Topic: Exploratory Data Analysis (EDA)

The Statistical Process

School of Mathematics and Applied Statistics

UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Statistics involves . . .

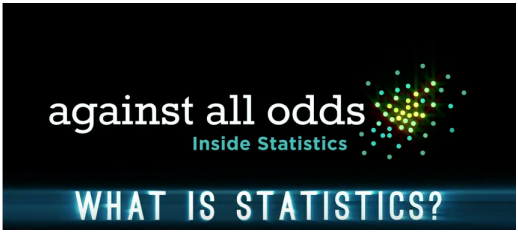
- collecting data about real life processes;
- presenting and describing data;
- formulating models which allow for chance variation;
- fitting models to data, checking assumptions, and making predictions;
- making decisions in the presence of uncertainty.

Probability theory provides the foundation for all of the above.

What is Statistics?

Statistics is the art and science of making sense of it all!

Video: 6.23mins https://www.youtube.com/watch?v=wG8L_C2OMu8



Notes

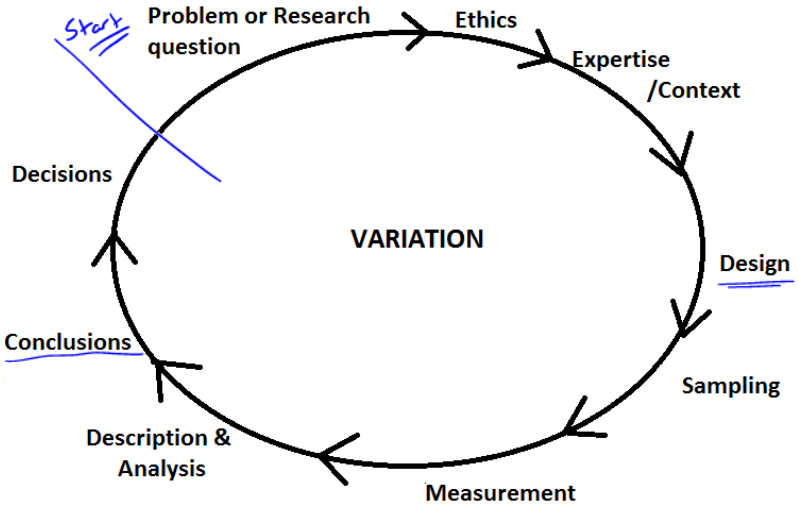
Statistics is . . . . . ?

Statistics is a study of variability in the world around us

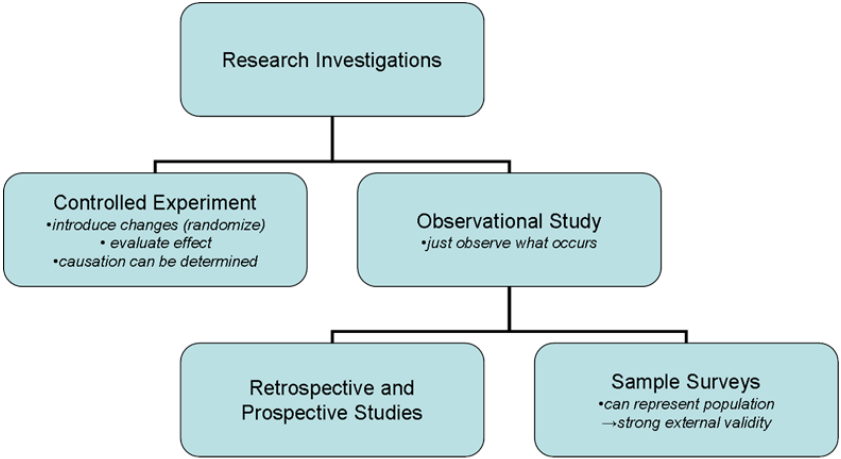
- What **brands** of mobile phone are most popular, least popular?
- How much **time** does the typical uni student spend on FaceBook per day?
- How many **texts** does the typical uni student send per day?
- What proportion of **emails** are spam?

If all things were constant we would not be studying them.

