Introduction to Statistics

MACSI
Mathematics Applications Consortium for Science & Industry

Why learn statistics?

Data are being generated on a huge scale in every area of our lives, e.g. Netflix, Facebook, marketing, product development, sports, etc.

How do we extract valuable information and draw meaningful conclusions from these data?

Statistics is the science of learning from data.

It gives us the tools to answer some very interesting questions!

What is statistics?

MACSI
Mathematics Applications Consortium for Science & Industry

1. Data collection

The MOST important step but often ignored or given little thought.

Data collected in a very ad hoc way, without thinking about the problem at hand.

There is nothing statistics can do if the data are poor quality!

2. Descriptive statistics

Exploring the data.

NB - Data visualisation.

Creating numerical summaries of the data.

Calculating appropriate summary statistics, e.g. means, variability, proportions, etc.

3. Statistical inference

Interpreting and drawing conclusions from the data in the presence of variability.

Generalising the results.

Hypothesis testing, confidence intervals, predictive modelling, etc.

Goal of statistics

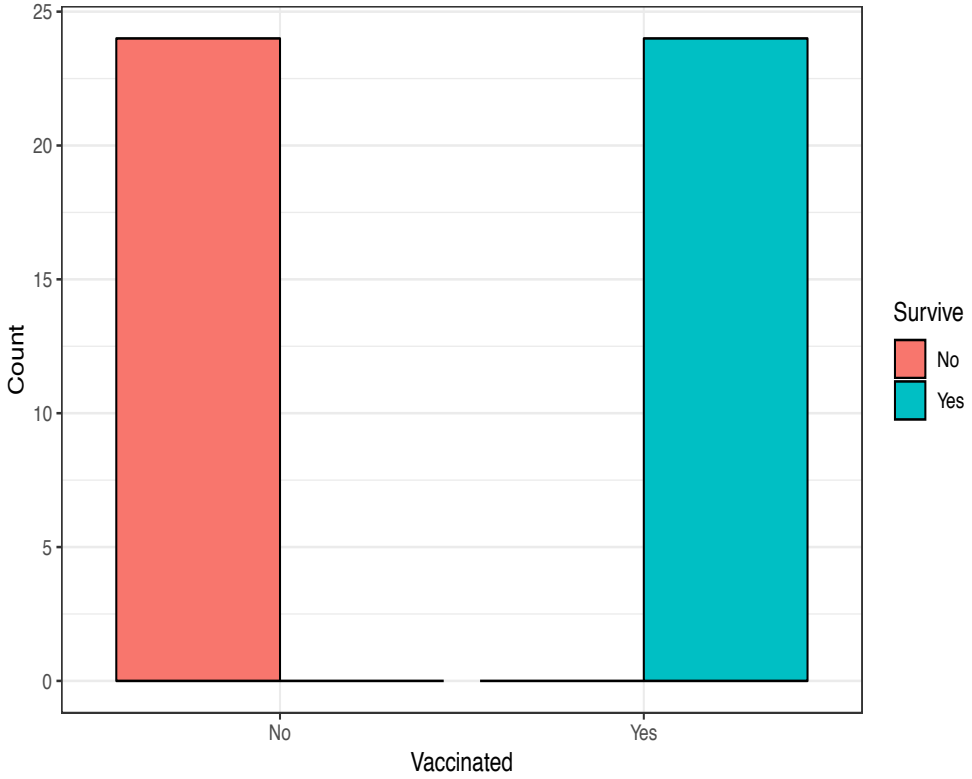
Statistics deals with sets of numbers that have something in common but differ from individual to individual.

Different values => *variability*.

Variability can obscure the patterns (signal) which are of interest to us.

Statistics helps us to distinguish the pattern (signal) from the variability (noise).

Example: Anthrax vaccine



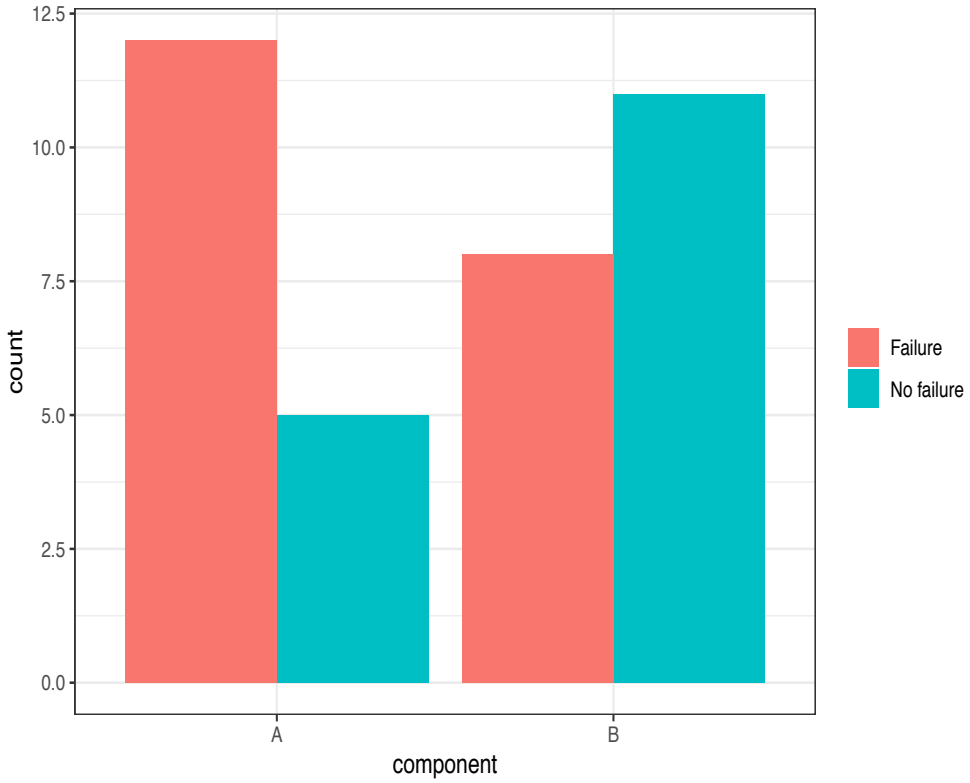
All vaccinated survived and all un-vaccinated died.

No variability is present.

Pattern is clear.

MACSI
Mathematics Applications Consortium for Science & Industry

Example: Phone components



Different failure rates for A and B.

Variability is present.

Is the difference due to component (pattern) or chance (variability)?

MACSI
Mathematics Applications Consortium for Science & Industry

Where does variability come from?

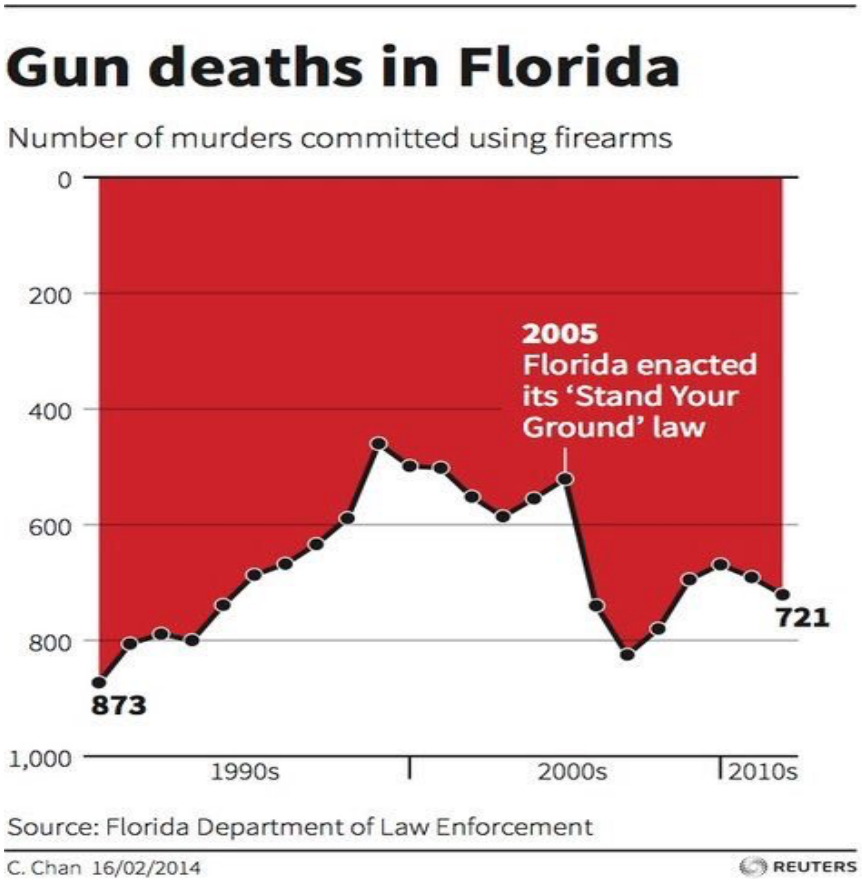
1. Natural variation, e.g. genetic, environmental.

   Try to average out sources of variation that we cannot control by carrying out data collection in a well-planned way.

2. Measurement process.

   Measure something => some associated error, e.g. rounding error, instrumentation may not be particularly sensitive, human error, etc.

Statistics cautionary tales!



MACSI
Mathematics Applications Consortium for Science & Industry

Statistics cautionary tales!

Health

Processed meats do cause cancer - WHO

By James Gallagher
Health editor, BBC News website

26 October 2015 | Health | Share

Processed meats - such as bacon, sausages and ham - do cause cancer, according to the World Health Organization (WHO).

MACSI
Mathematics Applications Consortium for Science & Industry

Interesting reads

Ben Goldacre's blog www.badscience.net and book *Bad Science*.

www.sciencedaily.com

https://understandinguncertainty.org/

https://statsandstories.net/

https://callingbullshit.org/

Twitter: @justsaysinmice; @f2harrell; @stephensenn; @d_Spiegel; @statsepi

MACSI
Mathematics Applications Consortium for Science & Industry

Data collection

MACSI
Mathematics Applications Consortium for Science & Industry

Sampling

We require data to perform any statistical analysis.

Typically we are interested in a very large group of individuals (not necessarily people!) on which we can measure characteristic(s) of interest.

This large group is called the *population*.

Sampling

Often it is not practical to measure everyone/everything in the population (cost, time, etc.)

Take a subset of the population of interest to learn about the population as a whole.

This subset is called a *sample*.

NNB! The sample must be *representative* of the population.

Example

Interested in whether (pharmaceutical) tablet weights are meeting a required specification.

Does it make sense to measure all tablets?

No! Too costly and time consuming.

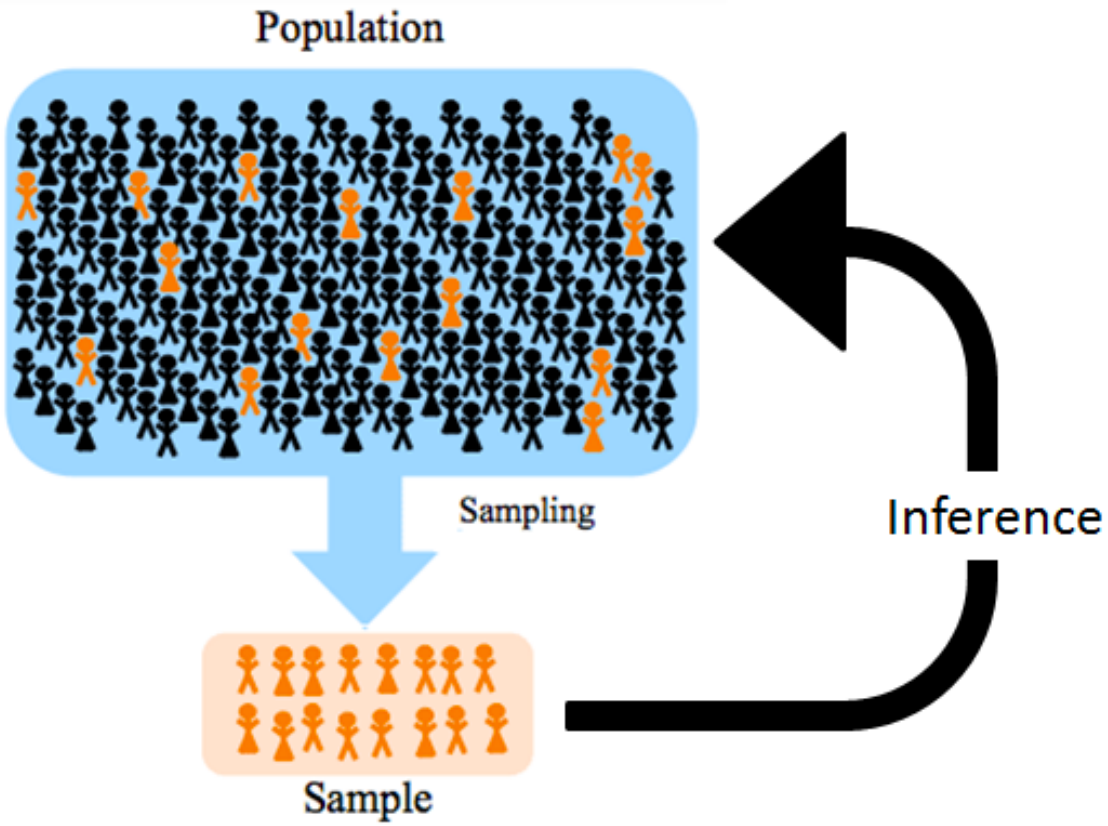
Terminology

Population: All items/individuals of interest.

Sample: A representative subset drawn from the population.

NOTE: The sample is not of interest in itself. Used to study and make inference(s) about the population.

Terminology



Terminology

Sampling/ experimental unit:	An individual member of the population of interest.
Sampling frame:	List identifying the individuals in the population.
Variable:	Something we measure for each individual but its value can change from individual to individual.

MACSI
Mathematics Applications Consortium for Science & Industry

Terminology

Parameter: A value (e.g. average, percentage, etc.) that we calculate for the *population*.

Parameter is typically *unknown* since it requires measuring the entire population.

Parameter is *fixed* (i.e. its value does not change).

Terminology

Statistic: A value (e.g. average, percentage, etc.) that we calculate for the *sample* and is our best estimate of the population parameter.

Statistic is *known* since it is calculated from the sample.

Statistic is *not fixed* (i.e. its value does change) as if we take another sample of the same size, the value is likely to change.

MACSI
Mathematics Applications Consortium for Science & Industry

Example

The ESENER survey examines how health and safety is managed in EU workplaces.

The survey was conducted in 2009, on companies in the 31 participating countries with 10 or more employees.

In each of the 36,000 establishments surveyed, the survey asked whether safety experts were used in the establishment. 71% of companies used a safety expert.

MACSI
Mathematics Applications Consortium for Science & Industry

Example

Population:	All companies with 10 or more employees in the 31 countries.
Sample:	36,000 companies in the study.
Sampling frame:	A list of all companies with 10 or more employees.
Variable:	Is a safety expert used? Yes/No
Parameter:	The true proportion of companies who use a safety expert.
Statistic:	Sample proportion = 71%.

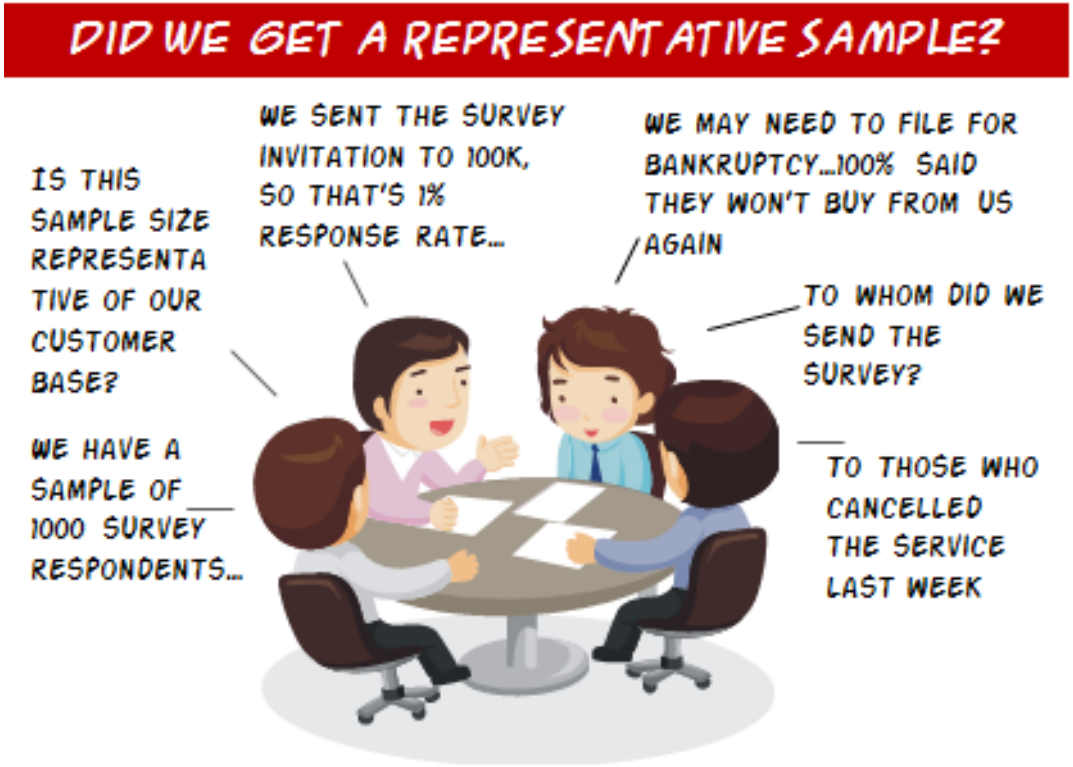
Representative sample

NB! A sample should be *representative* of the population of interest.

Sample should be a mini version of the population. No group of individuals in the population omitted or over-represented.

A non-representative sample => results not generalisable to the population as a whole and conclusions may be incorrect.

Representative sample



A large sample size is NOT the same as a representative sample

MACSI
Mathematics Applications Consortium for Science & Industry

Bias

Bias is the tendency to consistently over or underestimate the parameter of interest.

True weight of a tablet is 15g. Weigh 4 tablets and get weights of 16g, 18g, 21g, 17g => biased measurements.

NNB! Increasing the sample size does NOT eliminate bias.

Precision

Precision measures how close together (similar) sample statistics obtained from repeated sampling are.

Take 10 different samples (of same size) from same population and calculate the sample mean for each:

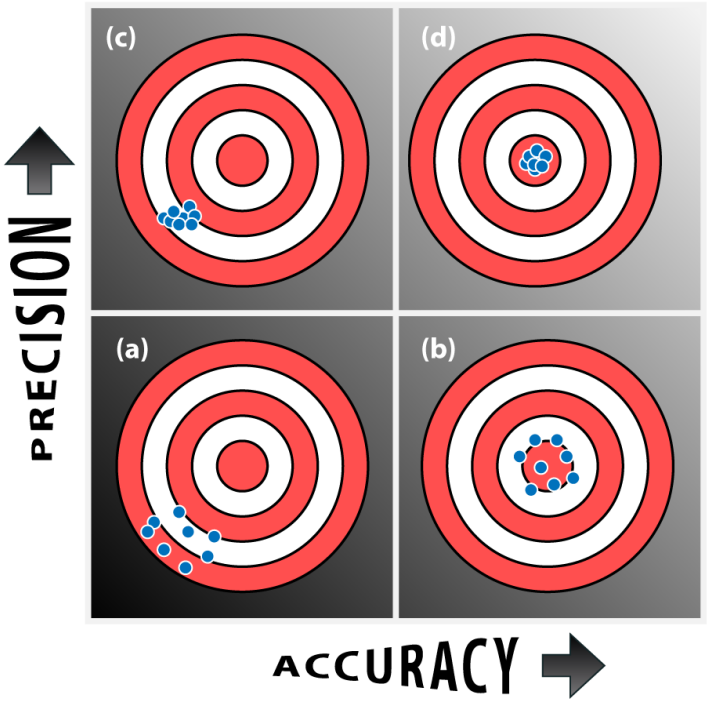
- Sample means close together => high precision (good!).
- Sample means are far apart => low precision (not good!).

