TABLE OF CONTENTS

Topic 1: Collecting data

Two types of variables in data sets:

- A **categorical variable** divides the cases into groups, placing each case into exactly one of two or more categories. The answer is one of three or a few categories.
- A **quantitative variable** measures or records a numerical quantity for each case. Numerical operations like adding and averaging make sense for quantitative variables.
- Categorical: defines groups
  - Eg, gender, year
- Quantitative: numerical measure
  - Eg, height, pulse, age
- We may use numbers to code the categories of a categorical variable, but this does not make the variable quantitative unless the numbers have a quantitative meaning.

The relationships between variables

- A variable is any characteristic that is recorded for each case.
- An **explanatory variable** is a type of independent variable. The two terms are often used interchangeably. But there is a subtle difference between the two. When a variable is independent, it is not affected at all by any other variables. When a variable isn't independent for certain, it's an explanatory variable.
- The **response variable** is the focus of a question in a study or experiment. An explanatory variable is one that explains changes in that variable. It can be anything that might affect the response variable
- The explanatory variable is used to understand or predict the values of another, the response variable
  - For example: does meditation help reduce stress?
    does sugar consumption increase hyperactivity?
    does the interest rate affect the exchange rate?

Key concepts

A <u>population</u> includes <u>all</u> the individuals or objects of interest

- A sample is being selected from a higher dimension, ie. The population
- T*he population is the source of the sample*

A sample from a population

Data are collected from a **sample**, which is a subset of the population.

- A sample consists of the cases <u>selected</u> into a dataset; a sample is a subset of the population
- The process of using a **sample** to gain information (to help understand more)about the **population** is called *inference.*
  - (Since we rarely have data on the entire population, a key question is how to use the information in a sample to make reliable statements about the population. This is called statistical inference.)
- **A sample is a representation of the population**
- If the sample is an acceptable representation on the population, then an inference can be carried out to get a perfect estimate
  - It would be expected to reflect similar characteristics to the population

Sample bias

- **Sampling bias** occurs when the method of selecting a sample causes the sample to differ from the population in some relevant way.

- To avoid sampling bias, we try to obtain a sample that is *representative* of the population. A representative sample resembles the population, only in smaller numbers.
- To avoid sampling bias, a **random** sample needs to be taken out
- The more representative a sample is, the more valuable the sample is for making inferences about the population.
  When choosing a **simple random sample** of *n* units, all groups of size *n* in the population have the same chance of becoming the sample. As a result, in a simple random sample, each unit of the population has an equal chance of being selected, regardless of the other units chosen for the sample.
- **Bias** exists when the method of collecting data causes the sample data to inaccurately reflect the population.
- The way questions are worded can also bias the results.
- not all methods of data collection lead to valid inferences.

Data collected to analyse a relationship can come from an experimental or observational study

Experimental

A researcher controls which case has what explanatory variable to come to a causal conclusion.

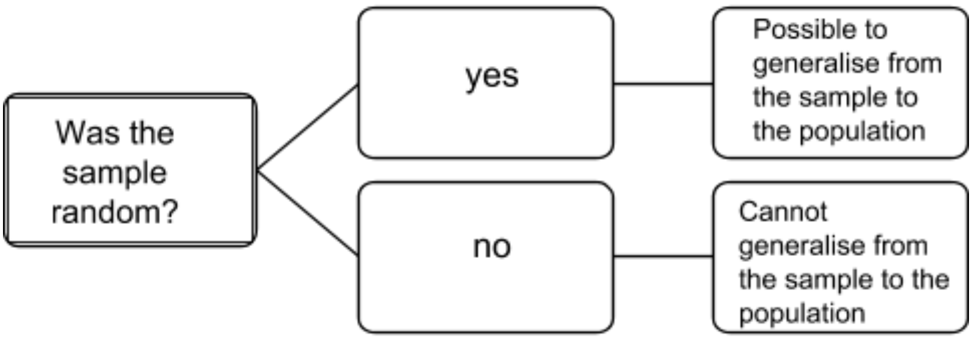
- The handling of different treatment groups in an experiment should be as similar as possible, with the use of blinding and/or a placebo treatment.
- The only want to infer a **causal relationship** between variables statistically is through data obtained from a randomized experiment.

