Introduction

Critical thinking with data

MATH1062

Table of contents

1 What is a statistician?

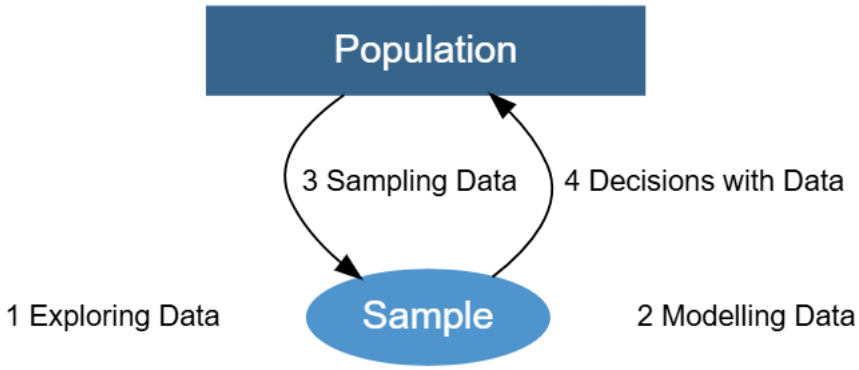
- A primary objective of the statistician is to answer a research question. Most of the time, the research question concerns with activities/behaviours/phenomena in a targeted **population**.
- Measurements with respect to the population is always very difficult to obtain, so statisticians aim to answer the research question using a **sample**.
- A majority of the statistics literature looks at how to 'best' infer population characteristics from a sample.

Example: what is the proportion of Australians who are taller than 180 cm?

- **Solution 1:** Survey all Australians (the population) and calculate the proportion.
- **Solution 2:** Measure the heights of a subset of the population. Build a model based on the observed heights and then answer the question using the model.

2 Course overview

- Fundamental statistical concepts.
- A number of useful statistical models.
- Probability and sampling.
- Decision with data.
- The R computing language – for all computational aspects and report written in the course.

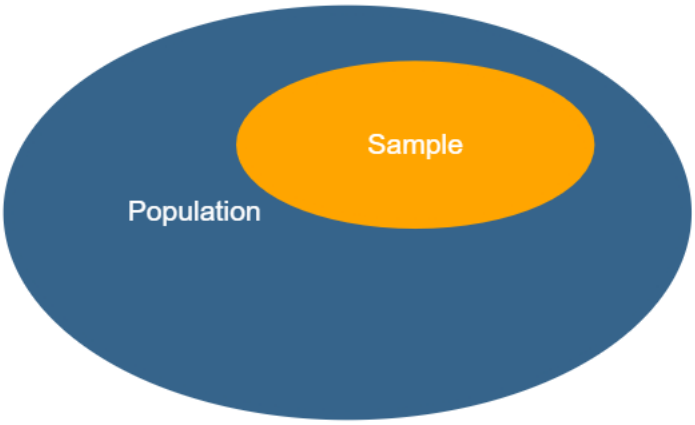


3 Plan for lectures 1 – 4

- Critical thinking with data
- Initial data analysis
- Graphical summaries
- Numerical summaries

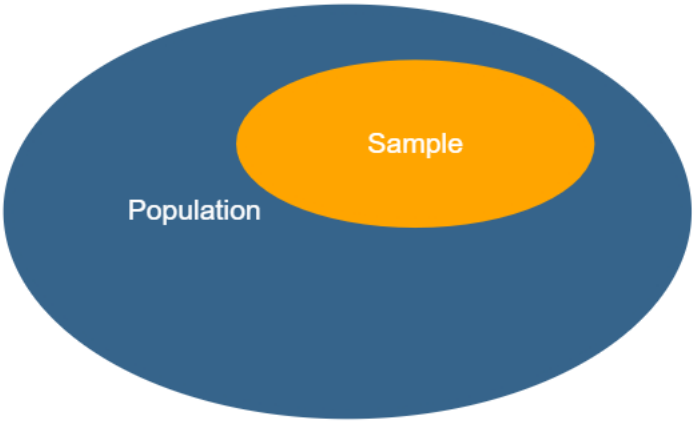
3.1 Population and sample

- The target population comprises all relevant subjects of interest.
- The sample is a manageable subset, selected to make the study feasible.