NNB! Increasing the sample size DOES improve precision.

Bias (accuracy) vs precision

Biased (not accurate)

Precise



Unbiased (accurate)

Precise

Biased (not accurate)

Not precise

Unbiased (accurate)

Not precise

MACSI
Mathematics Applications Consortium for Science & Industry

Sampling methods

Must choose a sample so that it is *representative* of the population of interest and avoids *bias.*

Two main types of sampling methods:

1. Non-probability sampling
2. Probability sampling

Non-probability sampling methods

Convenience sampling – select a group of individuals to be in your sample as they are willing to participate or it's easy to measure them.

Advantage – easy to do, cost efficient (maybe).

Disadvantage – can't evaluate how representative of the population the sample is.

Non-probability sampling methods

Judgement sampling – ask an expert to select a group of individuals they consider most representative of the population of interest.

Advantage – easy to do, cost efficient (maybe).

Disadvantage – quality of the sample depends on the expert.

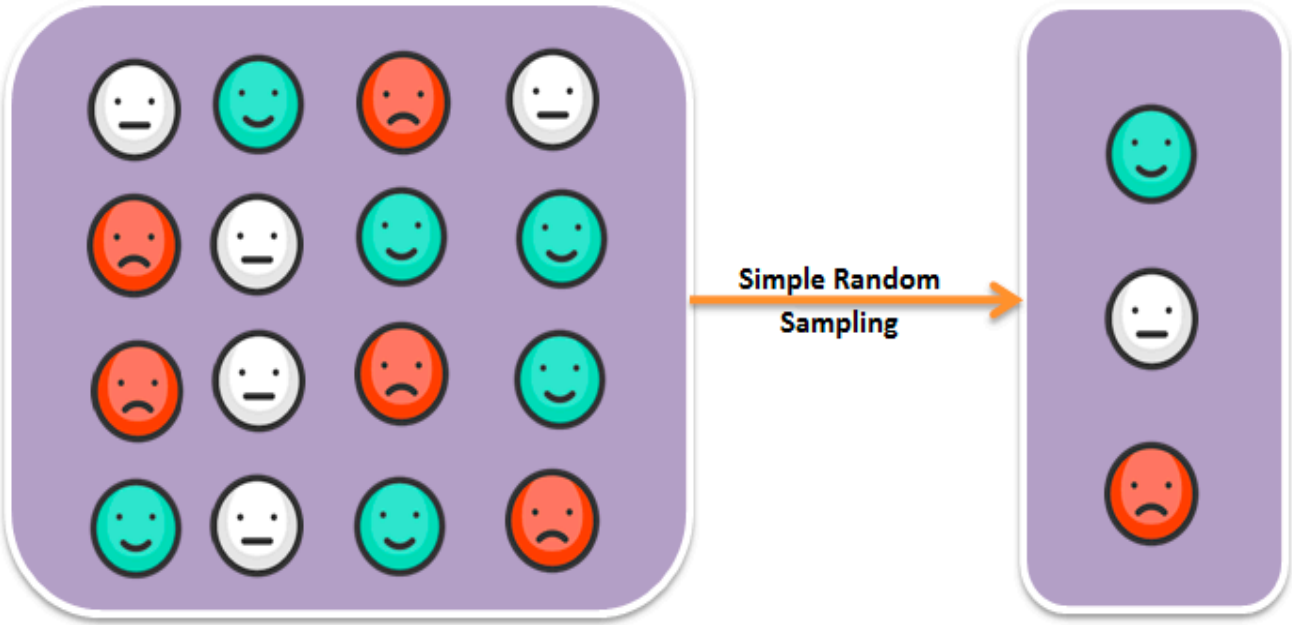
Probability sampling methods

Random sampling – Every member of the population has the same chance of being included in the sample. Every individual is selected to be in the sample independently of everyone else.

Advantage – protects against bias.

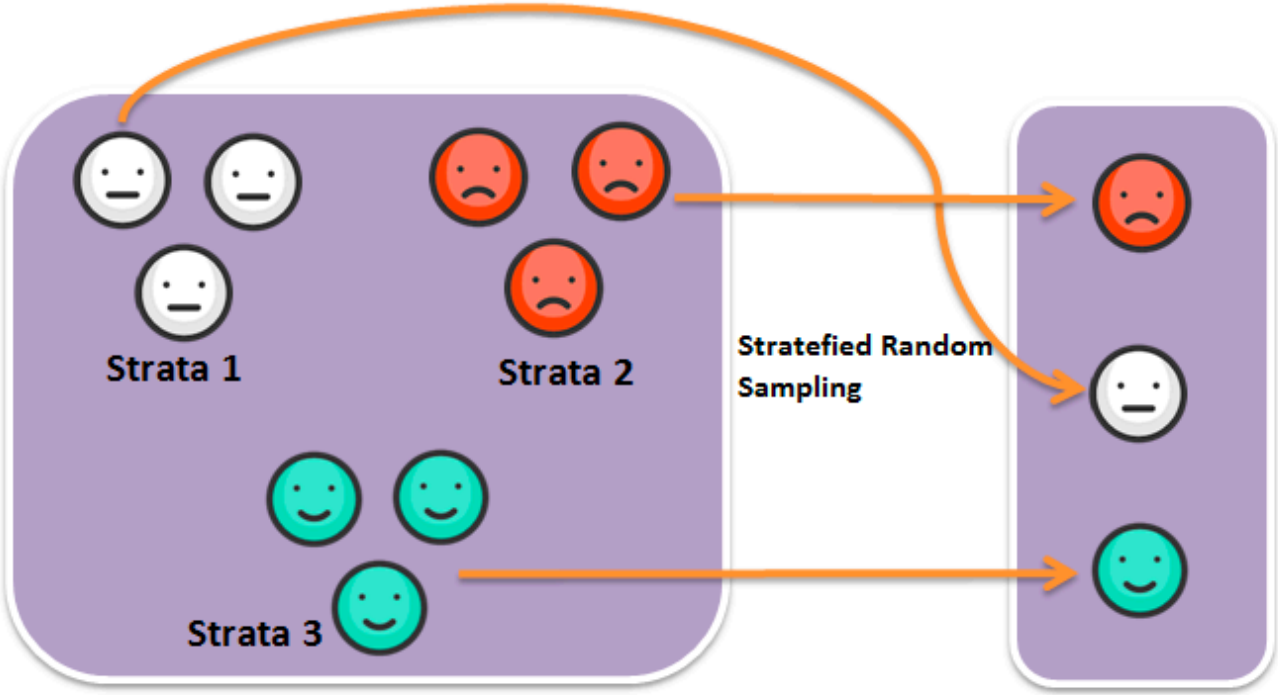
Disadvantage – harder to do.

Simple random sampling



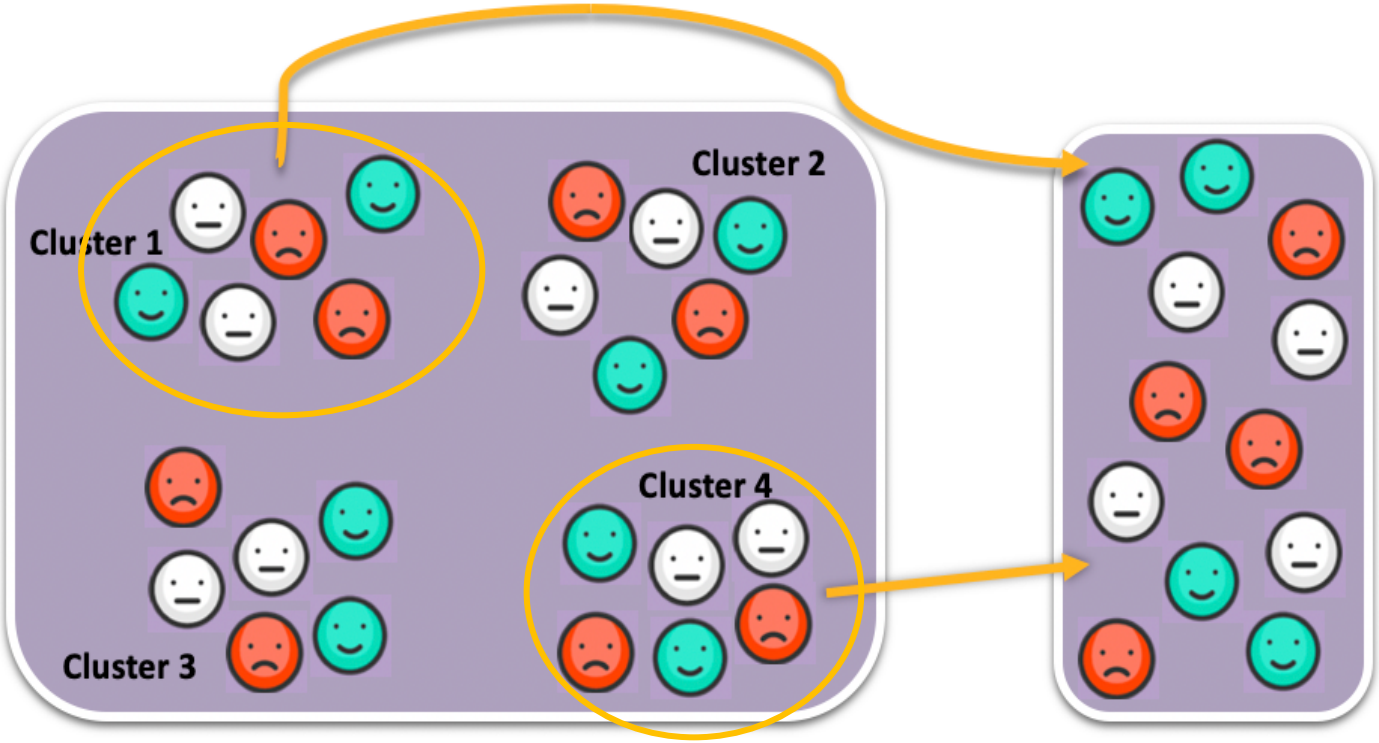
Source: www.datasciencemadesimple.com

Stratified random sampling



Source: www.datasciencemadesimple.com

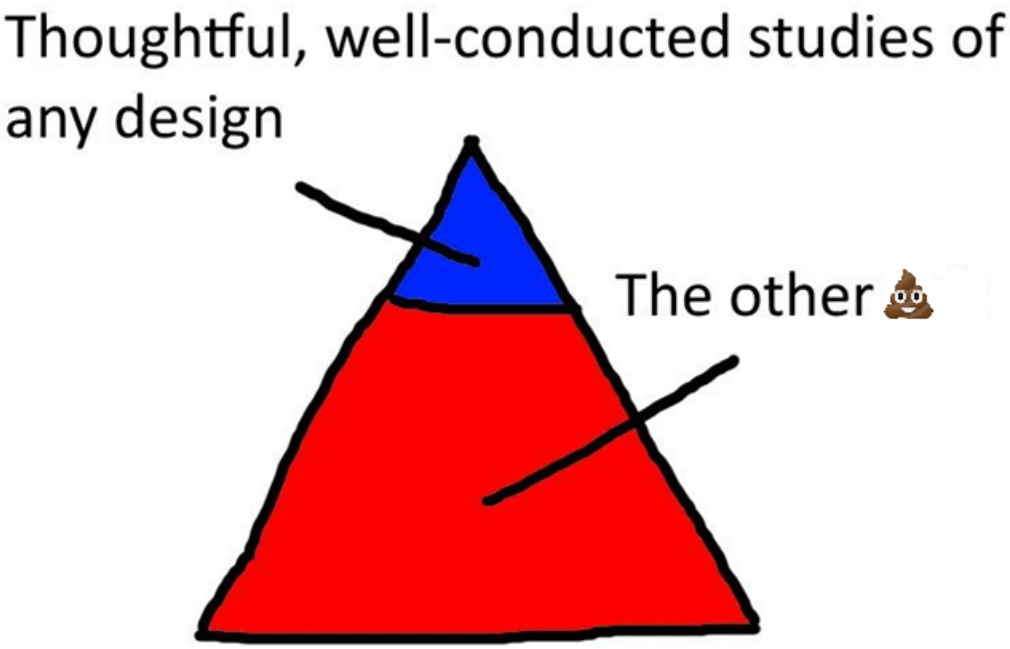
Cluster sampling



Experimental design

MACSI
Mathematics Applications Consortium for Science & Industry

Types of studies



Source: @statsepi

Types of studies

Studies are often carried out to collect data.

There are two main types of studies.

1. Observational studies: researcher collects information but does not influence events, e.g. opinion polls, National Cancer Registry studies.

2. Experimental studies: researcher deliberately influences events and investigates the effects of the intervention, e.g. lab studies, clinical trials, Design of Experiments (DOE) in manufacturing/pharma.

Types of studies

Prospective studies: data are collected forward in time from the start of the study.

Retrospective studies: data refer to past events and may be acquired from existing sources.

Longitudinal studies: examine changes over time (sometimes in relation to an intervention), e.g. The Irish Longitudinal Study on Ageing (TILDA), Growing Up in Ireland (GUI).

Cross-sectional studies: individuals are observed only once.

Observational studies

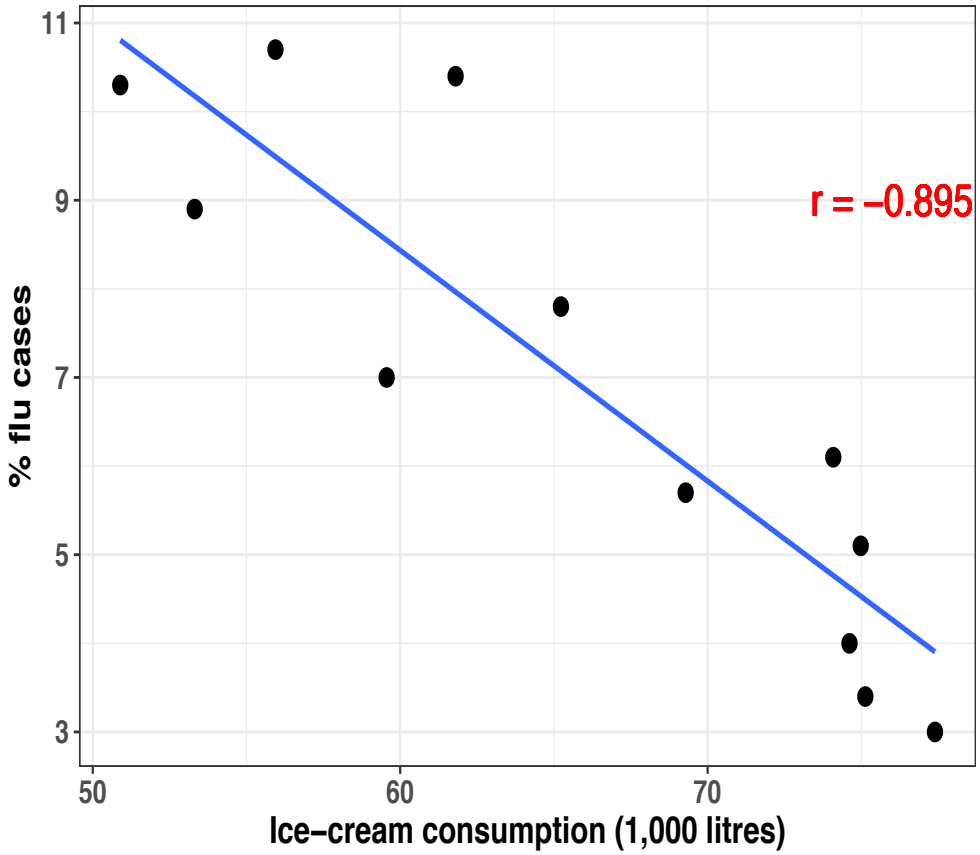
Describes aspects of a population by study of a sample of the population.

It is a snapshot of the population at a particular time.

Major limitation - they *describe* but do not *explain*, i.e. relationships cannot be interpreted as causal.

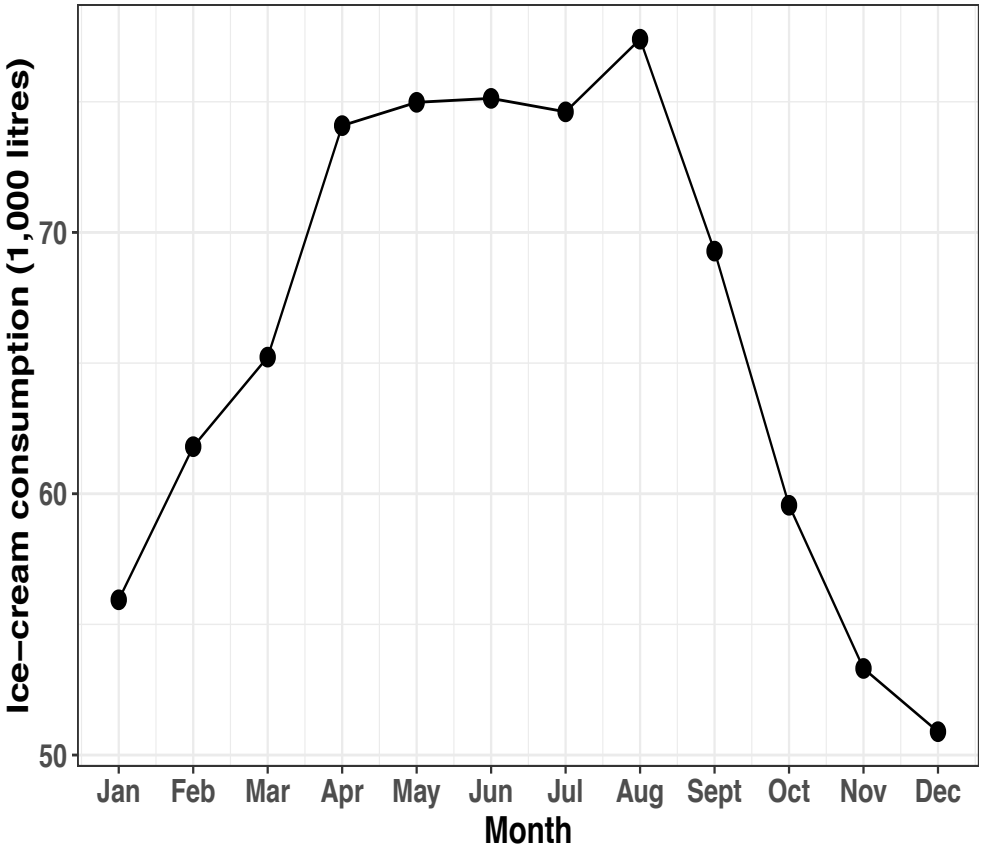
MACSI
Mathematics Applications Consortium for Science & Industry

Does ice-cream cure the flu?