Statistical Process



Research Designs



Issue: Quality of Evidence

Data Collection

- **Controlled experiments**
  - Investigator introduces a change into a process
  - Observes & evaluates effect of changes (variation)
  - Evidence of causation if factors properly controlled and/or association
  - Gold standard for evidence
- **Observational studies**
  - Investigator does not interfere with the process
  - Observes variation
  - Evidence of association
  - Weaker form of evidence
- **Simulations** 

Experimental Design

Simple between group design

- Randomly assign subjects to two groups
- Experimental and Control

  - Apply the treatment
  - Compare the outcomes
  - Determine the impact of the treatment

Many different designs (some better than others)

Observe and evaluate the effects of the **changes**

Controlled Experiments - Example

Video Plant Experiment Mark Drollinger (2:34min)

https://www.youtube.com/watch?v=VhZyXmgIFAo

My Notes:

-
-
-

Observational studies

Retrospective

- Look backwards in time
- eg current lung cancer patients
  identify habits in common

Prospective

- Select a group and follow them forward in time
- eg select and see who develops lung cancer

Observational Studies cont.

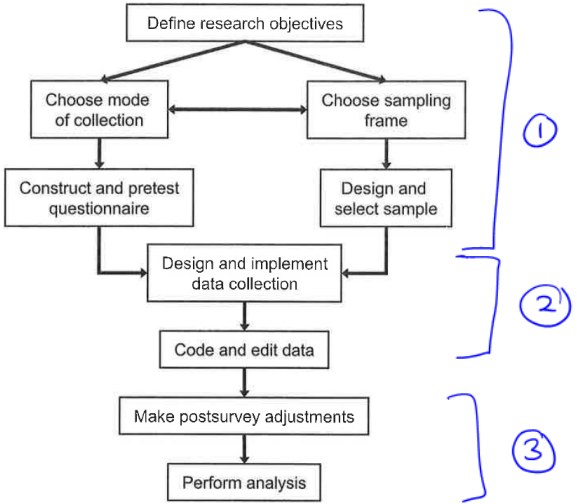
Sample surveys - common type of observational study

- represent population if sample well selected
- allow analysis of relationships for many groups in population
- strong external validity (generalisability)
- problems with internal validity
  - lack of control of other factors

Three Phases in the Sample Survey Process

1. Development - clear aims, design of questions, pilot testing etc
2. Operational - data collection, coding, editing
3. Analysis - exploratory, inference

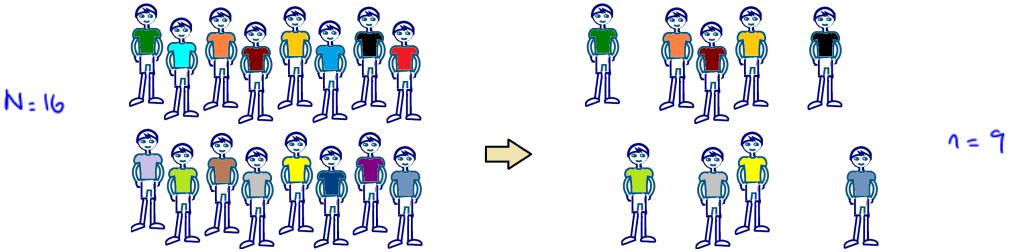
Sample Survey Process



Population vs Sample

A **population** consists of a group of units about which you wish to draw conclusions

- Eg people, businesses, hospitals, events etc.



A **sample** is a representative subset of a population

- used to draw conclusions about the population
- advantages of simplicity, cost reduction and timeliness

Populations

Defining the population of interest

- Definition of units
- Scope
- Geographic coverage
- Reference period

Eg: Who comprises the population of Doctors in Illawarra?

- Define what type of doctor?
- Do they need to live in the Illawarra or work in the Illawarra?
- Individual doctors or practices?
- What period - financial year?
- Does working part-year meet the definition?

Sampling Frame

The Sampling Frame

- Is a list of units in the population
  Eg White pages, Electoral roll, Uni admin list of students
- Need to know how to access them
- Often lists do not correspond exactly to the population of interest
  Eg Members of AMA versus all doctors; some students have withdrawn

Problems with lists

- omissions
- duplicates
- incorrect information, out of date etc

Survey Modes

Survey modes:

- Mail
- Telephone
- Field interview
- Internet

Sampling Designs

Video Statistics Learning Centre (4:53mins) http://www.youtube.com/watch?v=be9e-Q-jC-0

My notes:

- Simple Random Sampling
- Convenience Sampling
- Systematic Sampling
- Cluster Sampling
- Stratified Sampling

In Summary

- **Statistics** is the art and science of making sense of the variability we encounter in everyday life.
- The **Statistical Process** may include methods of data collection such as
  - Controlled Experiments
  - Observational Studies
  - Simulation
- Statistics uses sample data to make inference about populations using probability theory.