4 Population and sample

- A sample is a subset of the population.
- It should be representative of the target population (not **biased**).
- Large enough to give accurate information about the population.
- Ideally, the observations should be independent of each other.



5 Sources of bias

Bias may be defined as any systematic error (ie. not occurring randomly) which results in incorrect conclusions about the target population. Some types of bias include

- Selection bias
- Measurement bias
- Response bias
- Confounding

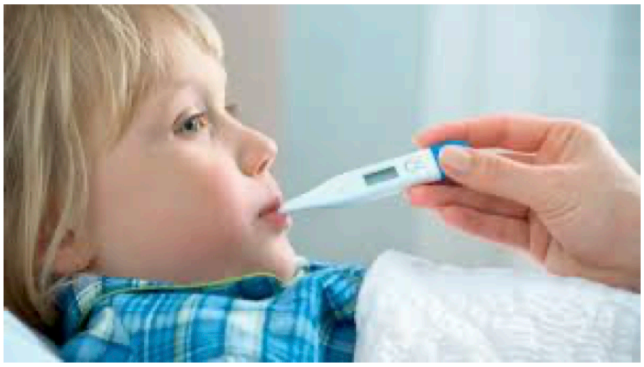
5.1 Selection bias

Selection bias refers to any systematic differences occurring in the way that subjects are selected for a study.

> **Note**
>
> E.g. In a height study, we accidentally selected a group of basketball players.

5.2 Measurement bias

Measurement bias refers to systematic differences in the measurement of variables.



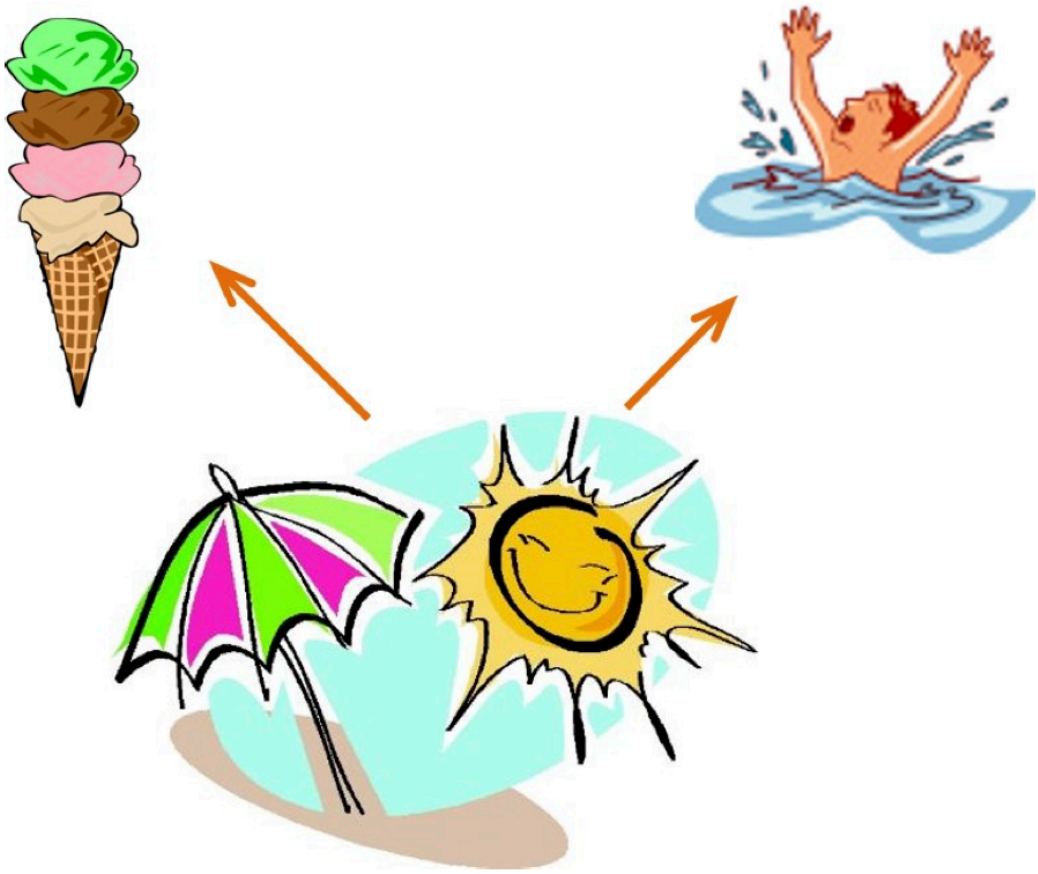
> **Note**
>
> E.g. In a human body temperature study, an in-ear thermometer is consistently higher than that of an oral thermometer.