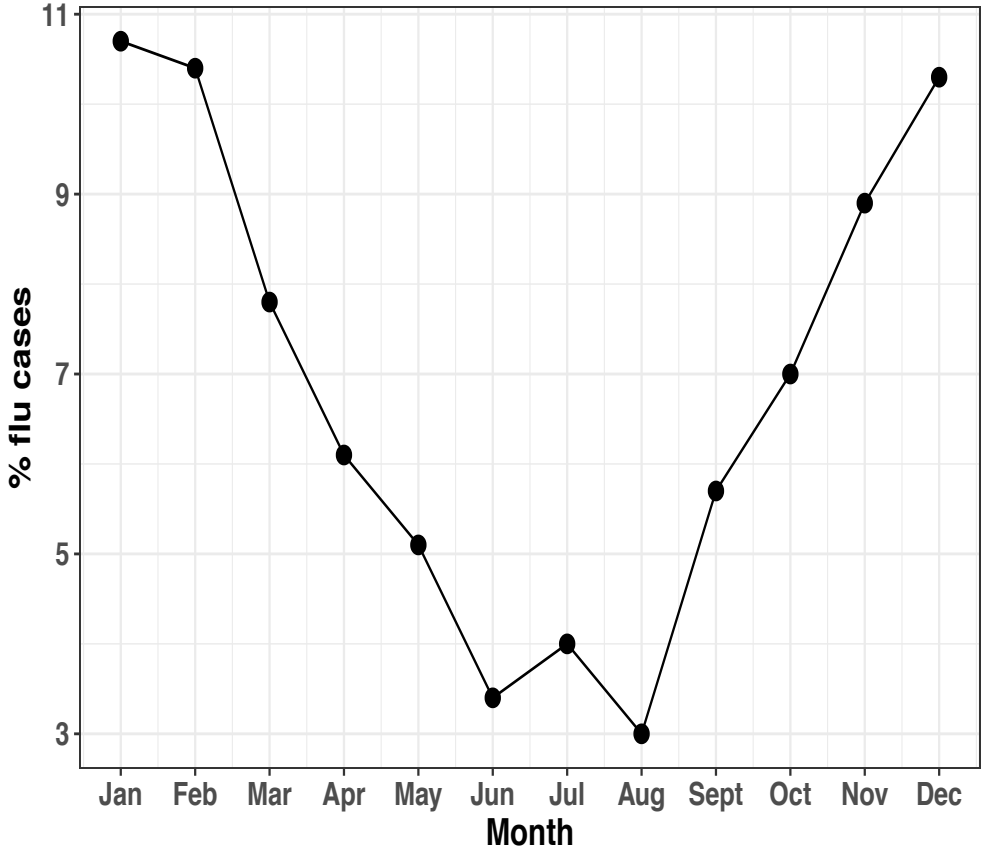
MACSI
Mathematics Applications Consortium for Science & Industry

Does ice-cream cure the flu?



MACSI
Mathematics Applications Consortium for Science & Industry

Does ice-cream cure the flu?



Observational studies

We can conclude that there is a relationship but we are just passive observers.

It is not clear which variables have an *effect* on which other variables => causality issue.

May be unmeasured factors that affect several of the measured variables, their resultant correlation suggesting causality between them.

Experimental studies

Widely used in medicine, science and engineering to:

- Identify new treatments for cancer/other diseases.
- Reduce time to design/develop new products or processes
- Improve current process performance
- Improve reliability of products.

MACSI
Mathematics Applications Consortium for Science & Industry

What is experimental design?

Collecting data on a system in a way that the data produced provide meaningful information about that system.

Organising an experiment to ensure that the right data, and enough of it, is available to answer the questions of interest as clearly and efficiently as possible.

Is the basic methodology for providing information on *causality*.

MACSI
Mathematics Applications Consortium for Science & Industry

Basic design concepts

The three most important aspects of experimental design are:

1. Randomisation;
2. Replication;
3. Blocking.

Randomisation

Randomisation = random assignment of treatments to experimental units to prevent bias.

Ensures each treatment is equally likely to be allocated to any given experimental unit.

Is used to try to reduce/eliminate the possibility of confounding effects that could render an experiment useless.

Guarantees the validity of tests of hypotheses.

MACSI
Mathematics Applications Consortium for Science & Industry

Example

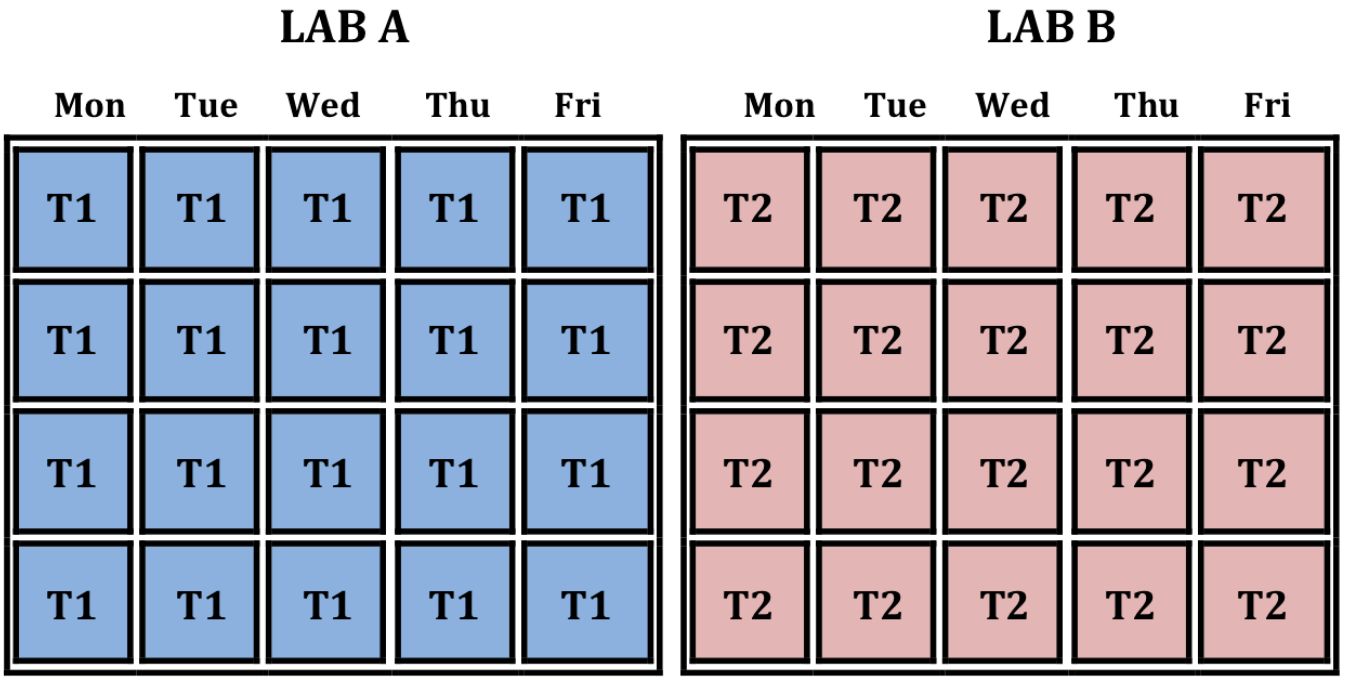
An experiment to compare the effects of two cholesterol lowering drugs. 40 patients (20 males and 20 females) randomly selected to receive a daily dose of either drug 1 or drug 2. Their cholesterol after one week was recorded.

1. *All 20 of the females receive drug 2 and all 20 of the males receive drug 1. Problem?*

2. *First the lab does the test for all 20 patients receiving drug 1 and then the 20 patients receiving drug 2. Problem?*

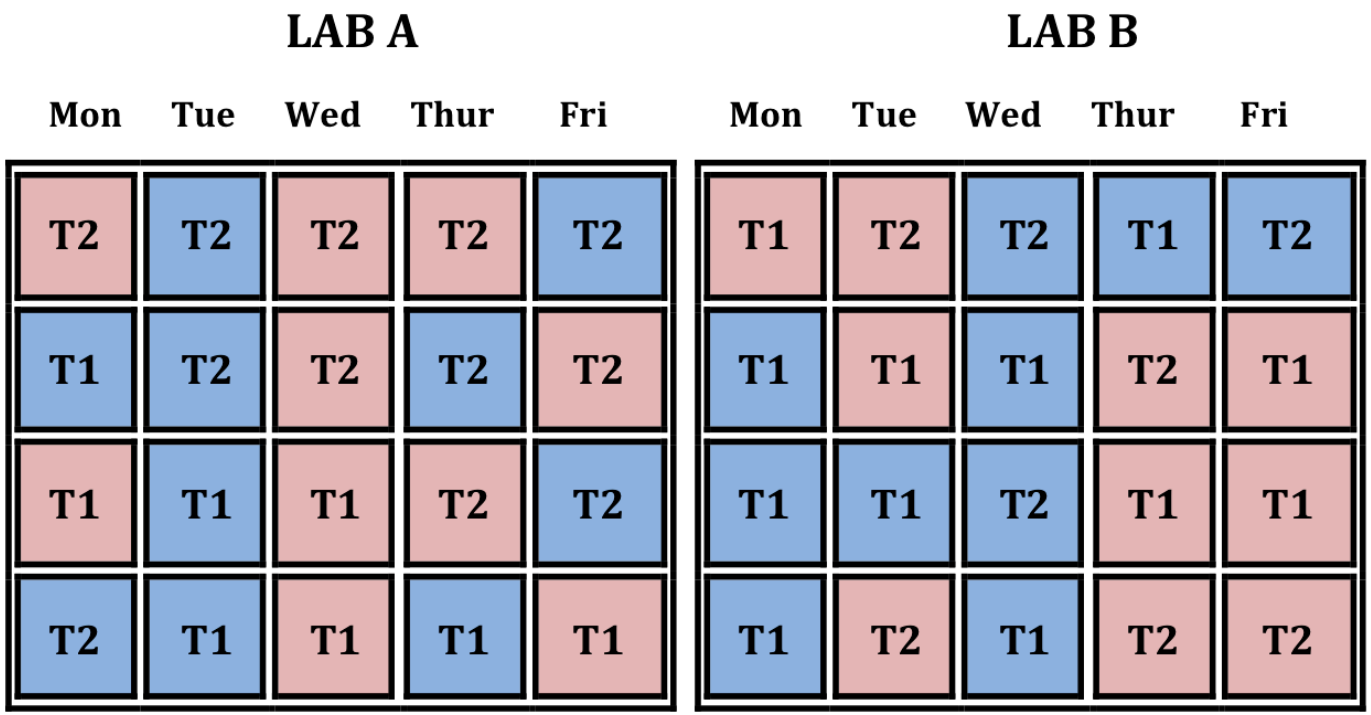
3. *Lab A does all the tests for drug 1 and Lab B does all the tests for drug 2. Problem?*

Bad experimental design



Blue = male, Pink = female, T1 = treatment 1, T2 = treatment 2

Completely randomised design



Blue = male, Pink = female, T1 = treatment 1, T2 = treatment 2

MACSI
Mathematics Applications Consortium for Science & Industry

Replication

Replication = start an experiment from scratch and repeat the entire experiment.

Is used to provide an estimate of how variable the experimental results are.

An estimate (e.g. a mean) has limited value without some statement of the uncertainty of the estimate.

Increases the confidence with which conclusions can be drawn about an experimental factor.

Example

An experiment was carried out to compare the effects of two settings (A and B) on a manufacturing performance score.

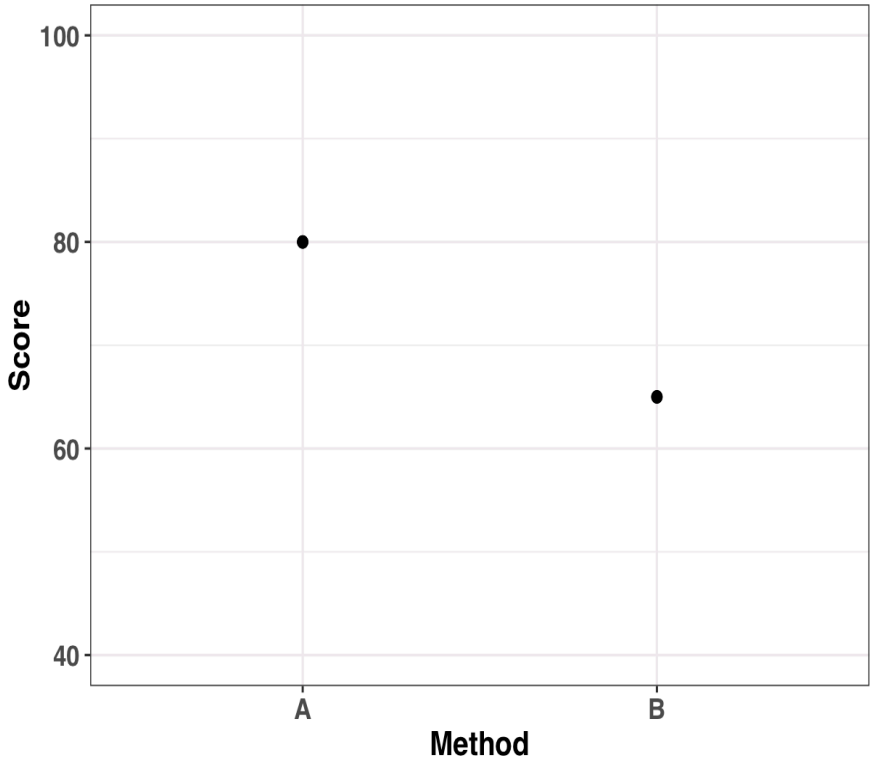
1. *One experimental unit is measured using setting A and setting B. What conclusions can be drawn?*

2. *30 experimental units are randomly assigned to setting A and 30 to setting B. What conclusions can be drawn?*

3. *300 experimental units are randomly assigned to setting A and 300 to setting B. What conclusions can be drawn?*

MACSI
Mathematics Applications Consortium for Science & Industry

Example – 1.

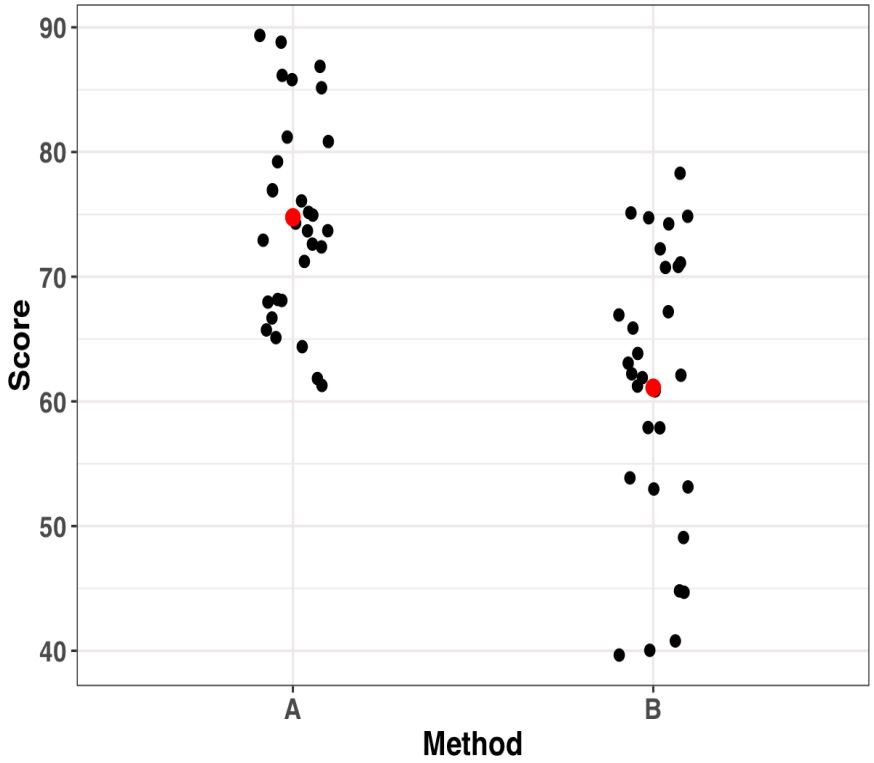


One one replicate.

No estimate of variability.

Can only say setting A gives a higher score than setting B.

Example – 2.



Have replicates.

Can estimate variability.

Can conduct a formal hypothesis test to check for differences between the average score for the two settings.

Example – 3.

Have lots of replicates.

Have a very good estimate of variability.

BUT now have an additional problem associated with very large sample sizes.

Want to compare the average score across settings. As sample size increases, the chance we conclude the averages are different (statistically) increases even though the difference is very small.

Statistical vs practical significance

If the experiment is sensitive enough a very small difference may be detected and lead to s*tatistical* significance.

However, such a small difference may not be of any *practical* significance.

On the other hand, the difference may be large enough to be of great practical importance but not be statistically significant because the experiment is not sufficiently sensitive (i.e. lacks power).

Example

A new antihypertensive drug is compared with a current drug. Change in blood pressure is noted for each person.

Average BP change new drug = 5.4 mmHg
Average BP change current drug = 5.3 mmHg

P-value = 0.001 => difference of 0.1 mmHg is statistically significant but is this difference *clinically significant?*

Example

A new antihypertensive drug is compared with a current drug. Change in blood pressure is noted for each person.

Average BP change new drug = 8.7 mmHg
Average BP change current drug = 2.3 mmHg

P-value = 0.14 => difference of 6.4 mmHg is not statistically significant but is this difference *clinically significant?*

MACSI
Mathematics Applications Consortium for Science & Industry

Blocking

If experimental units are homogeneous => the better the comparisons between treatments.

In many instances, experimental units are extremely heterogeneous.

To overcome this issue, *blocking* is used, where individuals within blocks are very homogeneous, but individuals in different blocks can be as heterogeneous as we like.

Example

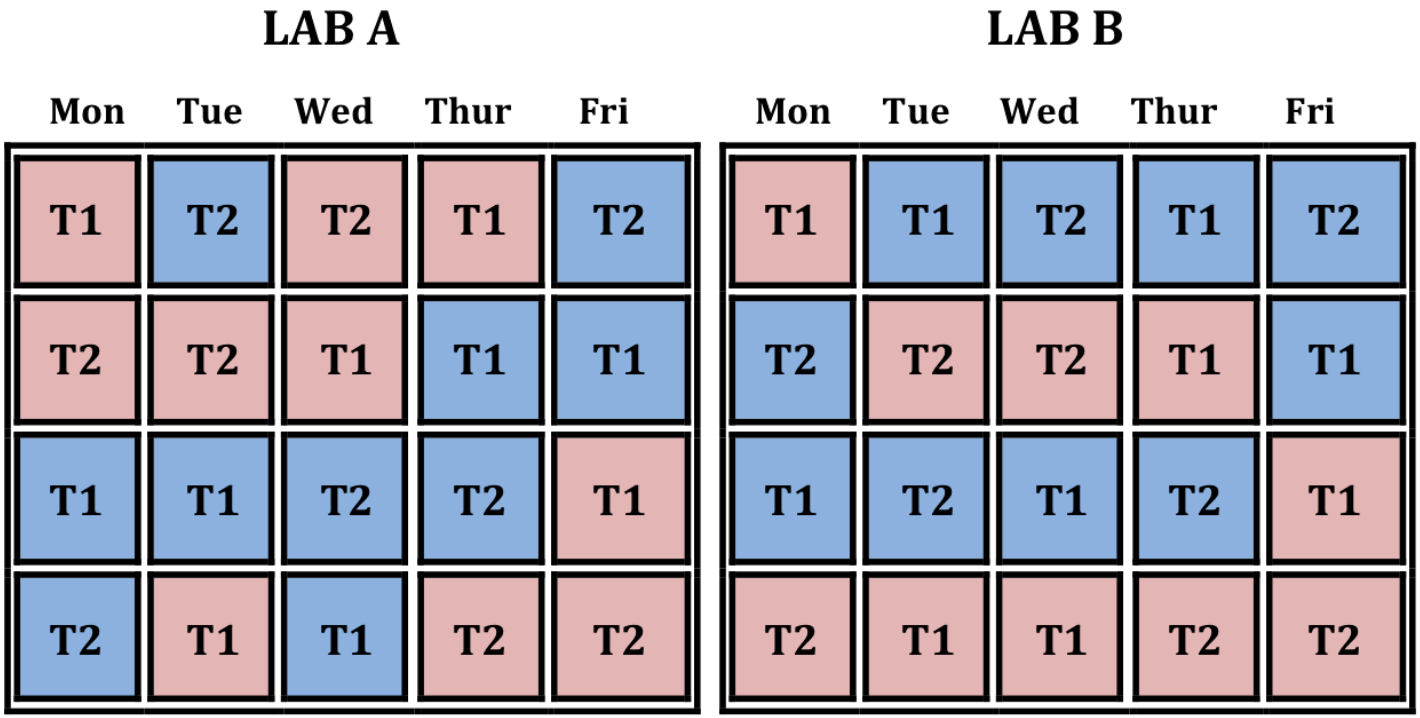
Let’s go back to the cholesterol example.

