ST 516: Foundations of Data Analytics

Types of Studies

Data Collection

- How are Data Collected?
- What Data are Collected?

Types of Studies

- Randomized Experiments
- Observational Studies

Data Collection

Data are collected in many different ways. Just a few examples are:

- Internet retail businesses collect data about customers and their purchases by gathering information at the point of sale, and/or by tracking customers "click paths" through a website.
- Environmental scientists measure all aspects of weather and climate across the whole globe using sensors on the ground, in the water, in the atmosphere, and remotely using satellite technology.
- Cancer researchers collect genetic data about individual types of tumors in the hopes of developing specific therapies and medications.
- Education researchers collect data about student performance under different types of teaching modalities.

Data Collection

It's important to think about *how* the data you are analyzing were collected:

- Were the data collected on all possible subjects or "units?" in some population that you're studying (for example, all customers? all types of tumors?)
- If the data were collected only on a subset of units in some population (we call this a sample of units, or simply, a **sample**), then how was that sample obtained (by convenience? randomly? only for a certain period of time?)
- What was the **protocol** for obtaining data on each of the sampled units? Was it always the same or did it change in some way?

Example

An internet retail company wants to learn how its site is used throughout the day, so it might collect data on the number of visits to its site at different times throughout the day.

1. Are the observational units the site visits or the customers or something else?

2. Is the time frame for the study only one day of the week or only one week or only one month?

3. Are *all* visits to the site to be observed or only some subsample of visits?

4. What if a customer visits the website for less than 2 seconds?

Data Collection

It's also important to think about what types of data were collected, how they may relate to each other, and how they relate to some question of interest.

Example Continued:

1. Do repeat visits to the website by the same customer count as *separate* visits, or just one visit?

2. Is "time of day" relative (i.e., to the customer's time zone) or absolute (i.e., the company headquarters time zone)?

3. Are visits to be classified as successful or unsuccessful relative to whether a purchase was made?

Types of Studies

We broadly classify the methods by which data are collected into two categories:

1. **Randomized Experiments**, in which researchers are able to assign two or more "treatments" or "interventions" to subjects or units in their study, and in which some chance mechanism is used to make that assignment.

2. **Observational Studies**, in which researchers are unable to assign treatments, and so instead they observe some system , an internet site, the genetics of a tumor) as it is.

The type of study that is used for data collection often determines an important component of the **scope of inference** of the study (more on this in a future lecture).

The Case Studies

The creativity case study from *The Statistical Sleuth* is an example of a **randomized experiment**—(a) the researchers controlled assignment of the treatments (intrinsic/extrinsic questionnaire); (b) a chance mechanism was used for this assignment.

The starting salaries case study, by contrast, is an example of an **observational study**—the researchers had no control over the gender groupings, and they just observed the starting salaries of people in those groupings.

Broadly speaking, there are many situations in which it is impossible and/or unethical to assign study units to a treatment (for example, you can't assign gender or cancer status).

Randomized Experiments

1. Researcher controls assignment of experimental units to groups.
2. A chance mechanism (for example, a random number generator) is used to make the group assignments.
3. By using a chance mechanism, the researchers can assert that differences between the experimental units are roughly balanced between the groups before the experiment begins.

Example: In the creativity case study, we can assert that the demographic make-up of the intrinsic and extrinsic groups is roughly the same before the start of the experiment.

Observational Studies

1. No researcher control over group assignments.
2. Possibility of **confounding factors** is high. A confounding factor is one that might provide an alternative explanation for any differences that appear between groups.

Example: In the starting salaries case study, it may be that the men at the bank generally had higher education levels than the women. Therefore, education level is then a confounding factor (or, we say that education level is confounded with gender).