5.3 Response bias

- Response bias can occur when the response rate to a survey is too low.
- This is because those who respond to a survey often have different characteristics or attitudes than those who don't respond.
- This is most common when sensitive issues are involved.

5.4 Confounding

Confounding occurs when the effect of one variable (X) on another variable (Y) is clouded by the influence of another variable (Z).



E.g., X: increase in ice cream sales, Y: increase in drowning incidents; Z: sunny and warm weather

> **!** Important
>
> Confounders can be hard to find, and can mislead about a cause and effect relationship.

6 Study design

There are two main types of study designs

An **observational study** is one in which there is no treatment imposed by the investigator.

- We simply observe.
- Data are observed and recorded based on responses from subjects.

An **experimental study** is one in which the investigator has some control over the subjects by giving some kind of treatment.

- Explanatory variable (determinant) is perturbed, behaviour of dependent variable (response) is noted.
- Data are observed and recorded based on responses from subjects.

6.1 Study design

Conclusions of a study depends on the design. Roughly speaking:

Observational studies allow us to infer **association**.

- We should be very careful about the implications: **association** is not **causation**
- Confounding variables are always a possible cause of ridiculous conclusions.
- E.g. 3 glasses of water can cure flu. Problem: I take my flu medicine with a glass of water, 3 times day.

Experimental studies allow us to infer causation.

- Usually more informative on the underlying mechanisms, since the researchers can eliminate external factors in their experiments.
- One must be very familiar with experimental designs and take into account of all sources of variations.
- E.g. Crops growth is better in soil with high nitrogen and locations with good sunlight. If an experiment only contains high nitrogen soil and good sunlight, then it is not possible to separate out these two effects.

6.2 Example: does smoking cause liver cancer?

A study finds that smokers tend to have higher rates of liver cancer. Does smoking cause liver cancer?

Study: Smoking Dramatically Increases Liver Cancer Risk

Article date: November 2, 2011

By Stacy Simon

A new study reinforces the link between liver cancer and the risk factors of smoking, obesity, and heavy drinking.

Researchers from the US and Europe studied 125 liver cancer patients to determine what risk factors were contributing to their disease. They compared them to 229 people without cancer who were matched by age, gender and other factors. The participants were all part of a European study group that was formed so researchers could investigate the role of biological, dietary, lifestyle and environmental factors in the development of cancer and other chronic diseases.

They found that almost half the cases of liver cancer in the study were associated with smoking, 16% were associated with obesity and 10% were associated with heavy alcohol consumption. Almost 21% of cases were associated with hepatitis C and 13% with hepatitis B.

The most common type of liver cancer, hepatocellular carcinoma, is a leading cause of cancer deaths worldwide. In many sub-Saharan African and Southeast Asian countries, it is the most common type of cancer. It's less common in the United States. Worldwide, the major risk factors for liver cancer are long-term infections with hepatitis B virus and hepatitis C virus. People with these infections are more likely to develop cirrhosis, a disease in which liver cells become damaged and are replaced by scar tissue. People with cirrhosis have an increased risk of liver cancer. In the US, most liver cancer is associated with alcohol-related cirrhosis and possibly non-alcoholic fatty liver disease.