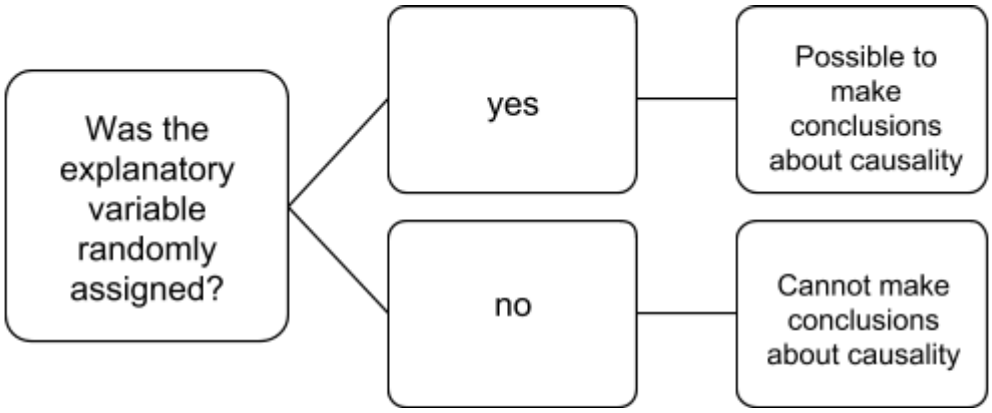
Observational

The researcher does not control the explanatory variable, they find people with the explanatory variables and observe them. You can not make a causal conclusion (a conclusion on causation) with observation.

- In an observational study, we need to be wary of confounding variables. A randomized experiment allows us to avoid confounding variables by actively manipulating the explanatory variables

<u>Two questions on collecting data</u>





Topic 2: describing data

Two variables are **associated** if the values of one variable tend to be related to values of the other variable (relationship)

- E.g. Spurious relationship (not valid relationship)
- **Association does not imply causation/ a causal relationship**

Two variables are **causally associated** if changing the value of the explanatory variable influences the value of the response variable (Nature of the relationship)

<u>Spurious relationships</u>
A spurious relationship can often result from a confounding variable

- A **third variable** associated with both the explanatory and the response variable is called a **confounding** variable.
- A confounding variable may offer a plausible explanation for an association between explanatory and response variables
- Causal association cannot be determined when confounding variables are present.

<u>Experiments</u>
In an experiment, the researcher **actively controls one or more of the variables**

- Experiments can be used to eliminate confounding variables
- **Experiments may be used to establish causation because they eliminate the confounding variables**

<u>Observational study</u>
In an observational study, the researcher **does not actively control the value of any variable, but simply observes them.**

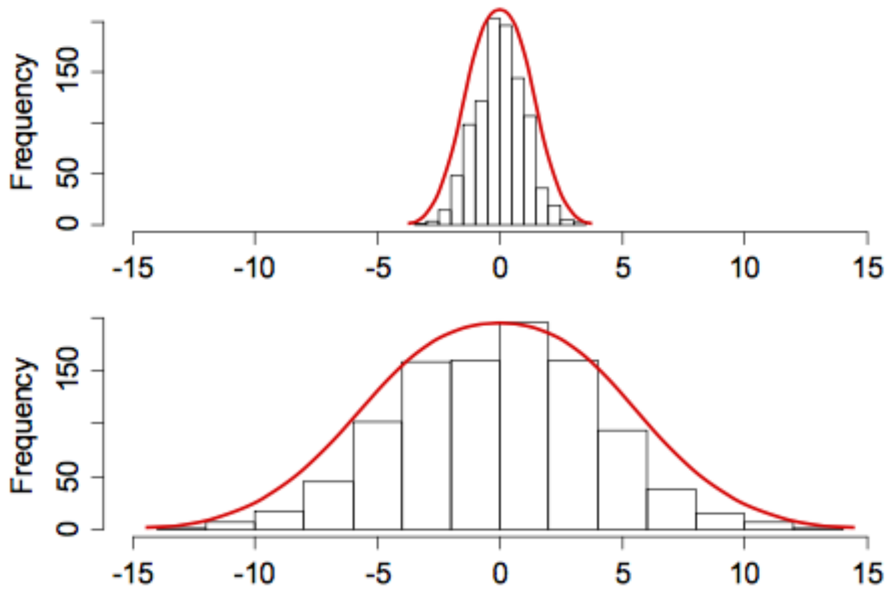
- Observational studies almost **always have confounding variables**
- **Observational studies can almost never be used to establish causation**

How to eliminate confounding variables…

- By **randomly** assigning values of the explanatory variable
    - A process referred to as **randomization**
- In an experiment, the researcher controls the assignment of one or more variables. This power can allow the researcher to avoid confounding variables and identify causal relationships, if used correctly.

