ST 516: Foundations of Data Analytics

Overview

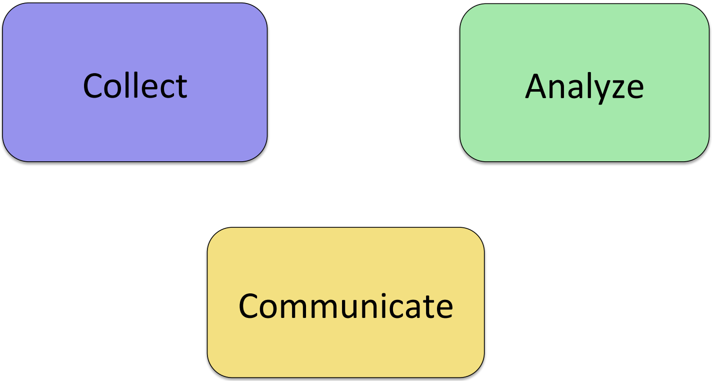
What is data analytics?

Goals of data analysis

Examples

Data science is learning from data

Data science starts with a question



Collect

Analyze

Communicate

Goals of data analysis

- Description: describe the data at hand
- Inference: generalize from our data to the population it came from
  - Estimation: the numeric value of a property of population
  - Hypothesis Testing: does the population have a certain property?
- Prediction (or classification): generalize from our data to future data

Example

We are interested in the support by voters of the current president. We sample 1000 voters and ask “Do you support the current president?”

- Description: what proportion of our sample replied “Yes”?
- Inference: from sample to population
    - Estimation: what proportion of all voters support the president?
    - Hypothesis Testing: is the proportion of all voters that support president greater than 50%?
- Prediction (or classification): If I sample another a voter at random, how likely are they to respond “Yes”?

Statistical Inference

This class focuses on the statistical foundations that let us use the data at hand to make conclusions about the population they came from.

Our conclusions are never certain! Why?

Statistical Inference

This class focuses on the statistical foundations that let us use the data at hand to make conclusions about the population they came from.

Our conclusions are never certain! Why? Because whenever we collect data, variability is unavoidable. If we repeated our study, the numbers we get would be different. By only making a partial observation on the population we can't be certain about the properties of the population. We want to recognize and quantify this uncertainty.

We'll talk about the conditions required to make these conclusions from a sample to a population, as well as how we quantify our uncertainty about the conclusions we make.

What motivates creativity?

Creativity case study 1.1.1 in the Statistical Sleuth

Writers were randomly assigned to receive either:

- an *intrinsic* questionnaire to stimulate a thought pattern that emphasized “doing something because it brings satisfaction”, or
- a *extrinsic* questionnaire to stimulate a thought pattern that emphasized “doing something because it brings external rewards” (money, respect etc.)"

After the questionnaire, writers wrote a haiku. The haikus were judged by a panel of poets on their creativity. The average score for each writer was recorded.

For each writer, the **response** (or outcome) is the creativity score, the **explanatory** variable, is the treatment, whether the writer received the *Intrinsic* or *Extrinsic* questionnaire.

What motivates creativity?

Does giving the extrinsic questionnaire increase the creativity scores? Hypothesis testing (inference)

By how much? Estimation (inference)

Are women paid less than men?

The starting salaries case study 1.1.2 in the Statistical Sleuth

The starting salary for all skilled, entry level clerical employees, hired by a bank between 1969 and 1977 is obtained. During this time there were 32 males and 61 females hired.

For each employee, the response is their starting salary, the explanatory variable is their gender.

Is the average salary of the female employees lower than the male employees? Description

Is the difference in salary between the male and female employees consistent with salaries being assigned at random (no gender discrimination?)? Hypothesis testing (inference)

Does seeding clouds increase rainfall?

An experiment was conducted to examine if seeding a cloud with silver iodide would increase rainfall.

On 52 days, a plane flew through a target cloud and injected its cargo. On 26 days, chosen randomly, the cargo was the chemical of interest, on the other 26 the cargo was empty.

Rainfall from the target cloud was measured by radar.

For each day, the response is the rainfall, the explanatory variable, is the treatment, whether the cloud was *Seeded* or *Unseeded*.

How much more rain results from seeding a cloud? Estimation (inference)

Does the use of stents reduce the risk of strokes?

Researchers gathered 451 volunteers at risk of stroke. Each patient was randomly assigned to either receive a stent (treatment) or not (control). Both groups also received medication, management of risk factors and help in lifestyle modification.

After one year, the recorded whether each patient had experienced a stroke.

For each patient, the response is whether they had a stroke or not, the explanatory variable, is the treatment, whether they had a stent or not.

Does using a stent reduce the risk of a stroke? Hypothesis testing (inference)