In an accompanying editorial, Morris Sherman, MD and Josep M. Llovet, MD clarify that smoking by itself does not cause liver cancer, but that it dramatically increases the risk, especially for people who have other risk factors, such as hepatitis B or C virus.

They conclude, "We should be counseling our patients who have other risk factors for hepatocellular carcinoma to quit smoking."

7 Association and causation

It is rather easy to establish association (that one thing is linked to another).

- Association may **suggest** causation. But association does not **prove** causation.
- We need to take **confounding** variables into account. They can mislead about a cause and effect relationship.

What could explain the fact that smokers have a higher rate of liver cancer?

- Smokers tend to drink more alcohol than non-smokers, and excessive alcohol consumption causes liver cancer.
- So the effect of smoking is confounded (mixed-up) with the effect of alcohol consumption.
- Here alcohol consumption is a confounding factor.

7.1 Strategy for establishing causation

- If a confounder is known, we can potentially add it as an additional variable. For example, if alcohol consumption is a potential confounding factor for smoking's effect on liver cancer, we can add an variable "drinking" with possible values:
  - "heavy drinker", "medium drinker", "light drinker", "lifetime abstainer"
- In some scenarios, for example, in clinical trials, it is possible to design **controlled experimental studies** by manipulating variables to test their effects.

8 Extra Example

8.1 Investigating Speed Limit

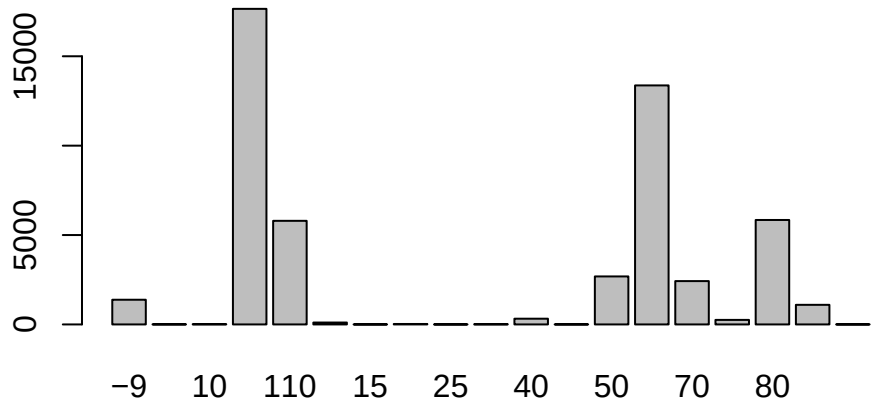
Speed Limit could be considered as a discrete, quantitative variable, but here it might be better to classify it as an (ordinal) qualitative variable. Why?

```
data = read.csv("data/2020fatalities.csv", header=TRUE)
SpeedLimit = data$Speed.Limit
Dayweek = data$Dayweek
Gender = data$Gender
table(SpeedLimit)
```

```
SpeedLimit
         -9         <40         10        100        110        130
       1381          6         17      17653       5796        106
         15         20         25         30         40          5
          1         26          2         13        321          3
         50         60         70         75         80         90
       2685      13371       2424        254       5843       1097
Unspecified
          2
```

8.2 Simple Bar Plot

```
barplot(table(SpeedLimit))
```



8.3 Statistical Thinking

What is curious about this data? Why?

- -9 indicates a missing value. Why speed category of <40 **and** speed limits of 10,15,20,25,30? We could 'clean' the data.

What was the most common speed at which a road fatality occurred? How might this affect public policy?

- Notice how many fatalities happen at high speeds (100km/hr+).
- Can we directly deduce from that lives can be saved by introducing stricter speed limits?