In a **randomized experiment**, the explanatory variable for each unit is determined *randomly*, before the response variable is measured.

- Different levels of the explanatory variable are known as **treatments**



- A bell curve is symmetric
- The mean (average) is always in the centre of a bell curve
- A bell curve has only one mode/ peak
    - One peak is *unimodal*
    - Two peaks is *bimodal*
- The larger the standard deviation, the more spread out the curve
- A bell curve follows the 68-95-99.7 rule (which provides a convenient way to carry out estimated calculations)
        - (The z score)
    - Approx. 68% of all the data lies within one standard deviation
    - Approx. 95% of all the data is within two standard deviations of the mean
    - Approx. 99.7% of the data is within three standard deviations of the mean

Not all normal distributions are bell shaped

Topic 3: Sampling distribution & Confidence intervals

We estimate a *population parameter* using a *sample statistic*
A statistic measures an attribute/characteristic of a sample

1. Average
2. Spread
3. Location
4. association

Notation
n is the number of cases in the sample or the **sample size**
  Eg. 134 movies-> n =134
x represents a variable (e.g., world gross)
  $x1$, $x2$, $x3$, . . . , xn represents the individual values of x
  $x1$ = 97.009, $x2$ = 201.897, $x3$ = 216.196, . . .
Average (measure of centre)
- common measures of entre are the **mean** and the **median**

sample mean

sigma = "the sum of . . ."

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

population mean

"mu"

$$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$$

population size

Lower case n (n)= sample
Upper case n (N)=population

The **median (m)** is the middle value when the data is ordered

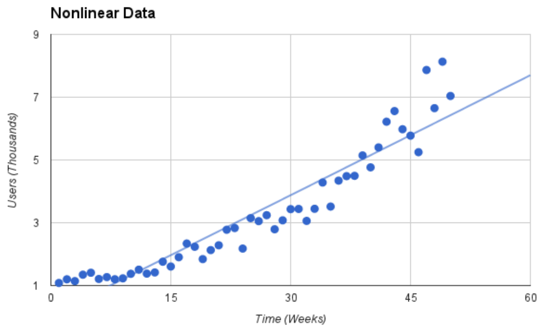
- The sample median splits the data in half
- To find the sample median of a variable x
    1. Order x from highest to lowest values
    2. Pick the middle value
    3. [if n Is even, the median is the mean of the two middle values]

An **outlier** is a value that is notable distinct from other values in a dataset
A statistic is **resistant** or **robust** if It is relatively <u>unaffected by outliers</u>

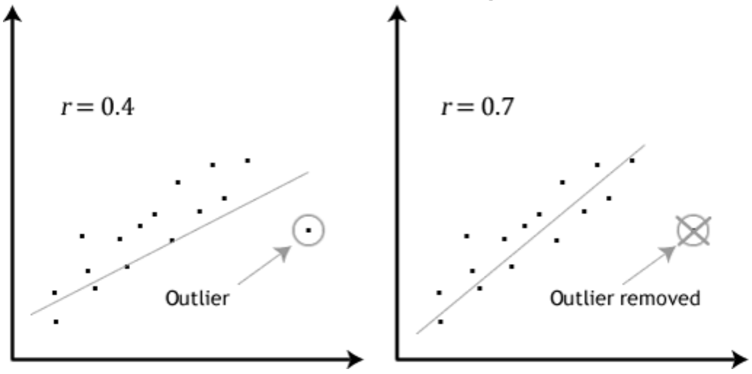
- The median is resistant (not affected by outliers); the mean is not (affected)

Spread is the measure of dispersion
The **standard deviation** of a <u>quantitative variable</u> measures the spread of the data