There are two treatments: Drug 1 and Drug 2.

Labs A and B perform the tests over different days of the week.

Day of the week is a blocking factor accounting for variability between test measurements done of different days of the week.

Completely randomised design



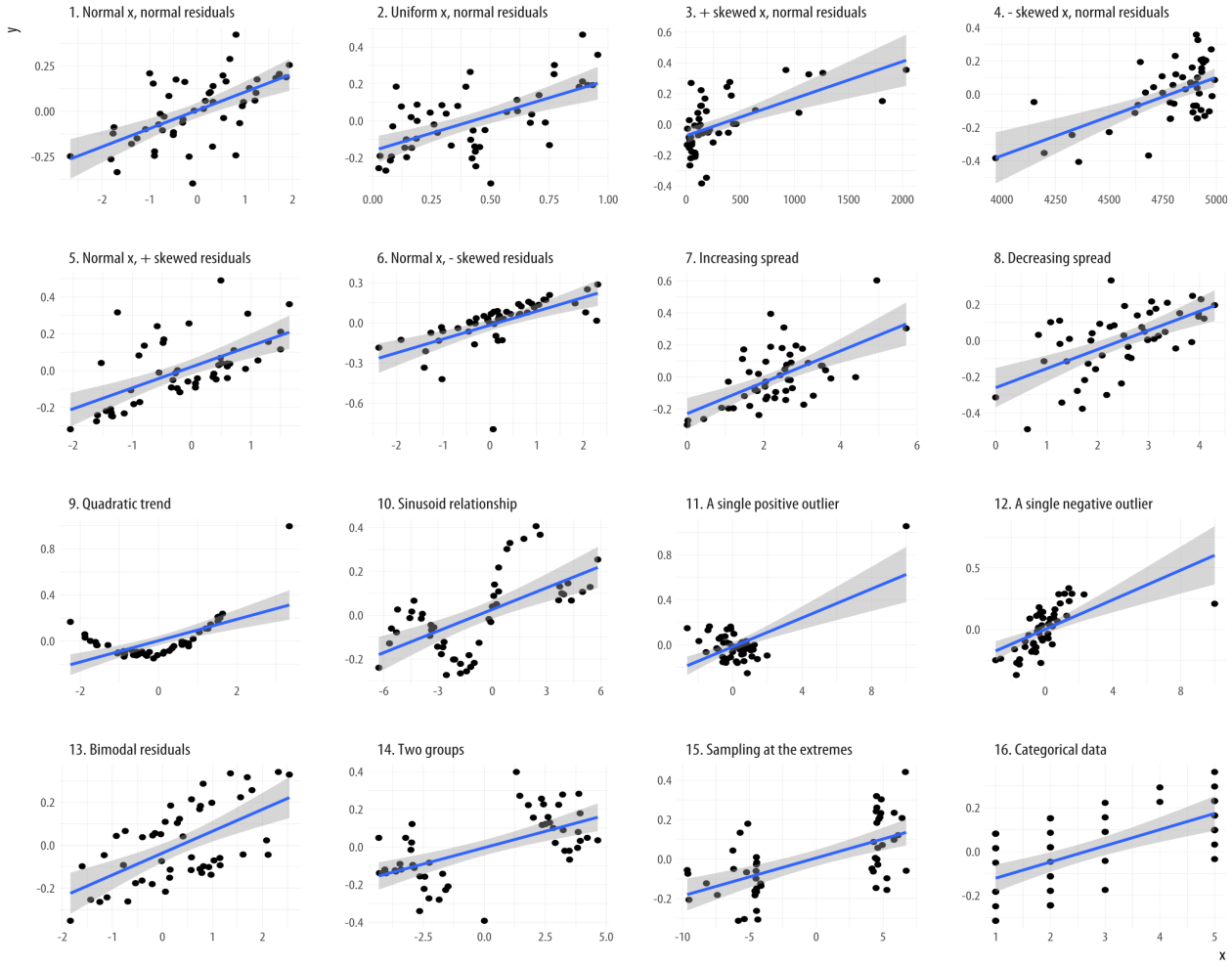
Blue = male, Pink = female, T1 = treatment 1, T2 = treatment 2

Types of experimental designs

1. Completely randomised design.
2. Matched-pairs design.
3. Repeated measures design.
4. Factorial design.
5. Fractional factorial design.
6. Split-plot design.
7. Etc.

Visualising and summarising data

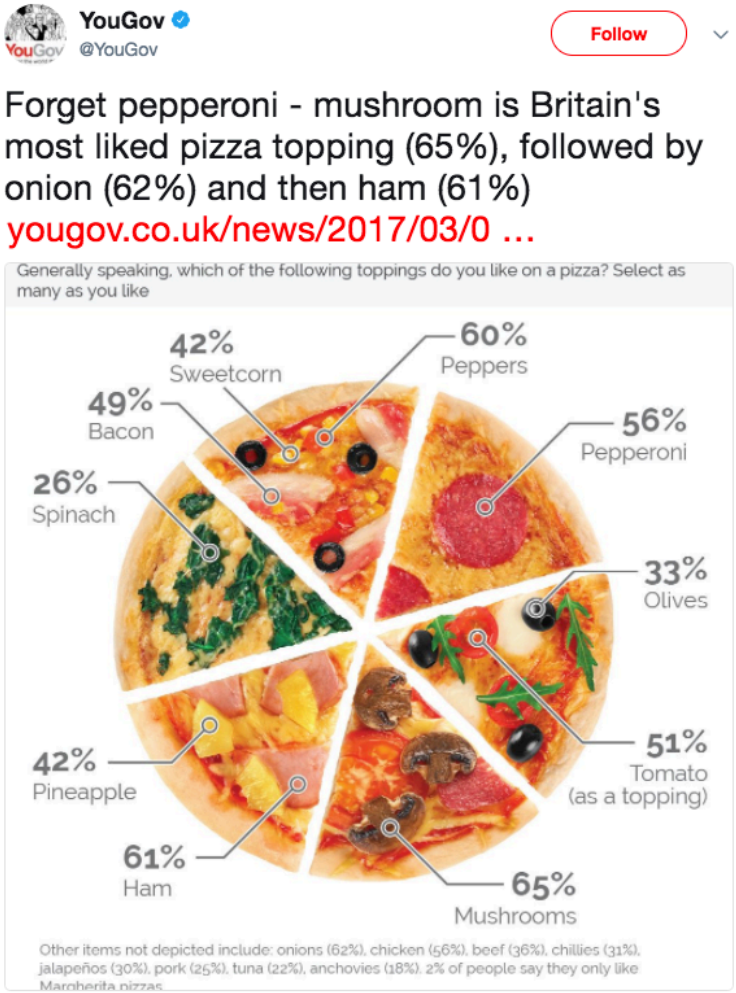
Why visualise data?



MACSI
Mathematics Applications Consortium for Science & Industry

Bad graphics

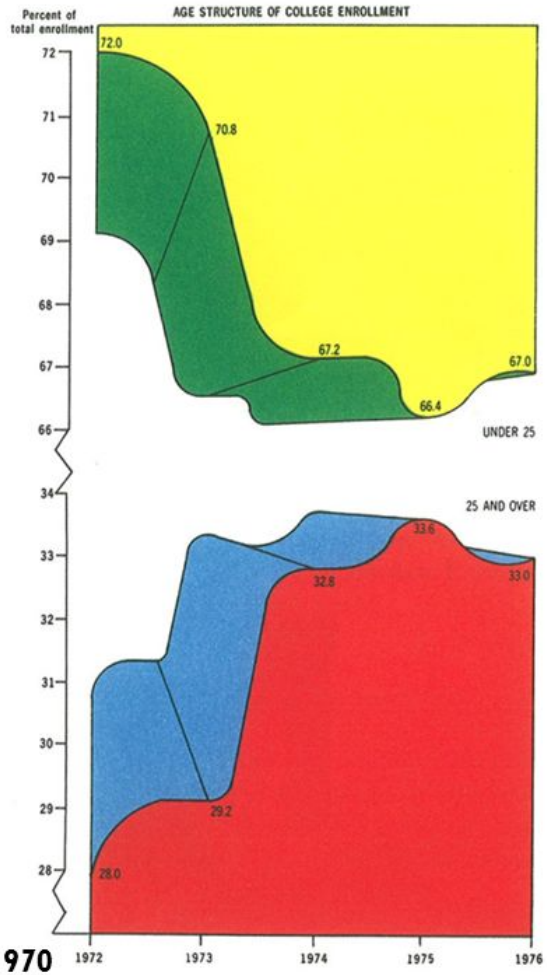
Worse graphics



MACSI
Mathematics Applications Consortium for Science & Industry

Worst graphic ever

"This may well be the worst graphic ever to find its way into print"



American Education, 1970

MACSI
Mathematics Applications Consortium for Science & Industry

Data types

Two main types of data:

1. Categorical/qualitative data - labels/categories (non-numeric).
2. Quantitative data – numeric.

These data types can then be further classified into sub-types.

Categorical data

Has two sub-types:

1. Nominal – labels/categories have no ordering, e.g. male/female, yes/no.

2. Ordinal – labels/categories have a meaningful ordering, e.g. excellent, very good, good, poor, very poor.

Ordinal data has more information that nominal data.

Quantitative data

Has two sub-types:

1. Interval – numeric, no unique zero point, e.g. time of day, temperature.

2. Ratio – numeric, has a unique zero point, e.g. weight, height.

Ratio data contains the most information.

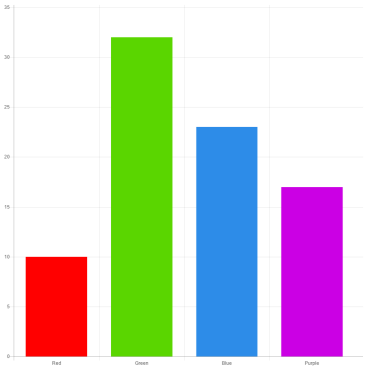
1 & 2 can be further classified as discrete (whole numbers) or continuous (contain decimal places).

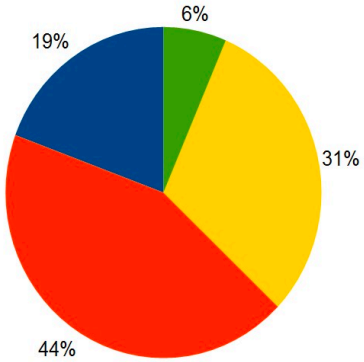
Example

Age (years)	Quantitative, ratio, discrete
Sex	Categorical, nominal
Education	Categorical, ordinal
Total bill amount (closest $)	Quantitative, ratio, discrete
Total pay amount (closest $)	Quantitative, ratio, discrete
Default (Yes/No)	Categorical, nominal

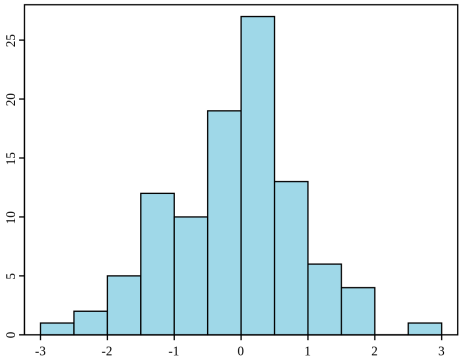
Graphical summaries

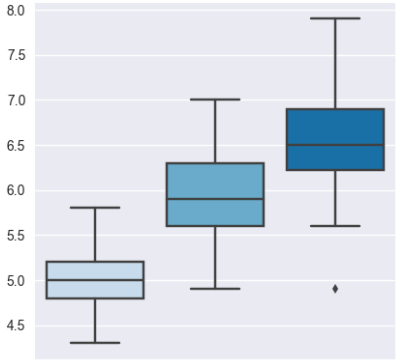
Categorical data:





Quantitative data:





MACSI
Mathematics Applications Consortium for Science & Industry

Graphing categorical data

In the credit risk dataset, 6943 people have primary level education, 9213 have second level education and 3228 have third level education.

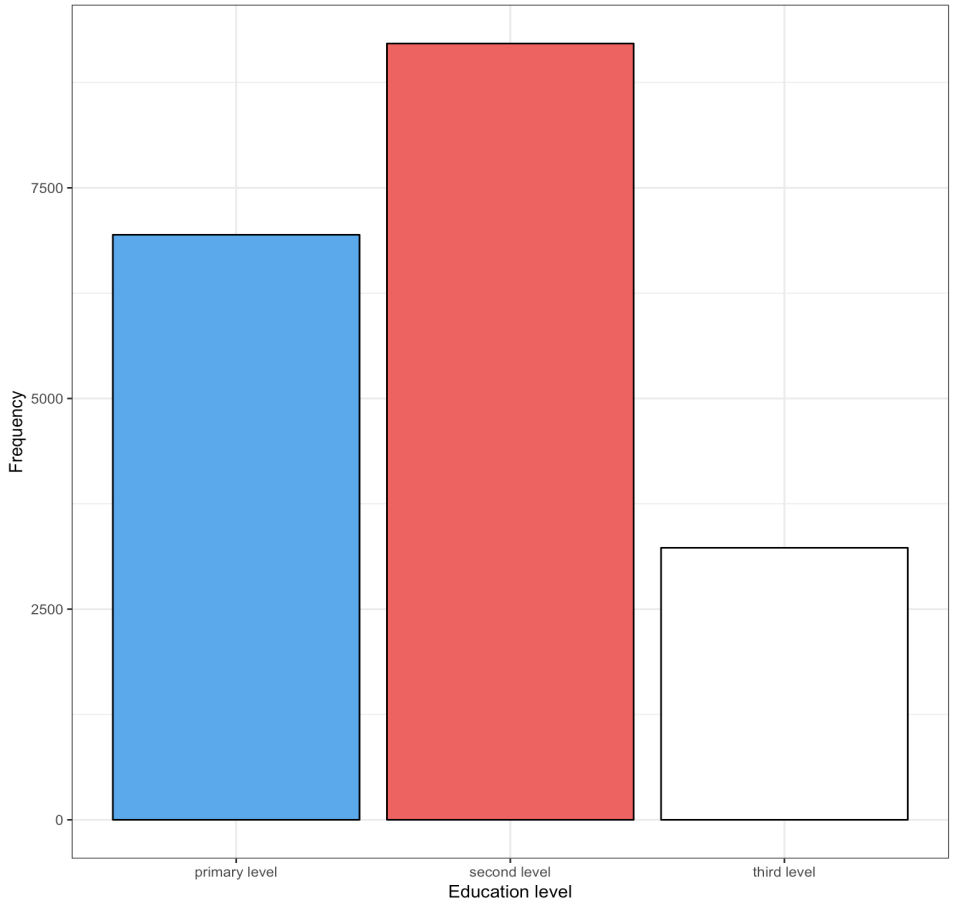
Create a graphical summary of the data.

Barchart

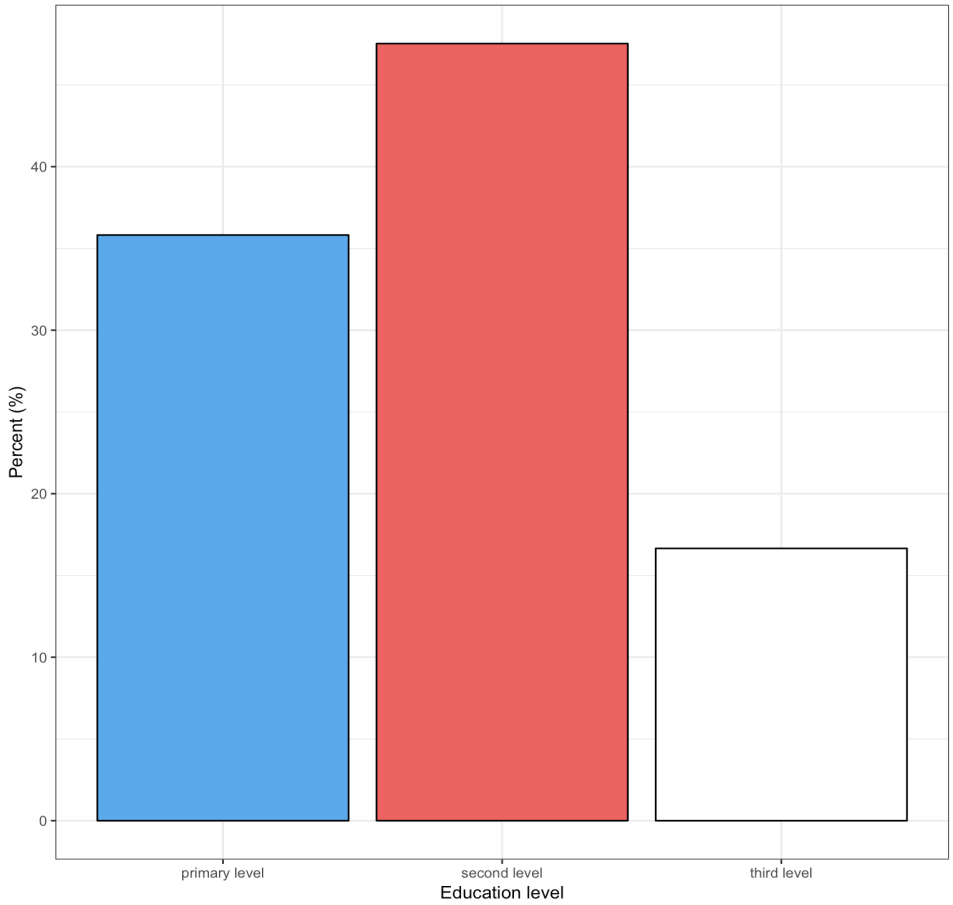
First create a frequency distribution.

Education level	Frequency	Relative frequency	%
Primary level	6943	0.358	35.8%
Second level	9213	0.475	46.5%
Third level	3228	0.167	16.7%
Total	19384	1.000	100.0%

Barchart



Barchart



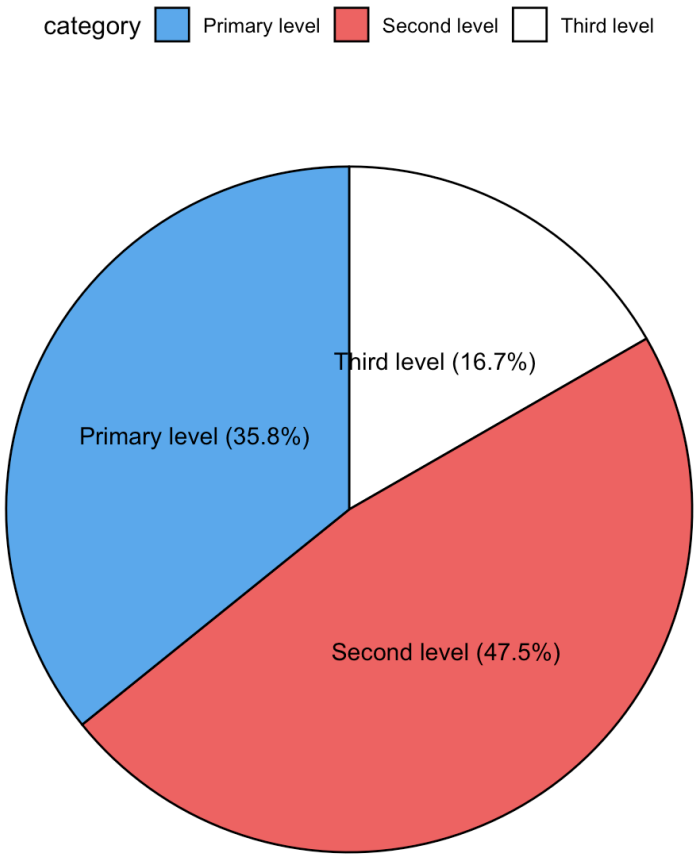
Creating a pie chart

To construct a pie-chart => need to decide on the size of the ‘slices’ in the pie.

