

STAT 3704

CH 1

Overture: Engineering Method and Data Collection

Collecting Data

It is important to think carefully about our data collection process and the type of data it will produce before we attempt to organize, and analyze that data. We need to take two main things into account:

1.) Methods:

- Retrospective Studies
- Observational Study
- Designed Experiment

2.) Sampling Techniques

- Simple Random Samples
- Stratified Sampling
- Systematic Sampling
- Cluster Sampling

Treatment and Control Group

- ▶ Treatment Group

Group that receives the treatment (or has the characteristic of interest)


- ▶ Control Group

Group that does not receive the treatment (or does not have the characteristic of interest)

Retrospective Studies

- ▶ Using Data or information that is readily available to us.
- ▶ **Advantages**
 - Cheap, Quick and Easy
- ▶ **Disadvantages**
 - We do not have control over how or what data was collected

Observational Study

- ▶ Observes individuals and measures variables of interest but does not attempt to influence the responses.
 - ▶ Subjects in the study are put into the treatment group or control group either
 - By their own actions
 - By the decision of someone else who is not involved in the research study.
- 

Important Terminology

▶ Time terms

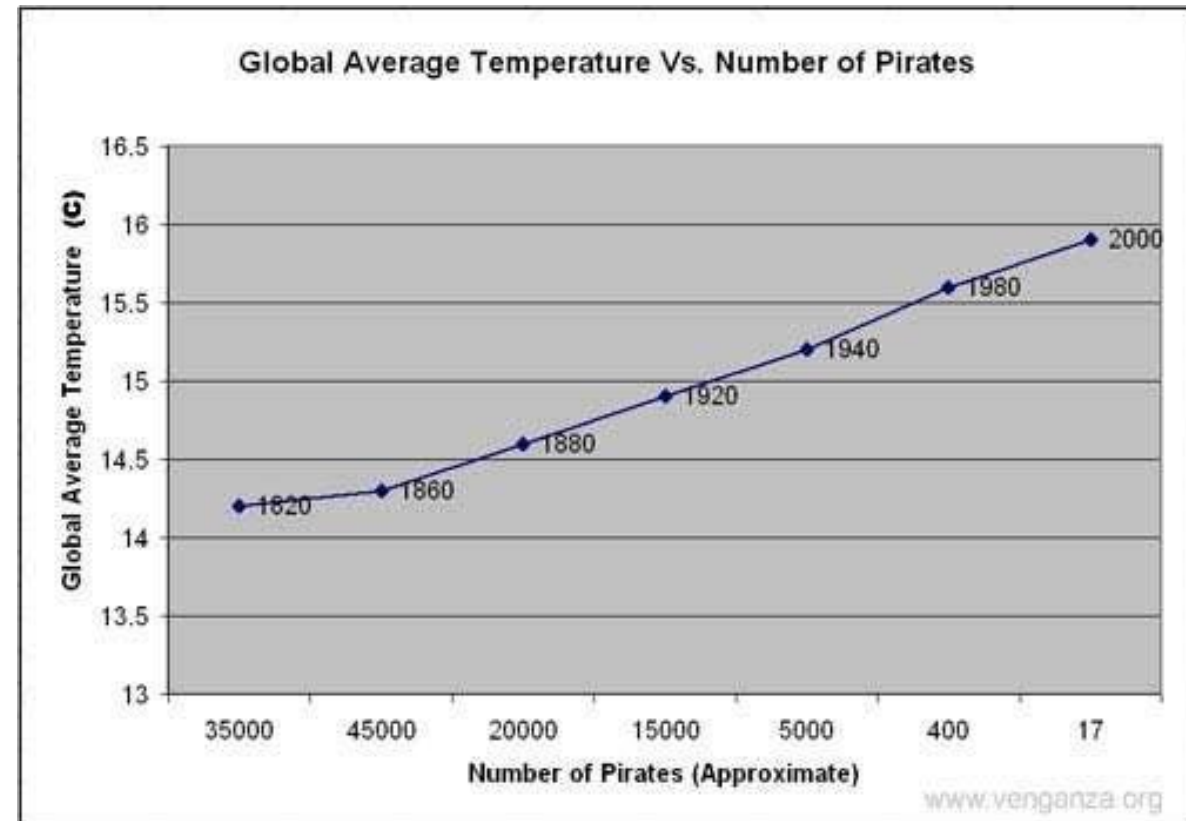
- Retrospective – A study that looks into the past
- Prospective – A study that deals with current data
- Longitudinal– A study that follows the same individuals over a long period of time

▶ Types of studies

- Cross sectional study – Surveys that collect data at one point in time (often prospective)
- Case–Control– Often a retrospective study for rare conditions
- Cohort Study– A Longitudinal Study following a group of similar individuals

Observational Study Examples

- ▶ Most (if not all) disease, addiction, or “bad outcome” study.
 - Cannot assign people to the treatment and control group randomly.
 - May not be ethical to control.
- ▶ Some interesting findings.
 - Number of TVs a person owns and life expectancy.
 - Cause of Global Warming...



Observational Studies

▶ Advantages

- It can detect associations between variables where the values of the variable have already been determined.
- Can be used to study variables that are impossible or unethical to control

▶ Disadvantages

- It cannot isolate causes to determine causations.

Confounding (Lurking) Variables

Confounding Variable

A variable that has not been accounted for but which is causing a difference in the groups being studied.

From Previous Example:

Higher income would lead to the ability to buy more TVs and most likely better healthcare. Therefore this might lead to longer life expectancy.

Designed (Controlled) Experiments

- ▶ Experiments deliberately impose a treatment on individuals and record their responses
- ▶ They compare the response to a given treatment versus:
 - the absence of treatment (control)
 - a **placebo**
 - another treatment
- ▶ Goals of Experimental Design:
 - Replication
 - Randomization
 - Control of Error
- ▶ We want to ensure that the treatment group and the control group are as similar as possible.

