# Statistical Methods

## Lecture 1 – Introduction
## [Ros17, Chp. 1]

**Luke Dickens**

**Autumn 2020**

# Data, Populations and Samples

## Brief History and Python

# Data Collection

Modern approach to learning about complex question: collect data

- You may begin with data...

  Statistics then used to describe, summarize and analyse this
- Otherwise, you must collect the data yourself
- Data collection can be subject to biases
- Good experimental design attempts to neutralise bias

  e.g. Placebo and double blind trials

# The Population and the Sample

There are two main types of data sets:

- **Population** comprises all the elements of a set of data. Each element is represented once. It represents the entire domain of interest.
- **Sample** comprises one or more observations drawn (at random) from the population. The nature of a sample depends on the sampling procedure (more later).

For the school age experiment in [Ros17, Sec. 1.1], they used census data from 1960 and 1980. What was the population of interest?

# The Population and the Sample

There are two main types of data sets:

- **Population** comprises all the elements of a set of data. Each element is represented once. It represents the entire domain of interest.
- **Sample** comprises one or more observations drawn (at random) from the population. The nature of a sample depends on the sampling procedure (more later).

For the school age experiment in [Ros17, Sec. 1.1], they used census data from 1960 and 1980. What was the population of interest?

Why was this better than the previous experiment based on tests at the end of a child's first year?

# Descriptive versus Inferential Statistics

We differentiate between two types of statistics:

- **Descriptive Statistics** – procedures used to summarize and describe characteristics of a set of measurements.
- **Inferential Statistics** – procedures used to draw conclusions from data

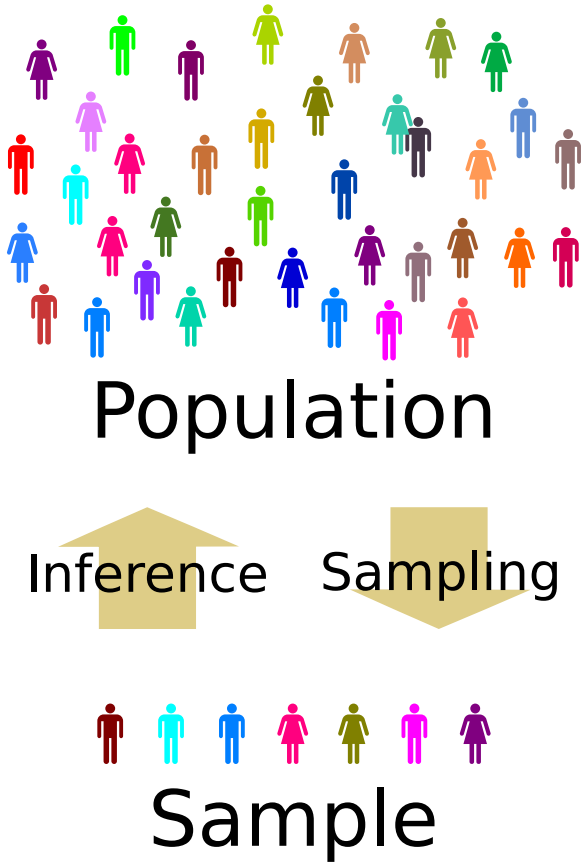
Inferential statistics makes inferences (draws conclusions, makes predictions, makes decisions) about the characteristics of a population from the sample.

This must account for randomness/chance in the observed data (sample).

# Inferential Statistics in a Nutshell

A general approach to inference (described in [MBB12]):

1. Specify question(s) and identify population
2. Decide how to select the sample
3. Collect the sample and analyse
4. Make an inference about the population (using information from Step 3)
5. Determine the reliability of the inference




*Man/woman images source from here*

Usually requires you to make assumptions about variability/ randomness in observations. Amounts to a **probability model**.

# Sampling and Representativeness

Samples are subgroups of populations to be studied in detail:

- to meaningully inform us about the population they must be **representative**

  – seeks to reflect the characteristics of the larger group
- biased samples are **unrepresentative**
- a sample cannot be representative unless chosen randomly

  . . . but what does that mean?
- stratified sampling - controls the representation of subpopulations

## Data, Populations and Samples

# Brief History and Python

# Statistics and John Graunt

The textbook describes the beginnings of data collection and analysis. John Graunt:

- analysed burials and plague deaths
- estimated the population of London
- created mortality table from "bills of mortality"

We'll look at these in more detail.

# Python

To look at these data, we'll use Python and Jupyter notebooks. To work with these you will need to:

- install Python $>$3.6
- familiarise yourself with basics
- install Jupyter
- practice with the module exercises

More detail on getting started can be found on the Moodle page.

# To the notebook...

*Book photo found here*

# References I

[MBB12] William Mendenhall, Robert Beaver, and Barbara Beaver, *Introduction to Probability and Statistics*, 14th ed., Duxbury Press, 2012.

[Ros17] Sheldon M. Ross, *Introductory Statistics*, 4 ed., Academic Press, 2017.