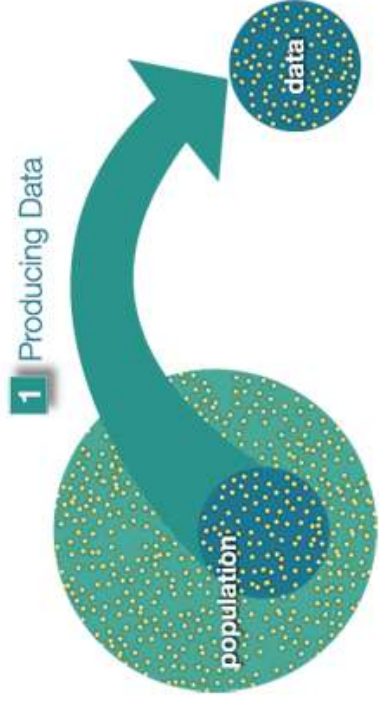# An Introduction to Statistics

Statistics plays a significant role across the physical and social sciences and is arguably the most salient point of intersection between diverse disciplines given that researchers constantly communicate information on varied topics through the common language of statistics.

In a nutshell, what statistics is all about is *converting data into useful information*. Statistics is therefore a process where we are:

- Collecting Data
- Summarizing Data, and
- Interpreting Data

The process of statistics starts when we identify what group we want to study or learn something about. We call this group the **population.** Note the word "population" here (and in the entire course) is not just used to refer to people; it is used in the more broad statistical sense, where population can refer not only to people, but also to animals, things, etc. For example, we might be interested in:

- The opinions of the population of U.S. adults about the death penalty
- How the population of mice react to a certain chemical
- The average price of the population of all one-bedroom apartments in a certain city

Population, then, is the entire group that is the target of our interest.

In most cases, the population is so large that as much as we want to, there is absolutely no way that we can study all of it (imagine trying to get opinions of *all* U.S. adults about the death penalty…).

A more practical approach would be to examine and collect data only from a sub-group of the population, which we call a **sample**. We call this first step, which involves choosing a sample and collecting data from it, *Producing Data.*

Since, for practical reasons, we need to compromise and examine only a sub-group of the population rather than the whole

Producing Data

1

Exploratory Data Analysis

2

data

Probability

3

population

Inference

4

---

population, we should make an effort to choose a sample in such a way that it will represent the population as well.

For example, if we choose a sample from the population of U.S. adults, and ask their opinions about the death penalty, we do not want our sample to consist of only Republicans or only Democrats.

Once the data have been collected, what we have is a long list of answers to questions, or numbers, and in order to explore and make sense of the data, we need to summarize that list in a meaningful way. This second step, which consists of summarizing the collected data, is called *Exploratory Data Analysis.*

Now we've obtained the sample results and summarized them, but we are not done. Remember that our goal is to study the population, so what we want is to be able to draw conclusions about the population based on the sample results. Before we can do so, we need to look at how the sample we're using may differ from the population as a whole, so that we can factor that into our analysis.