Depends on the relative frequency.

Primary level:	360˚*0.358 = 128.88˚
Second level:	360˚*0.475 = 171˚
Third level:	360˚*0.167 = 60.12˚

Pie chart



MACSI
Mathematics Applications Consortium for Science & Industry

Graphing quantitative data

What happens if our (quantitative) data just consists of a list of numbers?

Can't use a bar for each individual number - not sensible!

First need to group the data, and then plot.

Creating a histogram

In the credit risk dataset the age of each customer was recorded. Some of the data are given below:

26 57 37 29 28 35 51 24 49 29
39 26 23 23 27…

Construct a histogram of the data.

Creating a histogram

Need to convert the raw data into frequencies, i.e. partition the data into *bins*.

Calculate the range: 75 - 21 = 54.

Decide on the number of categories/bins. We'll use 11 bins.

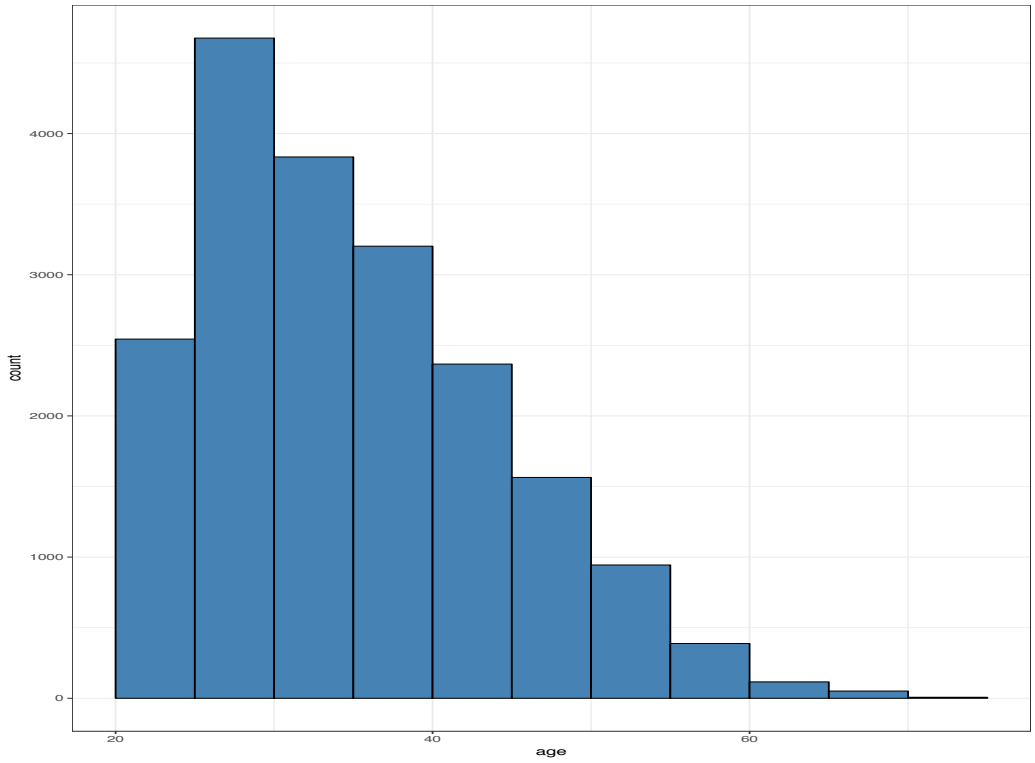
Calculate the width of the bins = Range/number of bins = 54/11 = 4.9 (use 5).

Histogram

Create the frequency distribution.

Age	Frequency	Relative frequency	%
20-24.99	2544	0.129	12.9%
25-29.99	4676	0.237	23.7%
30-34.99	3834	0.195	19.5%
…			
70-74.99	7	0.0004	0.04%
Total	19693	1.000	100.0%

Histogram



MACSI
Mathematics Applications Consortium for Science & Industry

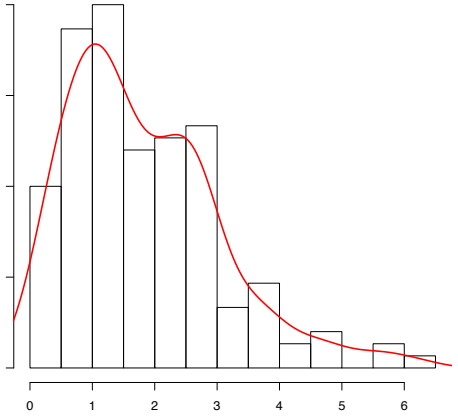
Properties of a histogram

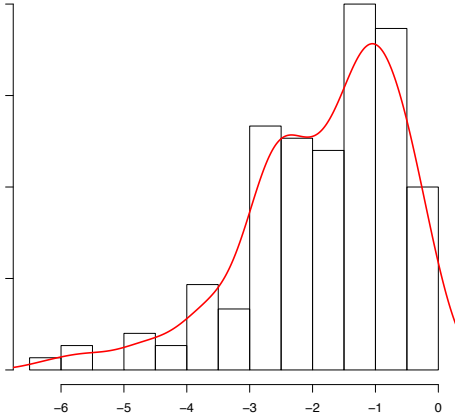
Areas of the bars are proportional to the relative frequency.

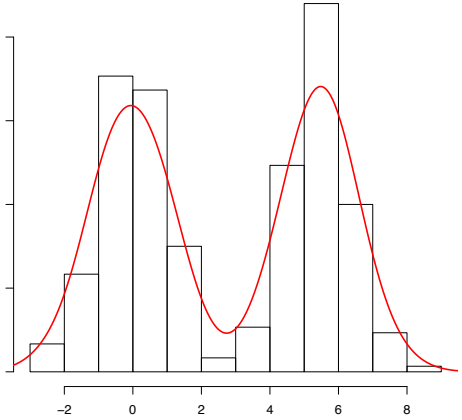
Shows the shape of the distribution (tops of the bars).

Interested in two properties – (1) is it symmetric, (2) is it bell-shaped?

Histogram and skewness







MACSI
Mathematics Applications Consortium for Science & Industry

Creating a boxplot

Boxplots are also used to summarise quantitative data.

Requires the five number summary:

1. Q_1 = 25th percentile (lower quartile)
2. Q_2 = 50th percentile (median)
3. Q_3 = 75th percentile (upper quartile)
4. Minimum
5. Maximum

Whiskers: Q_1 - 1.5 x IQR; Q_3 + 1.5 x IQR

Creating a boxplot

First order the data from smallest value to largest.

1. Q_1 = value in position $(n + 1)/4$
2. Q_2 = value in position $n/2$
3. Q_3 = value in position $3(n + 1)/4$
4. Minimum
5. Maximum

Whiskers: Q_1 - 1.5 x IQR; Q_3 + 1.5 x IQR

Example

For the age data:

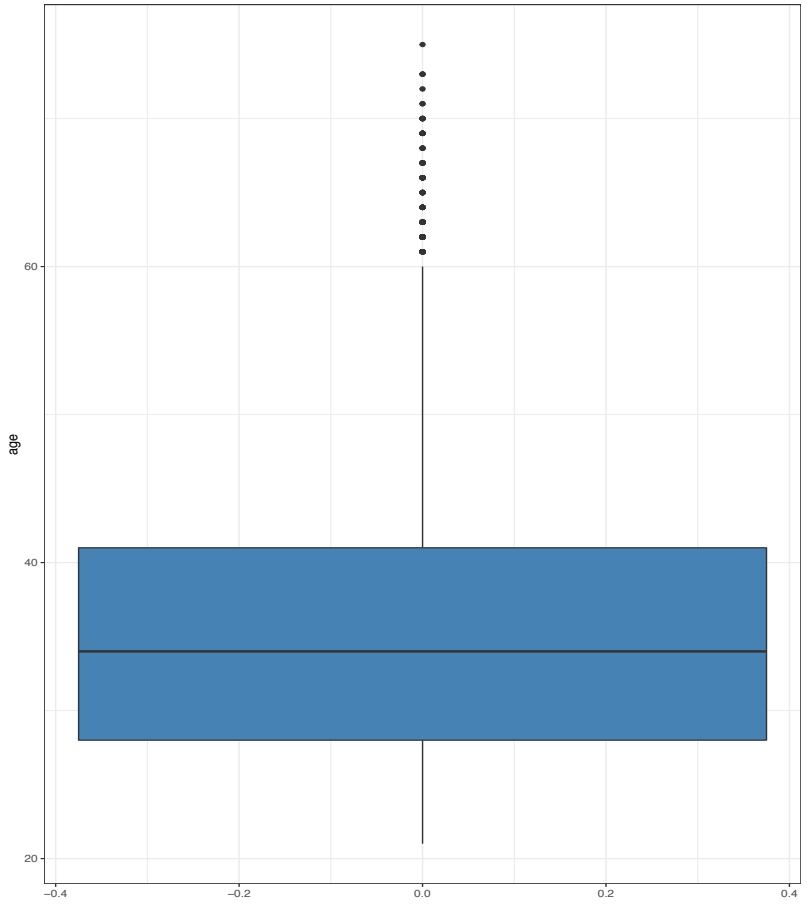
1. $Q_1 = 28$
2. $Q_2 = 34$
3. $Q_3 = 41$
4. 21
5. 75

Whiskers:

Lower: $28 - 1.5 \times (41-28) = 8.5$

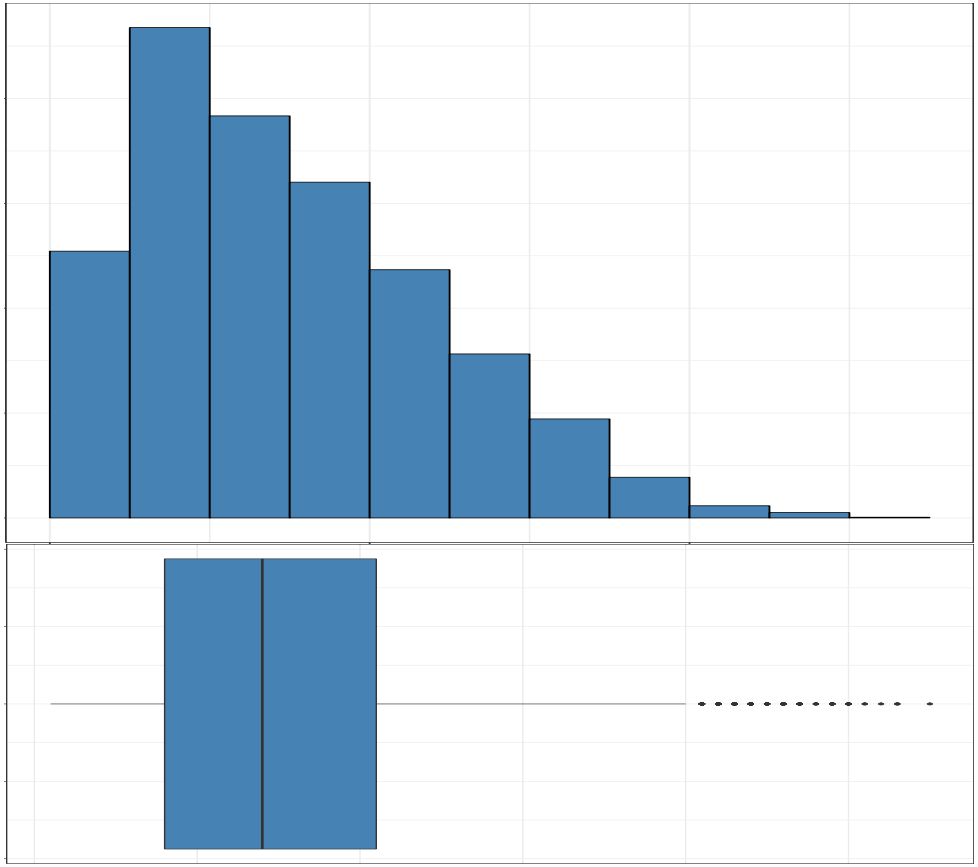
Upper: $41 + 1.5 \times (41-28) = 60.5$

Boxplot



MACSI
Mathematics Applications Consortium for Science & Industry

Link between histogram and boxplot



Numerical summaries

Main objective in statistics - infer something about a population based on a sample.

Summary measures help to characterise the data.

Numerical summaries of the population = *parameters*.

Numerical summaries of the sample = *statistics*.

Categorical data

Interested in the proportion belonging to a particular category.

$$p = \frac{\#\ pop.\ pts\ belonging\ to\ the\ category}{\#\ in\ population}$$

$$\hat{p} = \frac{\#\ sample\ pts\ belonging\ to\ the\ category}{\#\ in\ sample}$$

$\hat{p}$ (the sample proportion) is an estimate of p (the population proportion).

Quantitative data

Two main numerical summaries for quantitative data:

1. Measure of centrality (measure of the centre of the data).
2. Measure of variability (measure of the spread/dispersion of the data).

Have a sample of size *n* with each data point denoted:

$y_1, y_2, \ldots, y_n$.

Measures of centrality

1. Mean:
   - Sample mean: $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
   - Population mean: $\mu = \frac{1}{N}\sum_{i=1}^{N} y_i$
2. Median (middle number): 50th percentile (less sensitive to outliers).
3. Mode: Value which occurs most often in the data.

Example

For the age data:

1. Mean = $\bar{y}$ = 35.49
2. Median = 34
3. Mode = 29

Measures of variability

1. Variance:
   - Sample variance: $s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$
   - Population variance: $\sigma^2 = \frac{\sum_{i=1}^{N}(y_i - \mu)^2}{N}$
2. IQR = $Q_3 - Q_1$ (less sensitive to outliers).
3. Range = Maximum - minimum.

Example

For the age data:

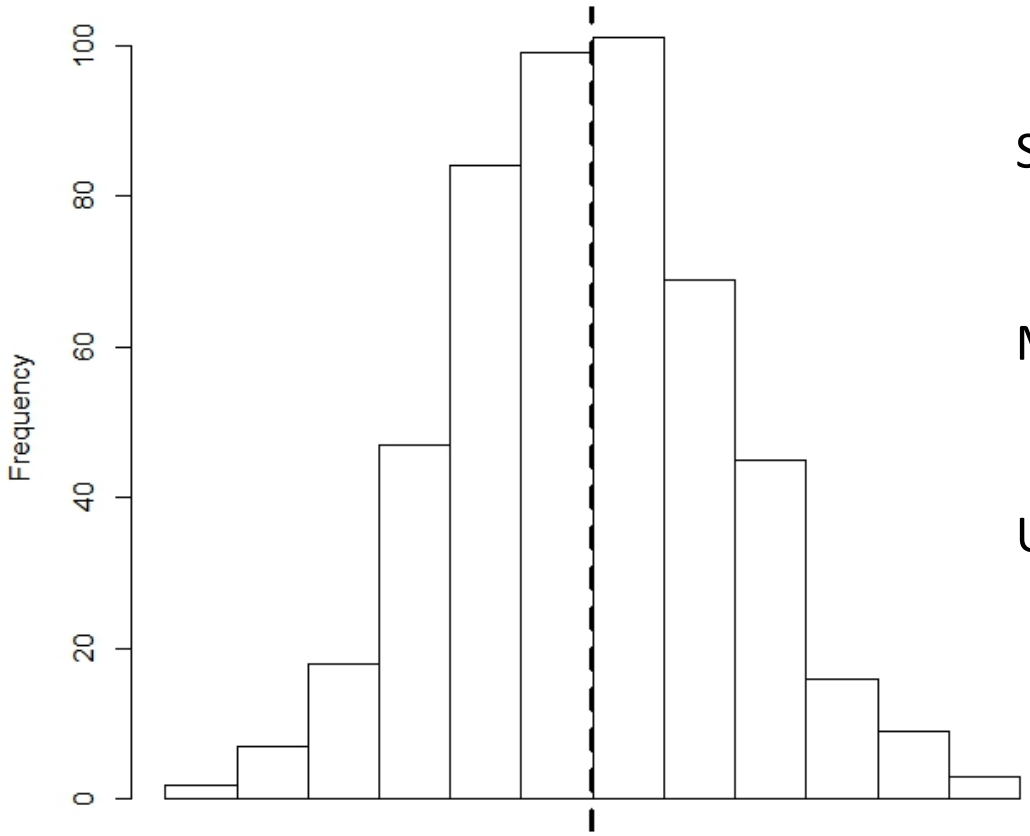
1. Variance = s^2 = 85.23 years^2

   Standard deviation = s = 9.23 years

2. IQR = 41 – 28 = 13

3. Range = 75 – 21 = 54

Which to choose?

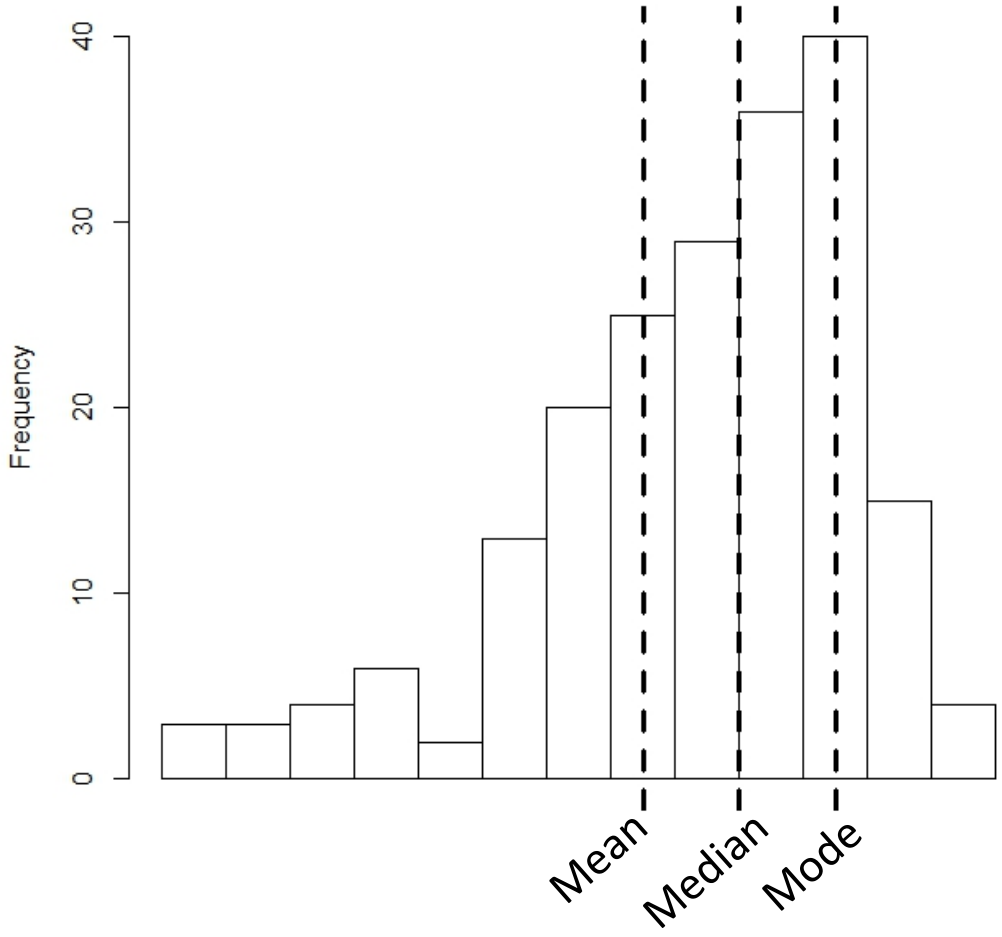


Symmetric and bell-shaped.

