Data Collection

DASC 512

Overview

- Study Design
- Sampling and Census
- Sources of Data
- Experimental Design
- Preparing for Data Collection

Study Design

- Recall that the first step in collecting data is to define the research question, including identifying the population
- Population: the entire group of interest
  - Who/what do you want to make inferences about?

Sampling and Census

From the population, we will gather a sample or (rarely) a census

- Sample: collection of data for a subset of the population
  - Representative: represents the population of interest
  - Random: this sample of cases is as likely to be selected as any other

- Census: all observations from the population

Sources of Data

- Potential sources of data may include:
  - ⊘ Anecdotal evidence: not suitable for scientific conclusions
    - "The plural of anecdote is not data."
  - Observational studies: careful collection of data without controlling the conditions
    - Cannot prove causation, but useful when conditions cannot reasonably be controlled
  - Experiments: the preferred source of data under controlled conditions
    - This is the only way to prove causation, but it is expensive and may be infeasible
  - Published datasets: using previously collected data in a new way is valid research!
    - Take care to ensure the data was collected appropriately and fits your research question

Data Biases

- Non-response bias: Those who respond may not be representative
  - According to Pew, only 6% of those surveyed by phone responded in 2018
- Response bias: Experimenters bias responses in some way
  - Double-blind studies and placebos reduce response bias
- Others
  - Misidentified populations
  - Sampling bias
  - Instrumentation bias

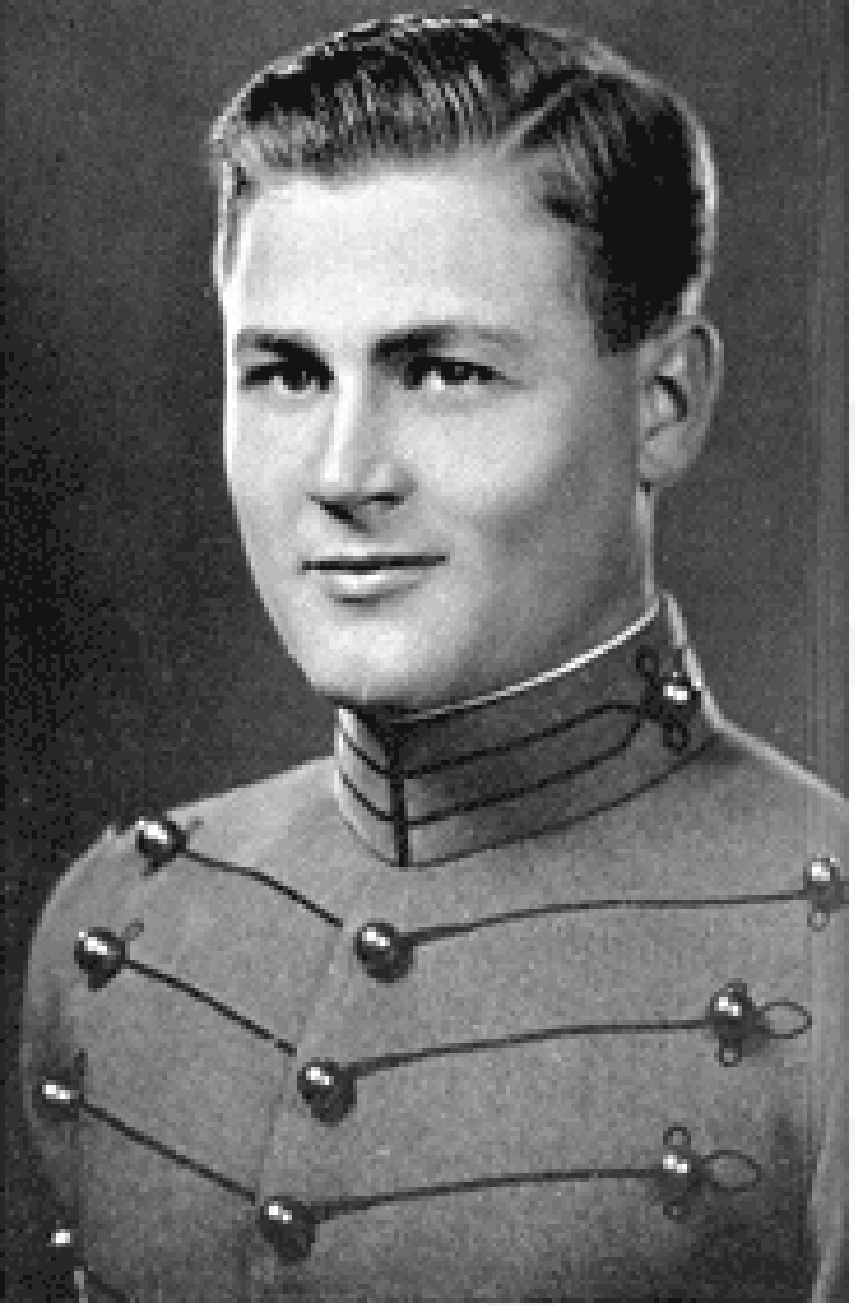
Experimental Design

- Factors or treatments: inputs that are controlled in the experiment
- Cofactors, nuisance factors, or confounding variables: uncontrolled inputs that much be measured if possible
- Principles of Experimental Design:
  - Controlling: confounding variables are controlled as much as possible
  - Randomization: assignment of each treatment should be equally likely
  - Replication: larger sample sizes and/or replicated experiments are preferable
  - Blocking: treatments are assigned equally to notable sub-populations

Preparing for Data Collection

A little forethought here saves your sanity tomorrow!

- How will you deal with errors in the data?
- How will files be named? Should be descriptive, sortable, unique.
- How will data in different files be combined? Does each EU have a unique ID?
- Are variable names stored in headers? Do they match across files?
- Where is data being stored? How often/when is it backed up?
- How will data be transported? If classified, are couriers designated?
- How will observations you didn’t plan for be captured?



On the Shoulders of Giants

- “Anything that can go wrong, will go wrong”
- Edward Aloysius Murphy Jr. (1918-1990)
- US Military Academy Class of 1940
- Attended AFIT in 1947
- R&D Officer at Wright-Patterson AFB 1947-1952

Recap

- Study Design
- Sampling and Census
- Sources of Data
- Experimental Design
- Preparing for Data Collection

Going forward

- This class is about “Applied Statistics” – and application requires critical thinking and communication
- Statistics is best applied with a skeptical view of the data
- Be careful not to overextend your conclusions