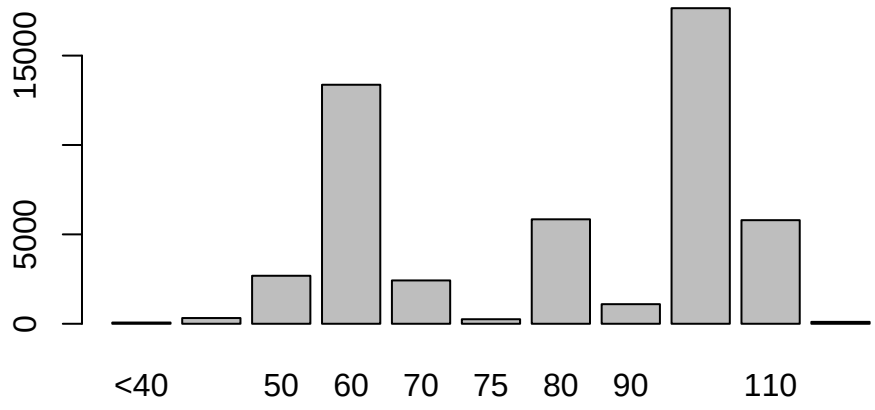
8.4 Clean data

```
SpeedLimit[SpeedLimit=="5"] = "<40"
SpeedLimit[SpeedLimit=="10"] = "<40"
SpeedLimit[SpeedLimit=="15"] = "<40"
SpeedLimit[SpeedLimit=="20"] = "<40"
SpeedLimit[SpeedLimit=="25"] = "<40"
SpeedLimit[SpeedLimit=="30"] = "<40"
SpeedLimit[SpeedLimit==-9] = NA

SpeedLimit = factor(SpeedLimit, levels=c("<40","40", "50", "60", "70", "75",
                                         "80", "90","100","110","130"))
```

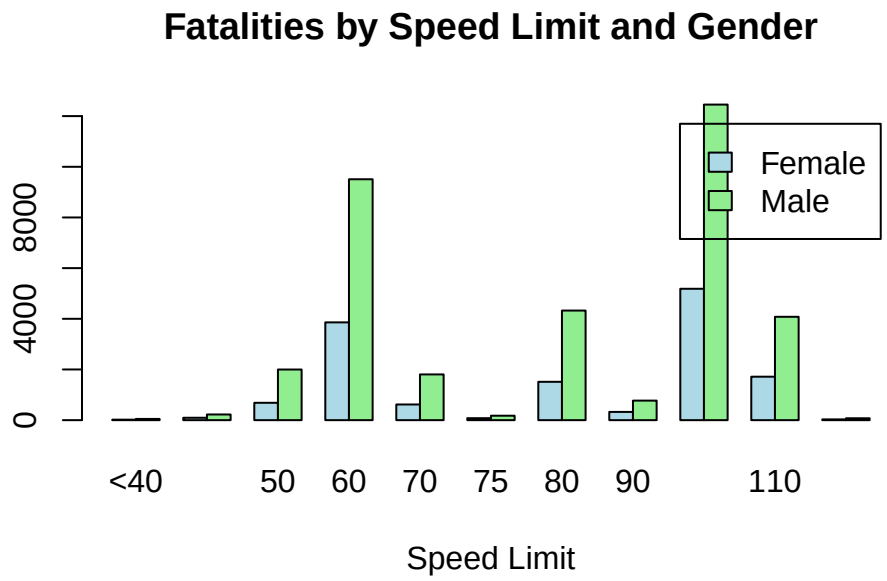
8.5 New barplot of SpeedLimit

```
barplot(table(SpeedLimit))
```



8.6 Double Bar Plot

```
#Gender = data$Gender
data2 = table(Gender,SpeedLimit)
barplot(data2[-c(1,4),], main="Fatalities by Speed Limit and Gender",
  xlab="Speed Limit", col=c("lightblue","lightgreen"),
    legend = rownames(data2[-c(1,4),]), beside=TRUE)
```



8.7 Statistical Thinking

Are there any interesting patterns?