Mean = Median = Mode

Use mean with std. dev.

Which to choose?



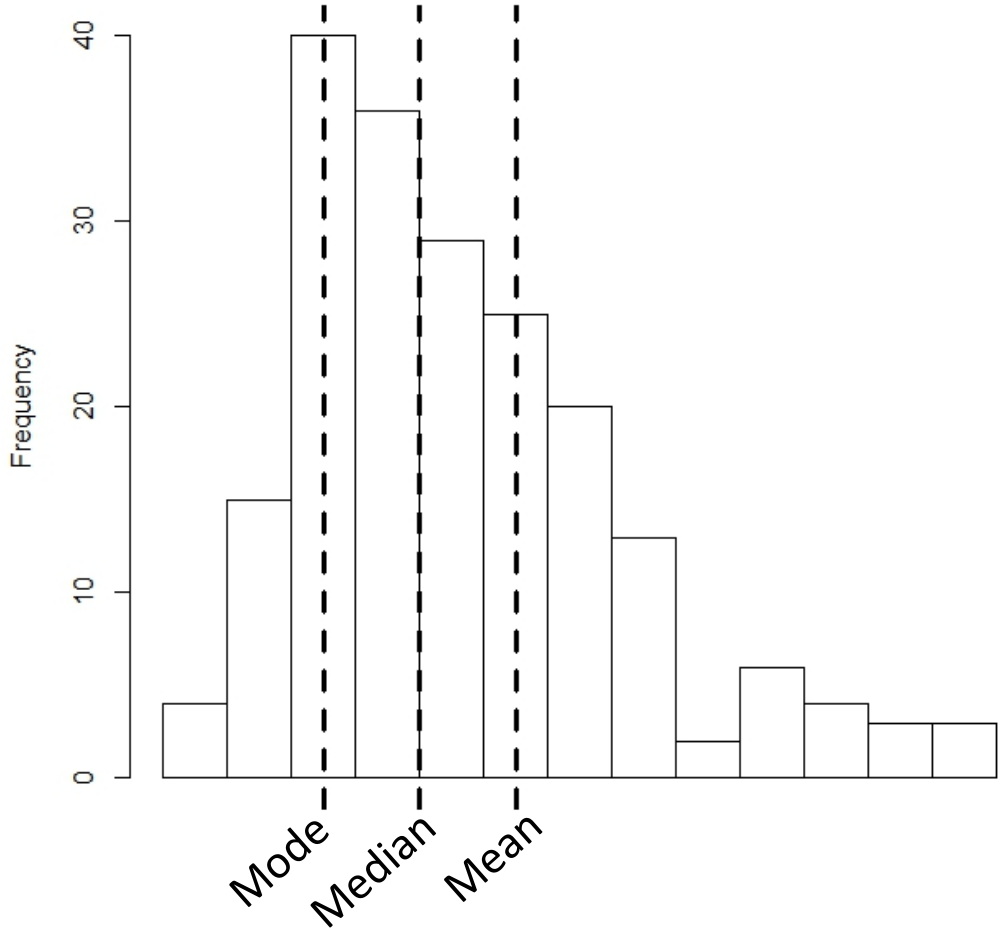
Left skewed.

Mean < Median.

Use median with IQR.

MACSI
Mathematics Applications Consortium for Science & Industry

Which to choose?



Right skewed.

Mean > Median.

Use median with IQR.

MACSI
Mathematics Applications Consortium for Science & Industry

Probability

Introduction

Probability is defined as the chance or likelihood of a particular outcome occurring and plays a key role in statistics.

Most people are familiar with the idea of probability, e.g. playing slot machines, card games, doing the lottery, betting.

Probability is also routinely used in insurance, investment banking and weather forecasting.

Introduction

Allows us to quantify how likely a particular (random) outcome is. It facilitates decision-making, which involves uncertainty.

- Should we invest in a company if there is a chance it will fail?
- What is the chance that a defective product on our production line will be detected?
- What is my risk of developing cancer given that I smoke?

Example

One in seven million chance

UNIVERSITY of Limerick statistician Dr Norma Bargary calculated that odds of this happening at one in seven million.

She said: "In Ireland, the chance of having a twin pregnancy naturally is around one in 80. However, the rate increases to around 1 in 65 if women are over the age of 35 and when including pregnancies that occur through fertility treatment.

Two-thirds of twins are non-identical, while a third are identical.

"Using the above figures, the chance of having one set of non-identical twins and one set of identical twins is approximately 1 in 20,000 and the chance of both sets of twins arriving on the same day (assuming that the chance of being born on every day of the year is equally likely) is about 1 in 7 million!" ■

MACSI
Mathematics Applications Consortium for Science & Industry

Example

What's the probability of winning Mega Millions?

Anthony Masters, a Statistical Ambassador from the Royal Statistical Society, says the probability of winning a Mega Millions jackpot is one in 300 million.

It's more likely…

- To flip a fair coin 28 times and get heads every time (one in 268 million)
- For two people to randomly dial the same eight digit phone number (one in 100 million).

And it's far more likely…

- To roll a six on a fair die 10 times in a row (one in 60 million)
- To die in a lightning strike (one in 114,000)

Of course, winning the lottery with a single ticket is very unlikely indeed, but of course the likelihood of someone winning the jackpot is high.

What is a probability?

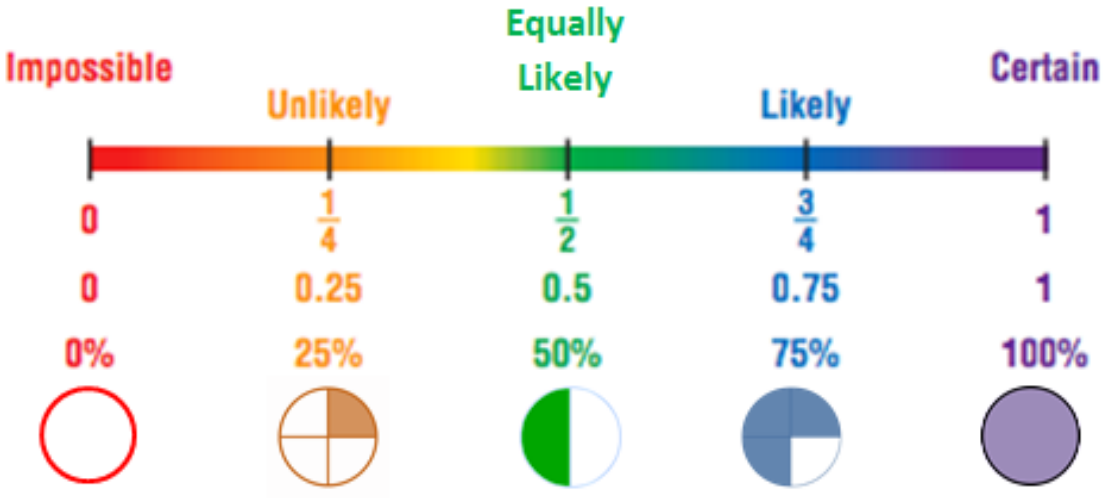
1. A probability is a value between 0 and 1.

Describes the chance or likelihood that a particular event will occur.

Typically expressed as a decimal (e.g. 0.02, 0.7, 0.5), as a fraction (e.g. 2/100, 70/100, 1/2), or as percentages (e.g. 2%, 70%, 50%).

What is a probability?

2. Probabilities measure how surprising an event is.



What is a probability?

3. Probabilities are used to model and analyse the outcomes of *random experiments*.

An experiment is any process that leads to a set of well defined results called outcomes.

Random experiment => each time you repeat the experiment, you can get a different result (outcome).

Assigning probabilities to events

There are 3 main approaches to assigning probabilities to events:

1. Classical: *k* possible outcomes. Each outcome is assigned the same probability $1/k$.

2. Subjective: probabilities assigned based on expert opinion.

3. Relative frequency (frequentist): Perform *n* identical experiments. Probability = *# times outcome is observed*/*n*.

Relative frequency

We have already seen how to calculate *relative frequencies* for data from a sample of size n.

Relative frequencies are also called *empirical probabilities* since they give us information about what outcomes are more or less likely than others.

As n gets larger and larger, this empirical probability will get closer and closer to the true probability.

Example

A machine fills containers with (pharmaceutical) tablets. Most containers are filled to the correct weight, however some can be underweight and others can be overweight. In order to determine the extent of the problem, a sample of 4000 containers was taken and the weight of each recorded. The following results were obtained.

Weight	*Num. of containers*
Underweight	*100*
Satisfactory	*3600*
Overweight	*300*
Total	*4000*

MACSI
Mathematics Applications Consortium for Science & Industry

Example

1. *What is the probability that a container is underweight?*

   Pr(Underweight) = 100/4000 = 0.025 (2.5%)

2. *What is the probability that a container is overweight?*

   Pr(Overweight) = 300/4000 = 0.075 (7.5%)

3. *What is the probability that a container is either underweight or overweight?*

   Pr(Over and/or under) = 0.025 + 0.075 = 0.1 (10%)

Combining probabilities

Often useful to calculate the probability of several events occurring.

Need some simple rules to do this:

1. Complement rule;
2. Addition rule;
3. Multiplication rule;
4. Bayes' theorem.

Useful resources

Probability cheatsheet available here: http://www.wzchen.com/probability-cheatsheet.

Mind on Statistics by Jessica Utts available here: https://www.cengage.com/c/mind-on-statistics-5e-utts/9781285463186PF/.

Resources on learning probability available here: https://projects.iq.harvard.edu/stat110/home.

MACSI
Mathematics Applications Consortium for Science & Industry

Random variables

A random variable is a numerical variable whose value depends on the outcome of a (random) experiment.

Outcomes of experiments are not necessarily numbers, e.g. heads or tails.

Useful to represent these outcomes using numbers.

Random variables

Random experiment => outcome can change each time we carry out the experiment.

A random variable (*Y*) is a number that describes every possible outcome of an experiment and whose value can change each time the experiment is carried out.

Random variables

Random variables can be:

1. discrete (can only have values that are whole numbers), or
2. continuous (can take any value in an interval).

Discrete random variables

Experiment	**Random variable *Y***	**Possible values of *Y***
Pick a number between 1 and 10	*Y* = number chosen	*Y* = 1, …, 10
Count the number of employees absent on Monday	*Y* = number absent	*Y* = 0, 1, 2, …
Count the number of scrapped items from production line	*Y* = number of items scrapped	*Y* = 0, 1, 2, …

MACSI
Mathematics Applications Consortium for Science & Industry

Discrete probability distributions

Can calculate the probability that each individual value of Y will occur.

The values of the random variable and the corresponding probabilities are called a *probability distribution*.

Y	y_1	y_2	…	y_k
Prob.	$\Pr(Y = y_1)$	$\Pr(Y = y_2)$	…	$\Pr(Y = y_k)$

Discrete probability distributions

The values $\Pr(Y = y_k)$ must satisfy the following conditions:

- $\Pr(Y = y_k) \geq 0$ (All probabilities are positive or 0.)
- $\sum_k \Pr(Y = y_k) = 1$ (The probabilities in the distribution must sum to 1.)

Example

The RSA collects data on the number of road deaths occurring each year in the 26 counties in the Republic of Ireland. The data and probability distribution for 2012 is as follows:

Y = # deaths	Frequency	$Pr(Y = y)$
0	2	2/26 = 0.08
1	1	1/26 = 0.04
2	3	3/26 = 0.12
3	3	3/26 = 0.12
4	3	3/26 = 0.12
5	3	3/26 = 0.12
6	0	0/26 = 0.00
7	5	5/26 = 0.19
8+	6	6/26 = 0.24
Total	26	1.00

Example

1. *What is the probability that a county has between 1 and 5 deaths?*

   $\Pr(1 \leq Y \leq 5) = 0.04 + 0.12 + 0.12 + 0.12 + 0.12 = 0.52$ (52%)

2. *What is the probability that a county has 7 deaths?*

   $\Pr(Y = 7) = 0.19$ (19%)

3. *What is the probability that a county has 2 deaths or fewer?*

   $\Pr(Y \leq 2) = 0.08 + 0.04 + 0.12 = 0.24$ (24%)

Discrete probability distributions

Common distributions for modelling discrete random variables:

1. Bernoulli distribution:- 1 trial with 2 possible outcomes (“success” or “failure”) where Pr(success) = *p*.

2. Binomial distribution:- *n* independent trials with 2 possible outcomes (“success” or “failure”) where Pr(success) = *p.* Gives probability of *r* successes in *n* trials. For example, the number of faulty products in a sample of size 10.

Discrete probability distributions

3. Poisson distribution:- used to model independent events that occur at constant rate in a given time interval. Gives probability of *r* events occurring in that interval. For example, the number of emails received in 1 hour.

4. Etc.

Continuous random variables

Examples so far have been for discrete random variables. What about continuous random variables?

More difficult to write down the probability for a continuous random variable since typically have many more unique numbers.

Therefore more interested in calculating the probability that a continuous random variable has a value in a certain *range.*

Continuous random variables

Experiment	**Random variable *Y***	**Possible values of *Y***
Return on investment in 1 year	Y = return in €	$-\infty \leq Y \leq \infty$
Time to failure of a machine part	Y = lifetime in hours	$Y \geq 0$
Fill a soft drink can	Y = number of ml in can	$0 \leq Y \leq 330$

MACSI
Mathematics Applications Consortium for Science & Industry

Continuous probability distributions

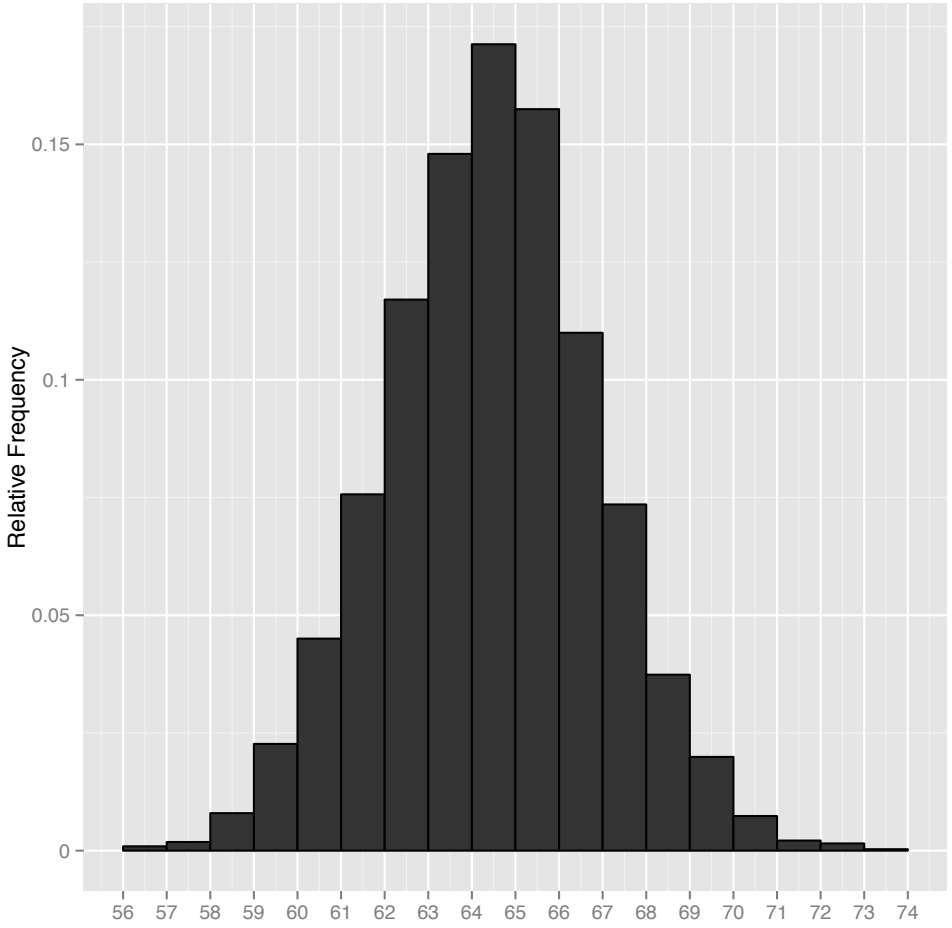
For continuous random variables, we can construct a frequency distribution and a histogram of our sample data.

There is a link between the histogram and calculating probabilities for a continuous random variable.