Finally, we can use what we've discovered about our sample to draw conclusions about our population. We call this final step *Inference.* This is the *Big Picture of Statistics.*
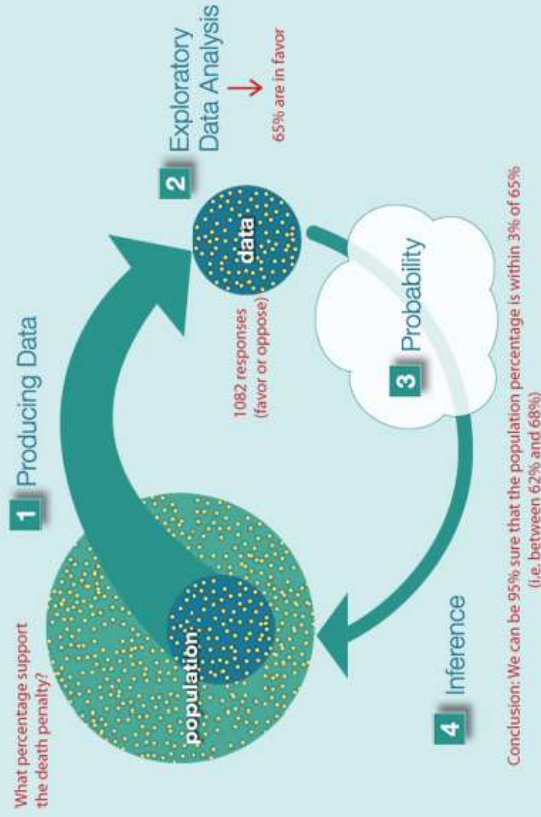
Since we will be relying on data that has already been produced, the focus of your individual project will be <u>exploratory and inferential data analysis.</u>
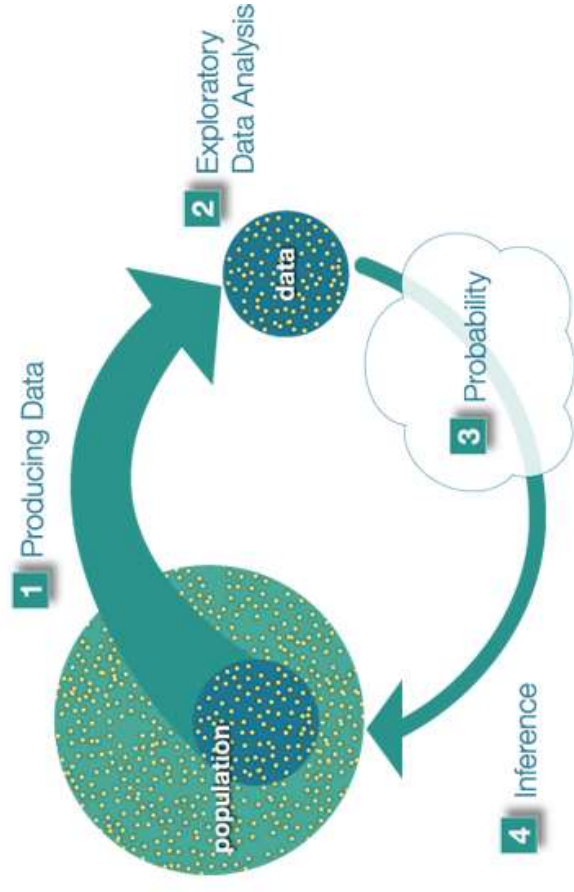
EXAMPLE

At the end of April 2005, a poll was conducted (by ABC News and the Washington Post), for the purpose of learning the opinions of U.S. adults about the death penalty.

1. **Producing Data:** A (representative) sample of 1,082 U.S. adults was chosen, and each adult was asked whether he or she favored or opposed the death penalty.

2. **Exploratory Data Analysis (EDA):** The collected data was summarized, and it was found that 65% of the sample's adults favor the death penalty for persons convicted of murder.

3. **Inference:** Based on the sample result (of 65% favoring the death penalty), it was concluded (within 95% confidence) that the percentage of those who favor the death penalty in the population is within 3% of what was obtained in the sample (i.e., between 62% and 68%). The following figure summarizes the example:

What percentage support the death penalty?

population

1 Producing Data

1082 responses (favor or oppose)

data

2 Exploratory Data Analysis

65% are in favor

3 Probability

4 Inference

Conclusion: We can be 95% sure that the population percentage is within 3% of 65% (i.e. between 62% and 68%)

Final Notes:

Statistics education is often conducted within a discipline specific context or as generic mathematical training. Our goal is instead to create meaningful dialogue across disciplines. Ultimately, this experience is aimed at helping you on your way to engaging in interdisciplinary scholarship at the highest levels.

MANAGING YOUR TIME

You will spend your time in this course working in a variety of ways. You will need to review codebooks and literature, perform data management, do analyses, make graphs, write and revise your findings and read and/or watch IBook lessons.

Lessons ■ Reviewing Codebooks ■ Reviewing Literature
Writing and Revising ■ Management, Analyses, Graphing