CMSC320: Introduction to Data Science

James Zhang[*]

February 1, 2024

These are my notes for UMD's CMSC320: Introduction to Data Science, which is an elective. These notes are taken live in class ("live-TEX"-ed). This course is taught by Professor Fardina Fathmul.

Contents

[*] Email: jzhang72@terpmail.umd.edu

§1 Big Data and Data Science Overview

The rise of data sicnece over the last 20 years is partialkly a result of big data. The 3 V's of big data are

- Volume: the amount of data from myriad sources
- Velocity: the speed at which big data is generated
- Variety: the types of data (structured, semi-structured, unstructured)

In this class, we will also explore the data lifecyle which includes stages such as data collection, data processing, exploratory data analysis, data visualization, analysis, hypothesis testing, machine learing, and insight and policy discussion.

§2 Experimental Design

Definition 2.1. **Experimental design** is process of planning, carrying out, and analyzing experiments to test a hypothesis. Experimental design is a crucial aspect of data science that focuses on planning and conducting experiments to gather meaningful data.

There are some rough experimental design steps in a research project

1. Define the problem or research question the problem aims to address
    - When it comes to predicting the future, choose option that maximizes your optimization criteria

Example 2.2

Online retail: you want to test whether the color of the button changes the click through rate.

- **Problem definition**: which version of the button maximizes CTR
- **Optimization crtiera**: maximizng CTR

When it comes to setting up a dataset for this, the sample is the number of website visitors.

- **Control group**: group with existing button color, no changes
- **Experimental group**: group that experiences the change that you want to test
- **Dependent variable**: CTR
- **Independent variable**: the color of the button

Definitions of variable, population, sample, independent variable, dependent variable

Definition 2.3. Confounder variable are extraneous variables that may affect the relationships between dependent and independent variables. An additional factor that is not controlled for but could potentially influence the outcomes.

Definition 2.4. Randomization minimize systematic confounding, reduce the risk of bias in either group being enriched for confounders, and help distribute confounding variables equally

Definition 2.5. Replication ensures certainty and suggests that confounders are less likely to be influencing the outcomes.

Methods for collecting data

- Observational

  Definition 2.6. Cross sectional studies: Collects data from many different individuals at one specific single point of time

  Definition 2.7. Retrospective: look back at case studies

  Definition 2.8. Prospective: following a cohort over a periods of time and observing changes

- Surveys: be careful of wording to not be biased

- Experiments

  Definition 2.9. Placebo effect: A person believes psychologically that a certain treatment is positively affecting them, even though no treatment was given at all.

- Artificial Simulations