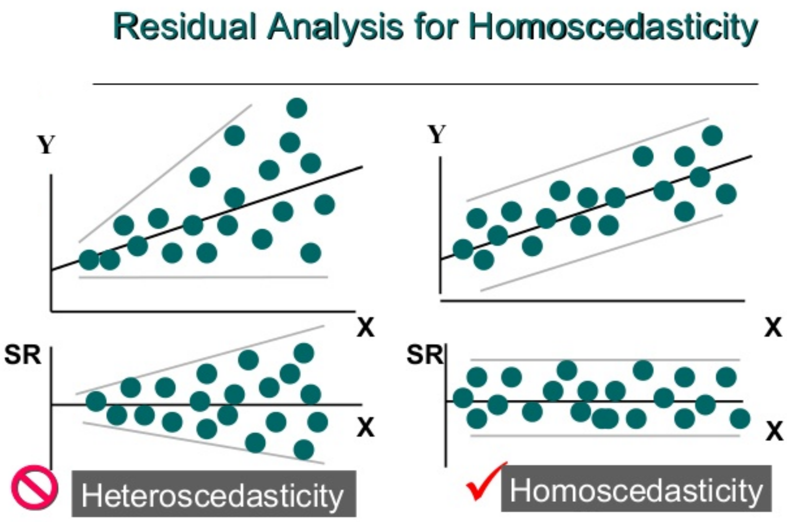
- Measures the distance of a 'typical case' from the mean
- The *larger* the standard deviation
    - The more spread out the data
    - The more variability in
    - the data



- If the data includes an outlier that affects the regression line:



- If the error terms are heteroscedastic instead of homoscedastic. I.e. the error terms get larger as *x* gets larger or as *x* gets smaller.

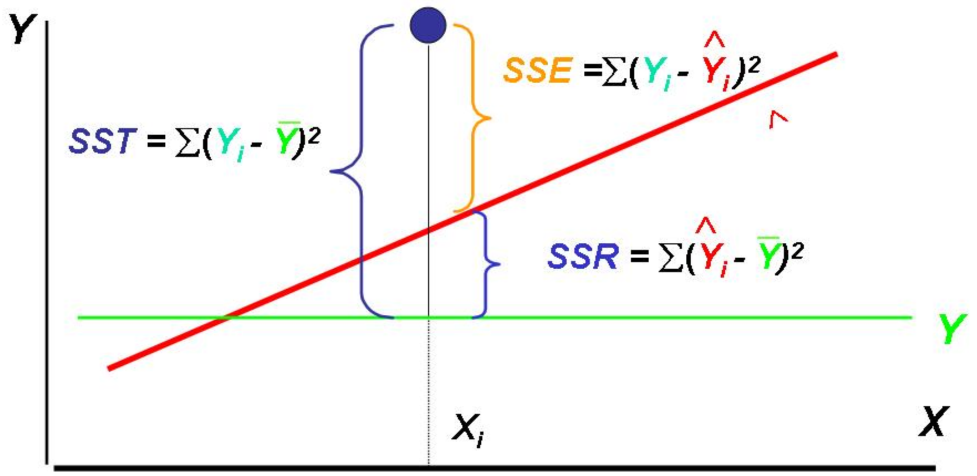


The variance in a regression mode can be broken down as follows:

$$SST = SSE + SSR$$

- Total variability in *y*: SST = $\sum_{i=1}^{n}(y_i - \overline{y})^2$
- Variability in *y* not explained by the model: SSE = $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- Variability in *y* explained by the model: SSR = $\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$

  Where $\overline{y}$ is the mean of y.



The strength and fit of the model can be measured by the R^2.

- $R^2 = \frac{SSR}{SST}$ (this is the portion of the variation in *y* that is explained by *x*)
- $0 \le R^2 \le 1$, where $R^2 = 1$ is the strongest fit, and $R^2 = 0$ is the weakest fit.
- $R = r$ (sample correlation)
- Therefore $r = +\sqrt{R^2}$ (if $b_1 > 0$) or $-\sqrt{R^2}$ (if $b_1 < 0$)

The conclusions you can make after estimating a regression model include:

- For a given x_i the predicted value of y_i (i.e. $\hat{y}_i$) is…
- When x increases by 1 unit, y increases/decreases by b_1 units on average
  - Note we are not making a conclusion about causation but merely an association.
  - Make sure you choose the units correctly.

For doing hypothesis tests and confidence intervals for β_0 and β_1 see the hypothesis testing and the confidence interval notes.

Probability

Basic Rules of Probability:

Complement Rules

$$P(not\ A) = 1 - P(A)$$

$$P((not\ A)|B) = 1 - P(A|B)$$

Multiplicative Rule

$$P(B) = \frac{P(A\ and\ B)}{P(B)}$$
$$P(A) = \frac{P(A\ and\ B)}{P(A)}$$

$$\therefore P(A\ and\ B) = P(B)P(B) = P(A)P(A)$$

Additive Rule

$$P(A\ or\ B) = P(A) + P(B) - P(A\ and\ B)$$