Example

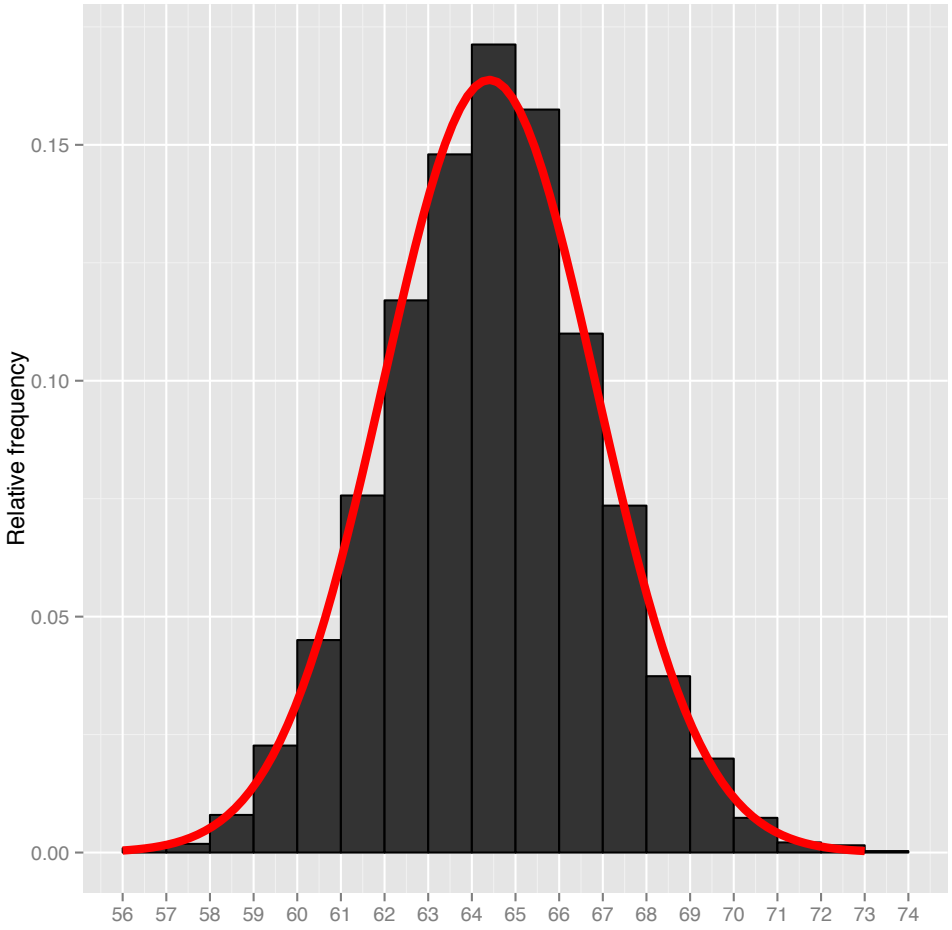
Length (mm)	Frequency	Rel. freq.
56 - 56.99	3	0.0009
57 - 57.99	6	0.0018
58 - 58.99	26	0.0080
59 - 59.99	74	0.0227
…	…	…
69 - 69.99	65	0.0199
70 - 70.99	24	0.0074
71 - 71.99	7	0.0021
72 - 72.99	5	0.0015
73 - 73.99	1	0.0003

Example



MACSI
Mathematics Applications Consortium for Science & Industry

Example



MACSI
Mathematics Applications Consortium for Science & Industry

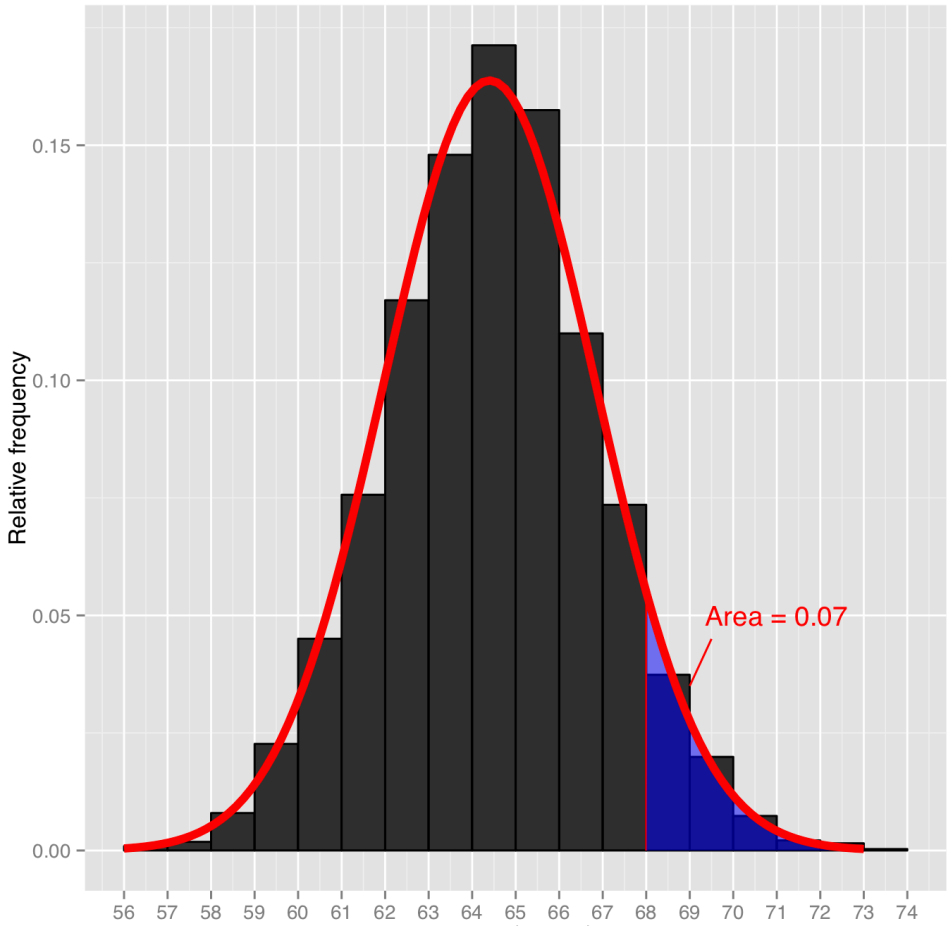
Density curves

The red line is called a *density curve*.

The area of each bar of the histogram gives the probability of having a length in a particular range.

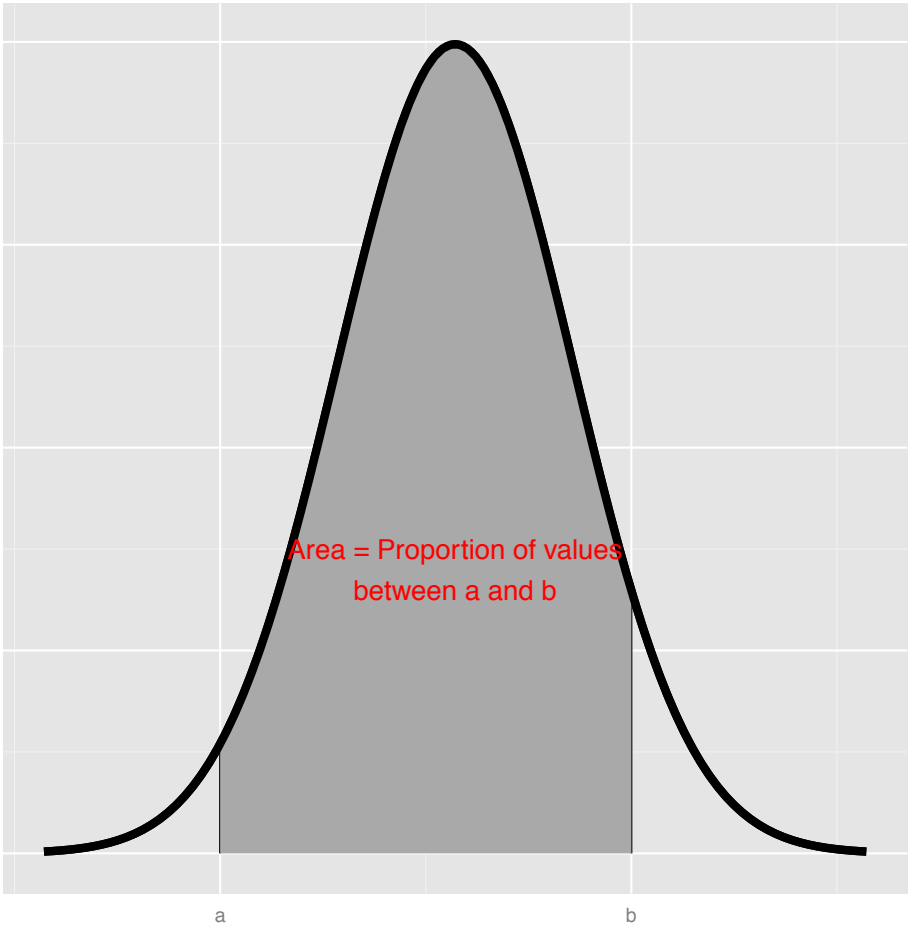
This area is equivalent to the area under the density curve over the same range.

Example



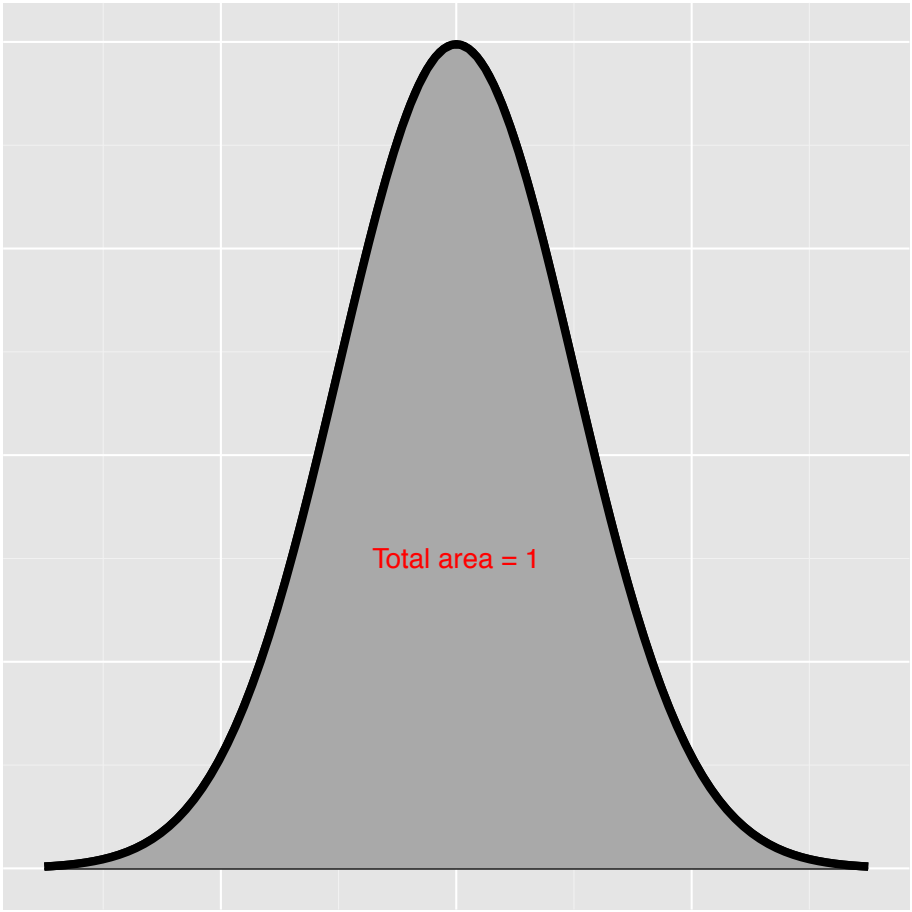
MACSI
Mathematics Applications Consortium for Science & Industry

Properties of density curves



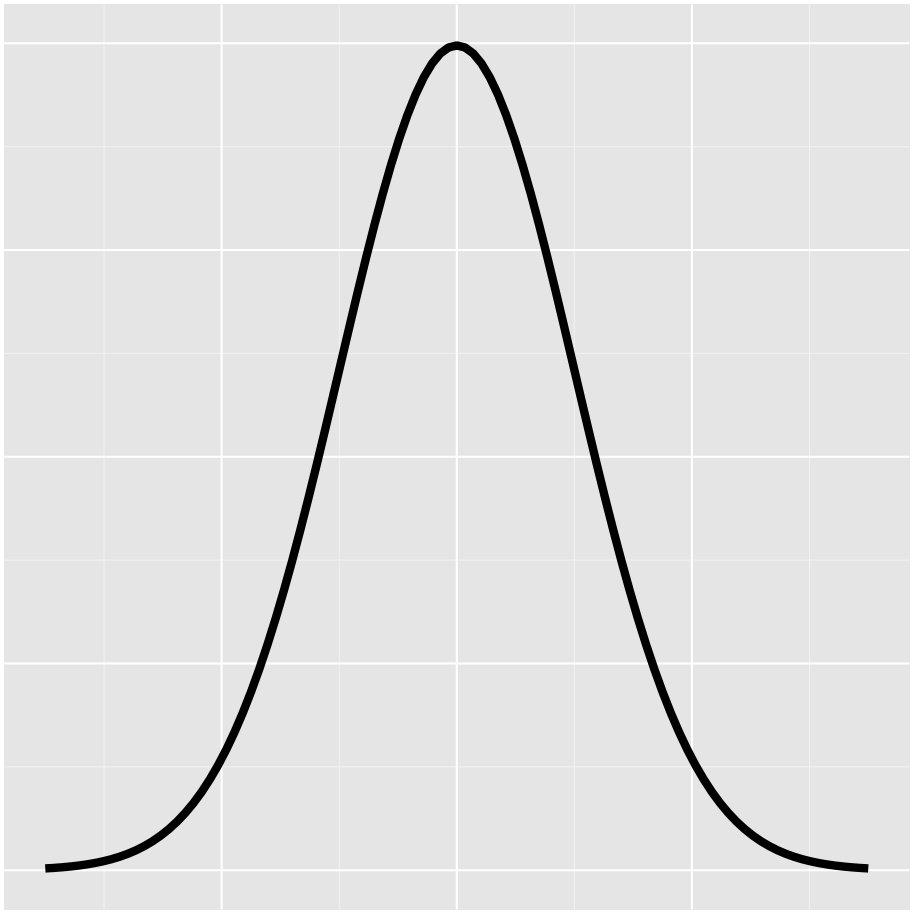
The probability of a value being between *a* and *b* = the area under the density curve between *a* and *b.*

Properties of density curves



The total area under the density curve = 1.

The Normal distribution



Many variables share an important characteristic: their distributions have (roughly) the shape of a Normal curve.

The Normal distribution

Is used widely in statistical modelling (hypothesis testing, regression, statistical process control, etc.).

Is described by the following formula:

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(y-\mu)^2}{2\sigma^2}\right\}$$

where *f(y)* is called the *density function* and is used to calculate the probability that *Y* has a value in a specified range.

MACSI
Mathematics Applications Consortium for Science & Industry

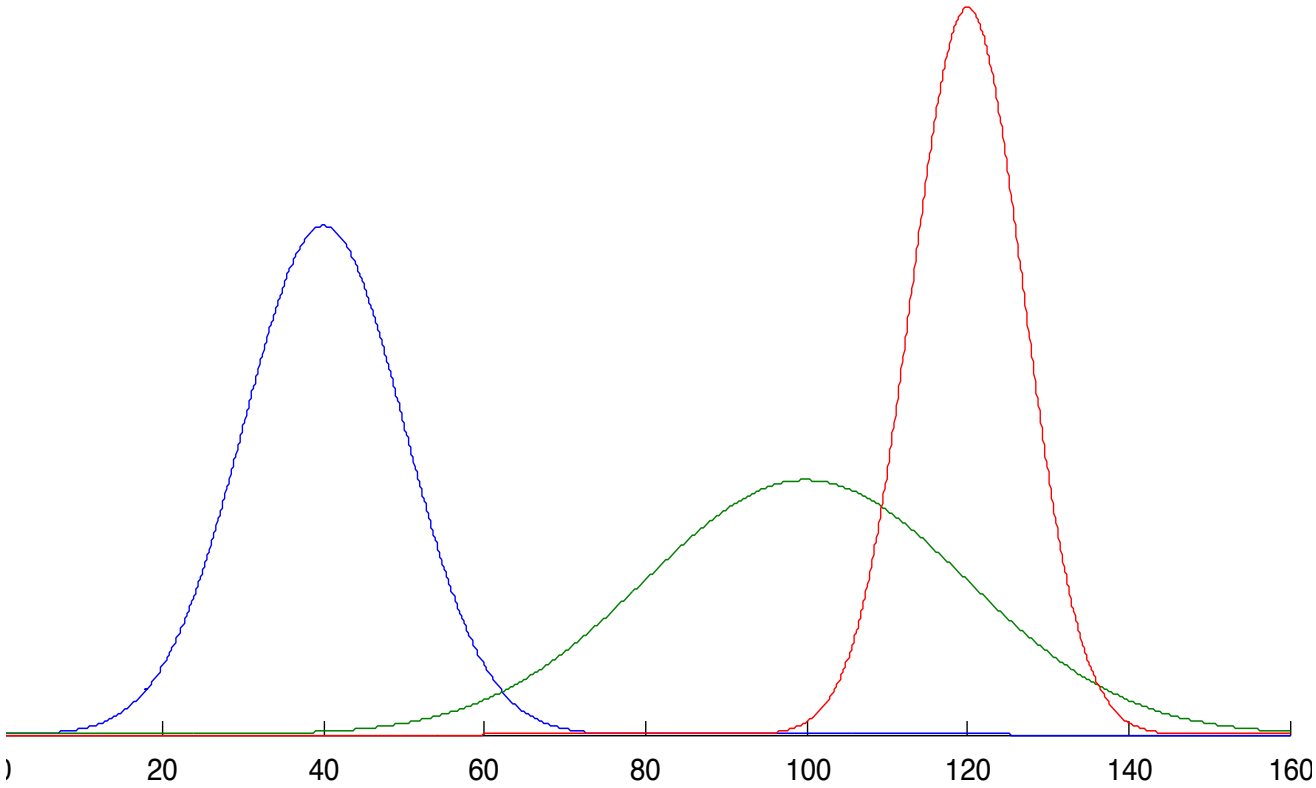
The Normal distribution

Is symmetric and has a bell-shape that is characterised by the mean μ, and the standard deviation σ.

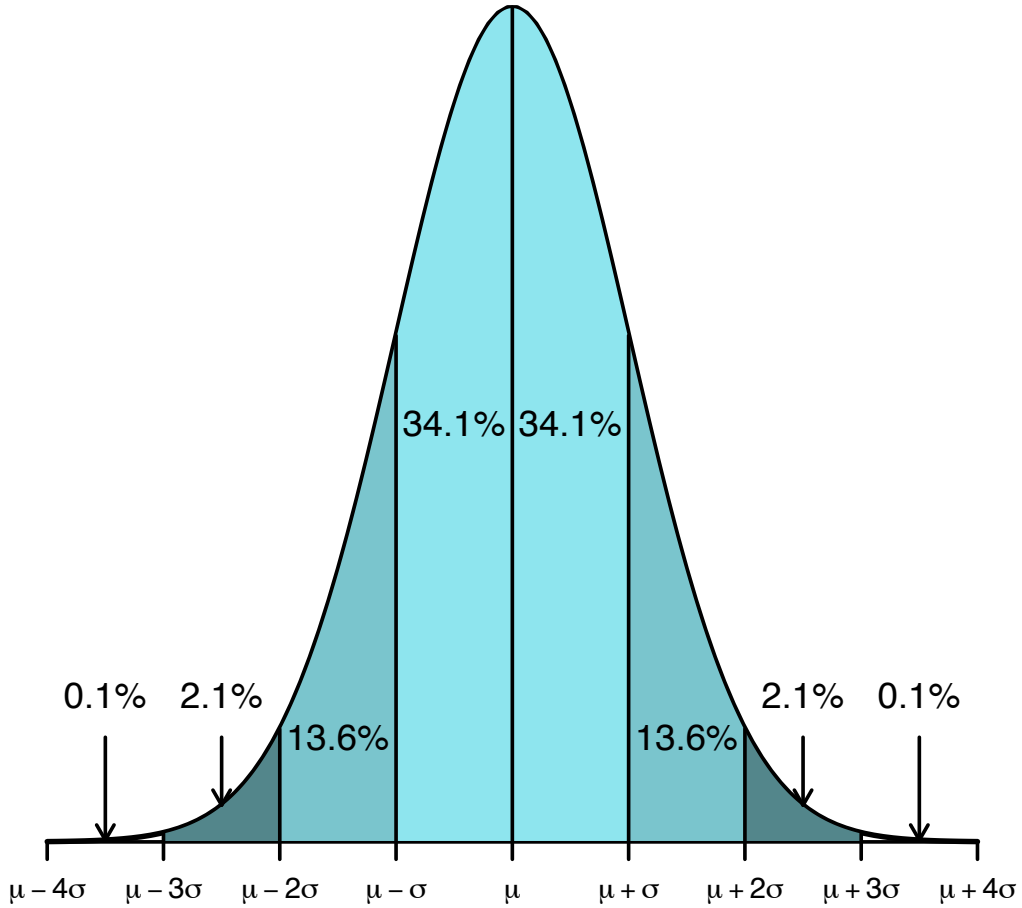
The mean governs where the peak of the distribution is. The standard deviation governs how spread out the distribution is.

Often write $Y \sim N(\mu, \sigma^2)$.

