Module 1 – Introduction to Statistics

Reading: Chapter 1 (sections not covered: 1.2.4, 1.2.5, 1.4)

Statistics

- the collection, organization, analysis, presentation, and interpretation of data
- "the science of data" – *Montgomery & Runger*
- "the logic of uncertainty" – *Joe Blitzstein*

Why do we need statistics?

We use statistics to understand phenomena that cannot be modeled analytically

- civil engineering: number of cars on a highway
- meteorology: trajectory of a hurricane
- industrial engineering: number of patients on a hospital and in-patient time
- history: establishing the authorship of the Federalist Papers
- finance: identifying trends and patterns in the economy to make investment decisions
- gambling: the historical roots of statistics are in gambling
- life: making decisions based on experience

- All these problems have one thing in common: **variability**
- This means that seemingly identical measurements & observations will produce different results
- We can't simplify these into analytical equations
- Instead, we use statistical methods to create models that can represent these phenomena
- Sometimes, statistics sheds light on nonintuitive phenomena

Monty Hall Problem

- Suppose you're on a game show and you're given the choice of three doors.
- Behind one door is a car, behind the others, goats.
- You pick a door (let's say #1).
- The host (who knows what's behind the doors) opens another door (let's say #3) which has a goat.
- The host asks: "Do you want to switch to Door #2?"
- Is it to your advantage to switch your choice of doors?

Sometimes statistics are not superficially intuitive. We need to analyze our problem mathematically to understand it properly.

Presidential Election Polls

- In 1948, Truman ran against Dewey in a close race.
- The Chicago Daily Tribune ran a telephone poll of how people would vote.
- Dewey won the poll decisively.
- The newspaper published the famous headline "Dewey Defeats Truman"
- Of course, Truman had actually won the election
- What happened?

This is an example of sampling bias. A sample needs to be representative of the population.

- Can you think of a more recent election where the outcome did not coincide with the polls?
- What happened there?

Problem-Solving

The focus of this class will be to solve problems. We will use the PPDAC model to apply the scientific method.

1. Define and understand the problem
2. Plan on what to study
3. Collect data
4. Perform an analysis to find patterns and generate hypotheses
5. Draw conclusions

Statistical Thinking

- When we incorporate variability into the problem-solving and decision-making process.
- Example: Determining the optimal route with multiple stops
    - Distance alone is not the only factor to consider
    - We also consider traffic, gas mileage performance, weather, etc.
    - Statistical thinking would incorporate all the above into determining optimal routes

Some Definitions

- experiment: any occurrence that has more than one possible outcome (also called random experiment)
- random variable: a variable that may have more than one possible value in an experiment
- outcome: the result of an experiment, which is a combination of one or more random variables
- sample space: the set of all possible outcomes in an experiment
- event: a subset of the sample space

- sample spaces and events can be studied using mathematical set theory
- in this class, we will also use a visual approach for sample spaces and events

Performing a Statistical Analysis

In statistical analyses, we often study a sample, and apply our conclusions to the population.

- this means that the sample must be carefully selected to ensure it is statistically significant
- sometimes it is not a matter of how much percentage of a population a sample is, but other factors come into play

There are three ways of studying a sample:

- retrospective study: uses historical data
- observational study: collects data from an undisturbed experiment
- designed experiment: the analyst creates and manages the experiment

Wait, we still haven't defined:

- data: information that can be quantitatively studied

Empirical Models

The result of the statistical analysis is to create an empirical model. This model can then be used to predict future observations.

Empirical models are made from data (observations). This means that the quality and quantity of data will affect our model, this is different from

Mechanistic models which are derived analytically.

- mechanistic models only work if we can quantify all factors that affect the experiment
- in most cases, not all factors can be quantified
    - examples: weather, human behavior
- therefore, empirical models are commonly used for phenomena with large variability

References

Blitzstein, Hwang, *Introduction to Probability*
Jones, "Dewey defeats Truman: The most famous wrong call in electoral history" in *Chicago Tribune*
Montgomery, Runger, *Applied Statistics and Probability for Engineers*
Spiedelhalter, *The Art of Statistics*
vos Savant, "Ask Marylin" in *PARADE Magazine*
Yang, *ENGR 3305 Data Collection & Analysis Lecture Notes*