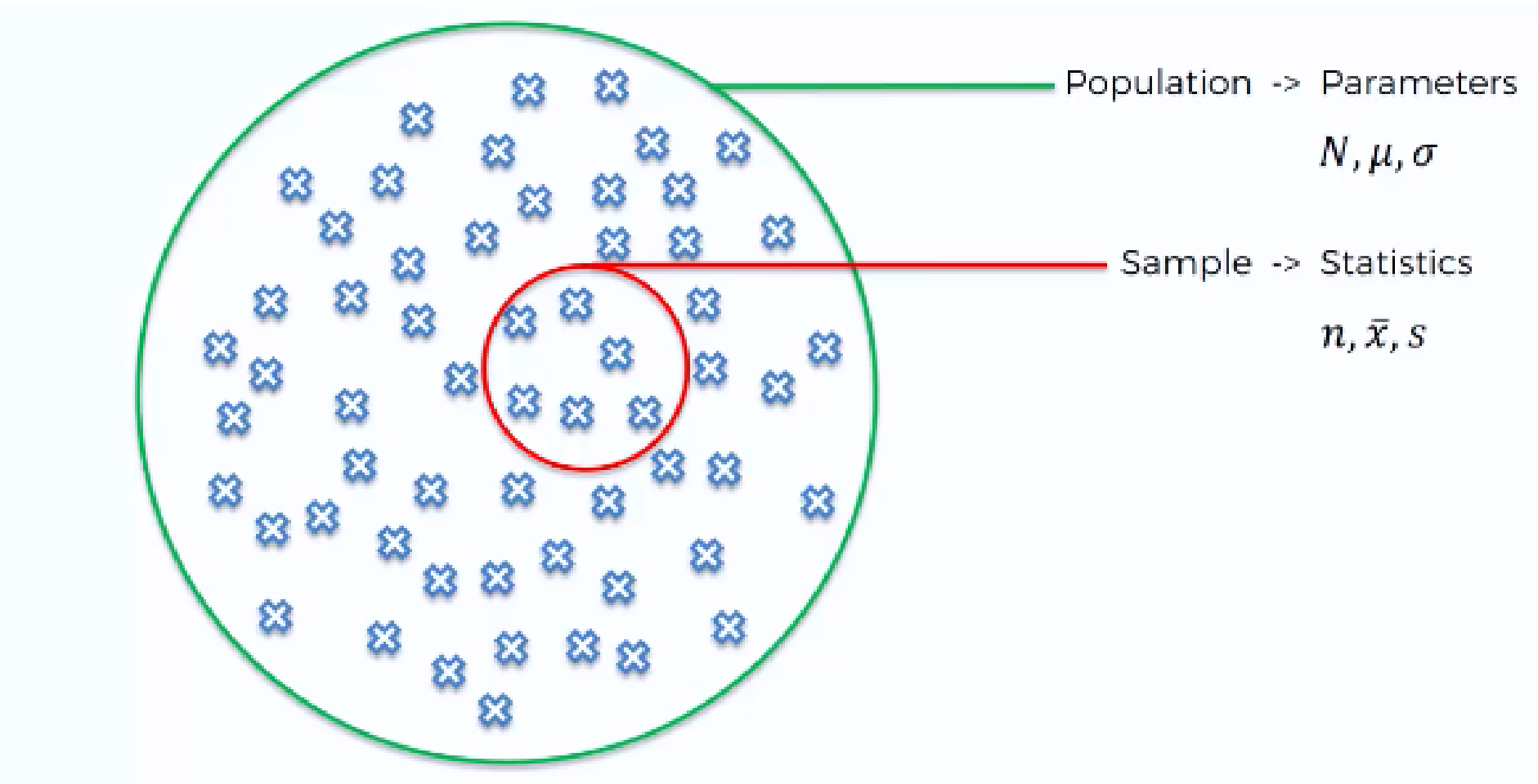

Figure 3: Depiction of parameters and statistics.

https://www.kdnuggets.com/2018/12/introduction-statistics-data-science.html

Table of Contents

Bayesian versus Frequentist Statistics

Another distinction that you will frequently hear in statistics is that of Bayesian versus Frequentist statistical methods.

- Bayesians treat population parameters as unobserved variables.
- Frequentists treat population parameters as unknown constants.

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 concepts from Frequentist statistical methods are...

- more popular in traditional psychological research
- required for understanding Bayesian statistics