Terminology

- ▶ Factors – Treatment(s) that we may be interested in
 - Levels – Specific values of factors
 - Treatment combination (Interactions) – With multiple factors, a specific combination of levels.
- ▶ Blinding– Independent party assigns subjects to treatment group
 - Double-blinding: Researchers do not know who is in the treatment group.
- ▶ Placebo – Harmless pill (or look alike) given in place of actual treatment
 - Placebo Effect – Reacting to a treatment after simply from being told you are receiving the treatment.

Designed Experiment Examples

- ▶ We have a manufacturing operation for carbon steel that depends on three factors. We want to test at the low and high end of each factor to examine the relationships.

Factor	Low	High
Steel Temperature (F)	120	150
Pressure (ATM)	2	3
Flow Rate (GPM)	100	150

- ▶ Suppose I want to test all treatment combinations. How many would I have?
 - $2*2*2=8$
 - This is called a 2^3 factorial experiment


Reboil Temp	Press.	Flow Rate
120	2	100
150	2	100
120	3	100
150	3	100
120	2	150
150	2	150
120	3	150
150	3	150

Controlled Experiments

▶ Advantages

- Can analyze causal relationships between variables and outcome because the researcher is able to control variables that influence the response.


▶ Disadvantages

- Cannot be done when the variables cannot be controlled.
 - Cannot apply in some studies for moral or ethical reasons.
- 

Common Experiment Designs

- ▶ Completely randomized–
- ▶ Block Design – Groups (or blocks) are defined by predetermined factors who are then randomized to treatments
- ▶ Matched Pairs – Finding a similar individual or unit to compare outcomes
 - Twins studies
 - Before–after, pre–post
 - Crossover
- ▶ Repeated Measures – Having individuals go through a single treatment more than once
- ▶ Experimental Designs
 - <https://www.youtube.com/watch?v=10ikXret7Lk>

Data Collection Method Summary

- ▶ Retrospective Study
 - Cheap and easy
 - No control over data collection
 - ▶ Observational Study
 - No control over environment.
 - Can show associations, but effected by confounding variables.
 - ▶ Controlled Experiment
 - Can show causality if done correctly (Replication, Randomization, Control of Error)
 - Cannot study certain things
- 

Populations and Samples

- ▶ Population – consists of all subjects or items of interest. It is the group being studied.
 - Parameter – A numerical measurement describing a Population
 - Census– the collection of data from every member of a population
- ▶ Sample – a group selected from the population (a subset of the population). It provides information used to infer information about populations.
 - Statistic– A numerical measurement describing a sample

Choosing a Sample to Represent a Population

- ▶ Survey Sampling and methodology studies the sampling of individual units from a population and associated sampling design techniques (e.g. questionnaire construction).
- ▶ Goals:
 - To produce data in a way that is designed to answer our questions.
 - Individuals in sample are **representative** of the population (that is, provide accurate information about the population)
 - Minimize:
 - Cost of obtaining the sample (money, time, personnel, etc.)
 - Bias (two types)

Bias

A survey method is biased if it has a tendency to produce an untrue value.

A couple types we will talk about:

- 1.) Measurement bias
- 2.) Sampling bias

Measurement Bias

Results from asking questions that do not produce a true answer. Measurement bias can occur in a variety of situations including:

- ▶ Self-reporting of personal data (Anecdotal Evidence)
- ▶ Wording effects:
 - Confusing, leading, or non-neutral wording
- ▶ Missing data, precision of numbers, percentages, scales used.

Sampling Bias

Occurs when a sample is used that is not representative of the population. Some causes could be:

- ▶ Voluntary response samples – Samples where people decide whether to respond. Ex– Internet polls
- ▶ Convenience sampling – only sampling individuals that are convenient
- ▶ Incentivized sampling
- ▶ Other issues:
 - Small samples
 - Non–responses

Surveys Face Growing Difficulty Reaching, Persuading Potential Respondents

	1997	2000	2003	2006	2009	2012
	%	%	%	%	%	%
Contact rate (percent of households in which an adult was reached)	90	77	79	73	72	62
Cooperation rate (percent of households contacted that yielded an interview)	43	40	34	31	21	14
Response rate (percent of households sampled that yielded an interview)	36	28	25	21	15	9

PEW RESEARCH CENTER 2012 Methodology Study. Rates computed according to American Association for Public Opinion Research (AAPOR) standard definitions for CON2, COOP3 and RR3. Rates are typical for surveys conducted in each year.

Simple Random Sampling (SRS)

▶ Idea:

- Every possible sample of size n out of a population of N individuals has an equally likely chance of being selected.
- Each individual in the population has the same chance of being chosen for the sample.
- Subjects are selected without replacement (subjects cannot be selected twice).

▶ Consider a simple random sample of size $n = 2$ from a population of $N = 4$.

- Population: {A, B, C, D}
- Possible samples: [AB, AC, AD, BC, BD, CD]

Other Sampling techniques

- ▶ **Systematic (Probability) sampling** –uses chance to select a sample, based on known selection probabilities.
- ▶ **Stratified sampling**– dividing a population into subgroups and then sampling equally from those subgroups.
- ▶ **Cluster Sampling** – Dividing a population into clusters and then randomly selecting all individuals from them.
- ▶ **Sampling Video**
 - <https://www.youtube.com/watch?v=QOxXy-l6ogs>

Context is Key

- ▶ The first time you see a data set, ask...
 - Who or what was observed?
 - What variables were measured?
 - How were the variables measured?
 - What are the units of measurement?
 - Who collected the data?
 - How were the data collected?
 - Why did they collect the data?
 - When were the data collected?