The Normal distribution

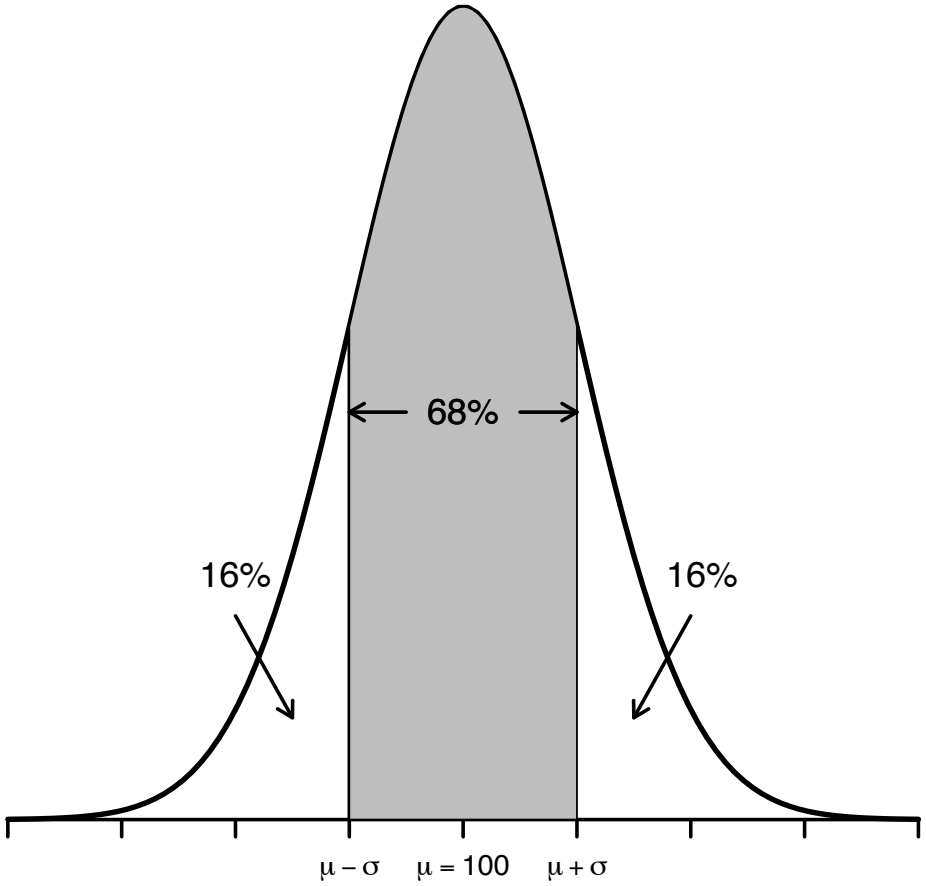


Properties of the Normal distribution



Example

IQ scores have a Normal distribution with $\mu = 100$ and $\sigma = 15$.

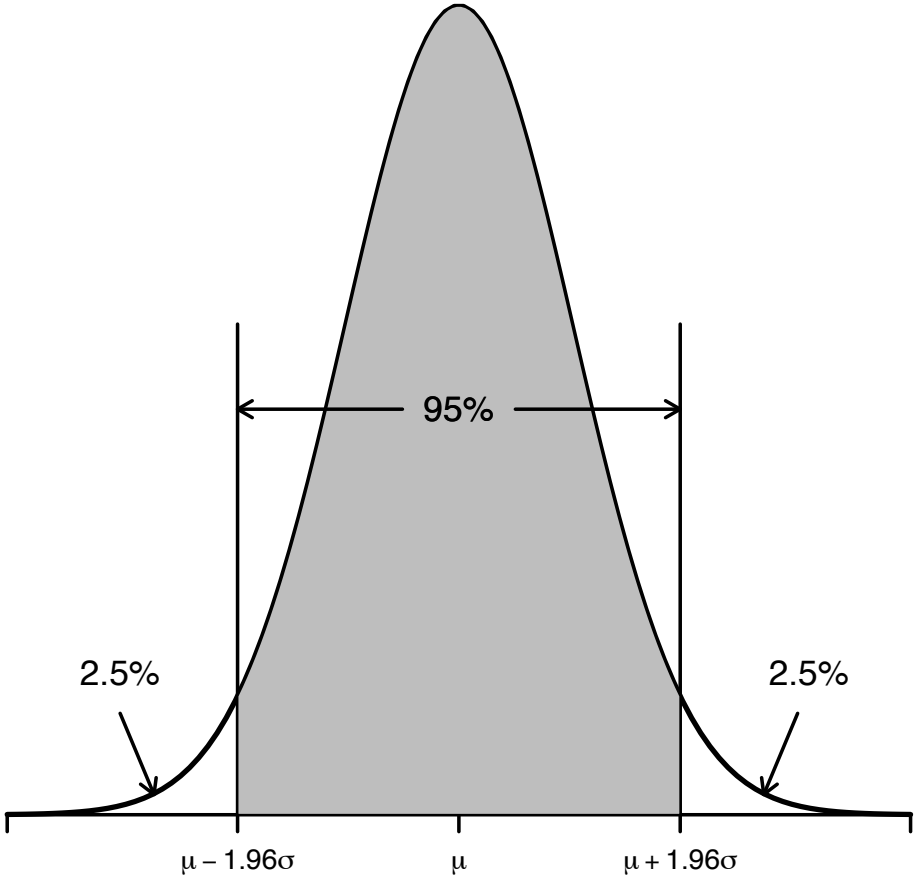


68% of people have IQ scores in the range:

$100 \pm 15 = [85, 115]$

Example

IQ scores have a Normal distribution with $\mu = 100$ and $\sigma = 15$.



95% of people have IQ scores in the range:

$100 \pm 2*15 = [70, 130]$

Areas under the Normal curve

We would like to calculate the probability that a Normally distributed random variable has a value in a given interval.

Use the density function formula to calculate the area under the Normal curve.

Areas correspond to probabilities and have been tabulated in statistical tables (and software).

Standard Normal distribution

There are an infinite number of Normal curves, with different means and standard deviations.

The standard Normal distribution has a mean of 0 and standard deviation of 1.

Standard Normal distribution

Convert any Normal distribution to the standard Normal using the re-scaling:

$$Z = \frac{Y - \mu}{\sigma} \sim N(0, 1)$$

which measures the number of standard deviations a value Y is from the mean.

Example

Pulse rates (Y) of people have a Normal distribution with mean 75 bpm and standard deviation of 8 bpm. Calculate:

1. *$Pr(Y \geq 75)$*
2. *$Pr(Y \leq 63)$*
3. *$Pr(61 \leq Y \leq 69)$*
4. *90% of people have a pulse rate below what value?*

Example

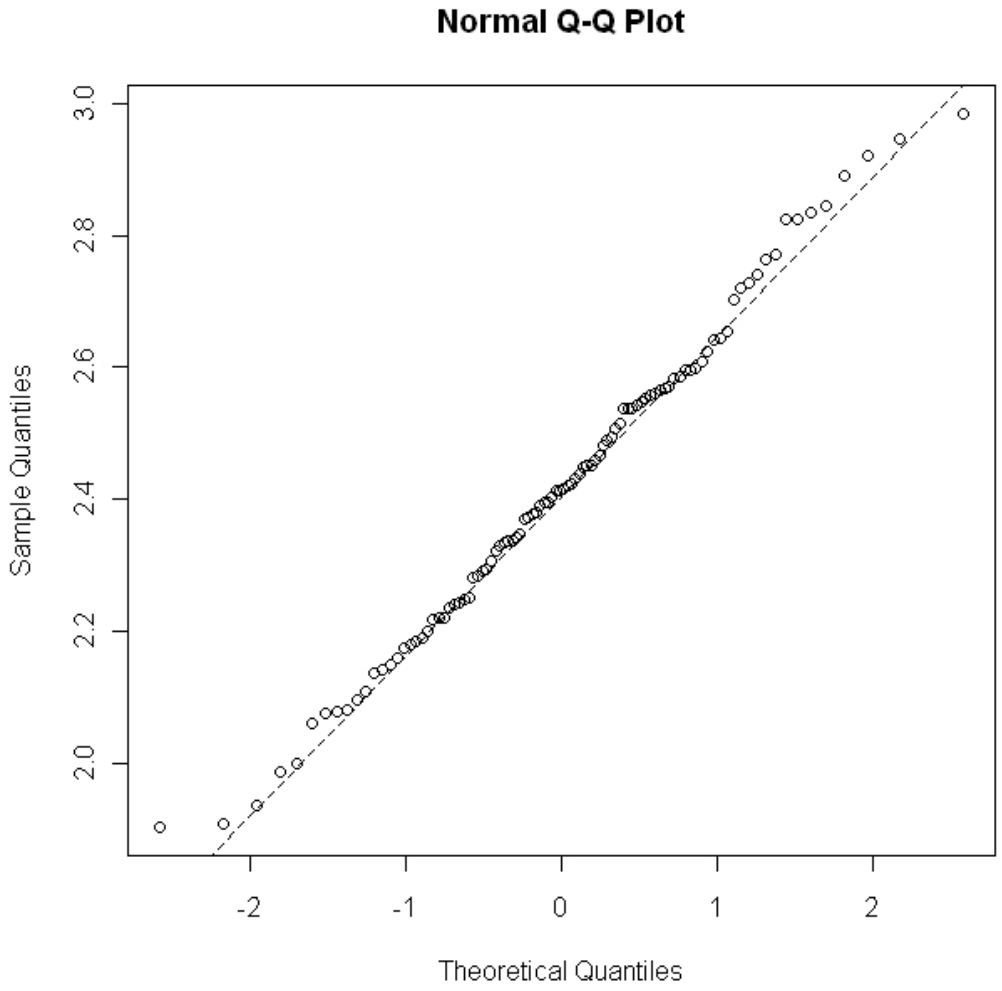
1. $Z = (75 - 75)/8 = 0 \Rightarrow$
   $Pr(Y \geq 75) = Pr(Z \geq 0) = 0.5$

2. $Z = (63 - 75)/8 = -1.5 \Rightarrow$
   $Pr(Y \leq 63) = Pr(Z \leq -1.5) = 0.07$

3. $Z_1 = (61 - 75)/8 = -1.75$
   $Z_2 = (69 - 75)/8 = -0.75 \Rightarrow$
   $Pr(61 \leq Y \leq 69) = Pr(-1.75 \leq Z \leq -0.75) = 0.19$

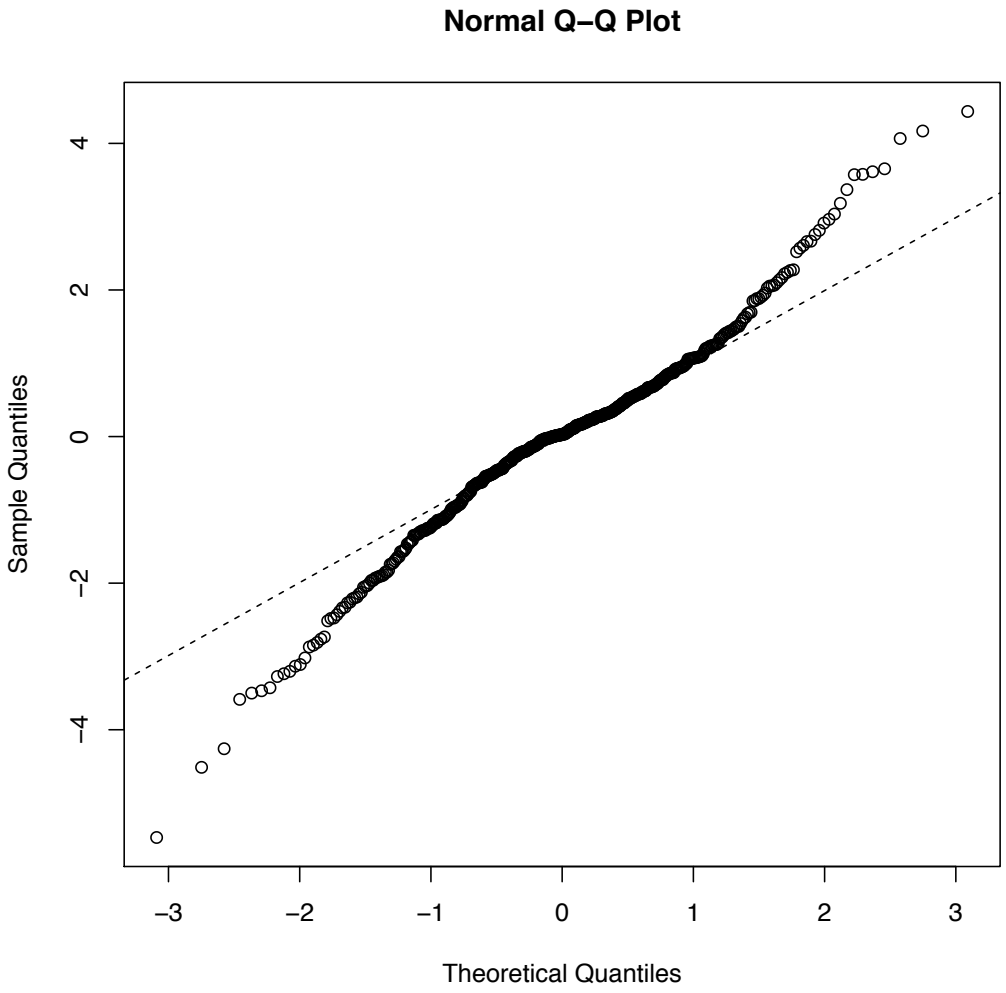
4. $1.28 = Z = (Y - 75)/8 \Rightarrow Y = 85.24$

Normal probability plot



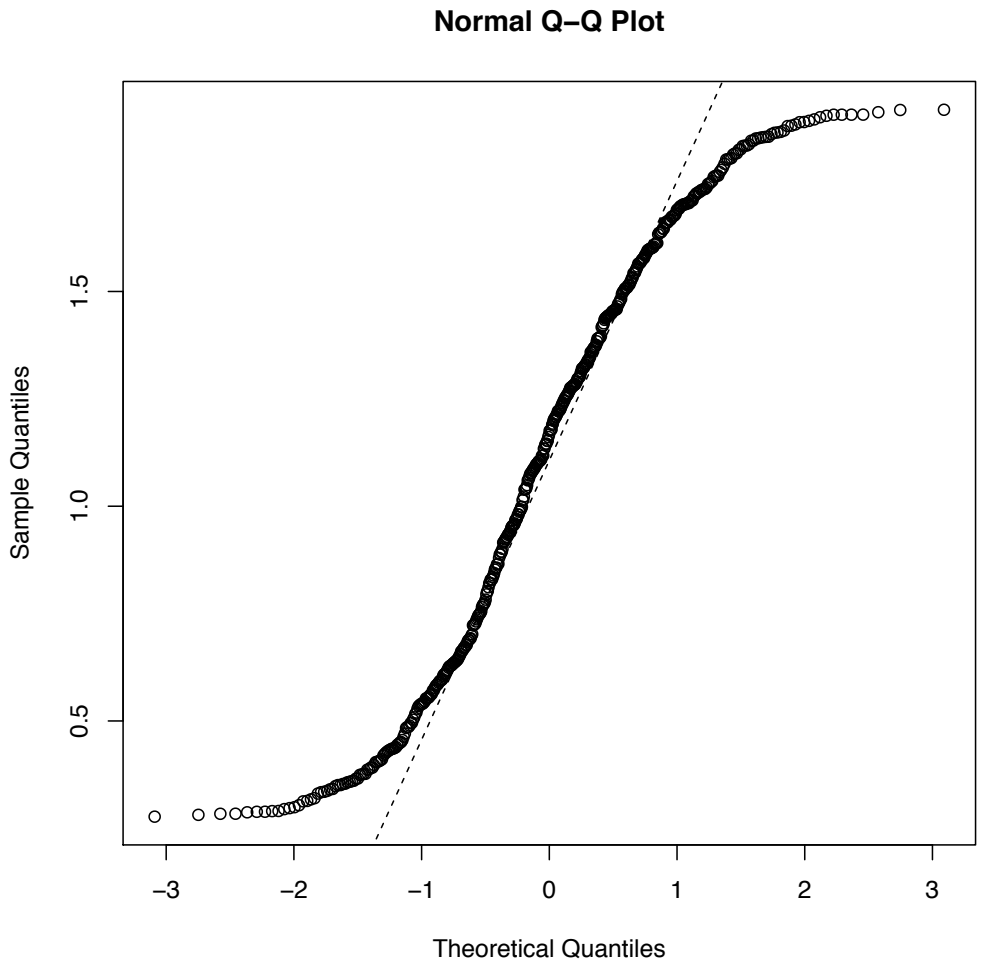
MACSI
Mathematics Applications Consortium for Science & Industry

Normal probability plot



MACSI
Mathematics Applications Consortium for Science & Industry

Normal probability plot



MACSI
Mathematics Applications Consortium for Science & Industry

Transforming variables

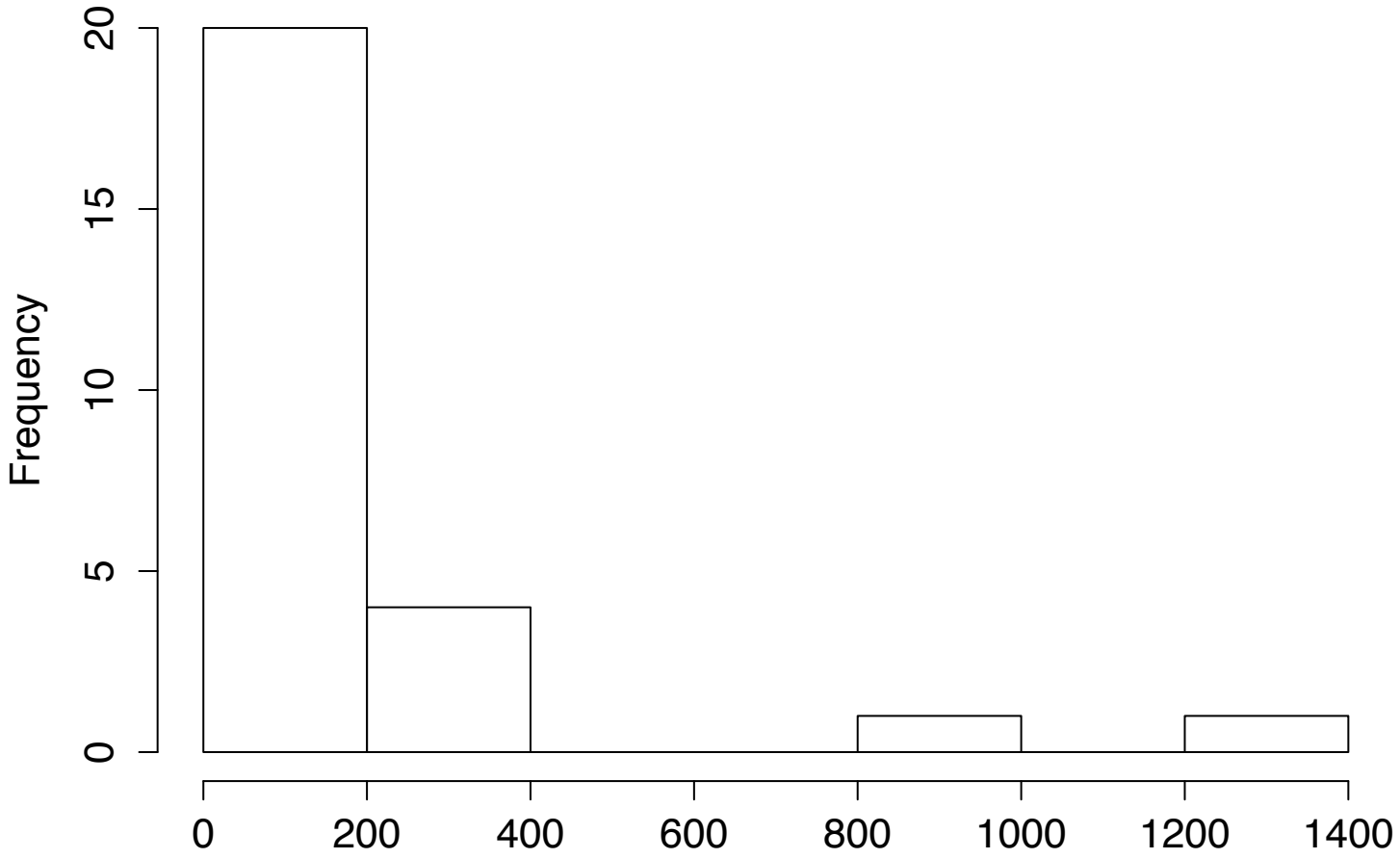
Sometimes data are not exactly Normally distributed.

However, many methods in statistics require that the data have a Normal distribution.

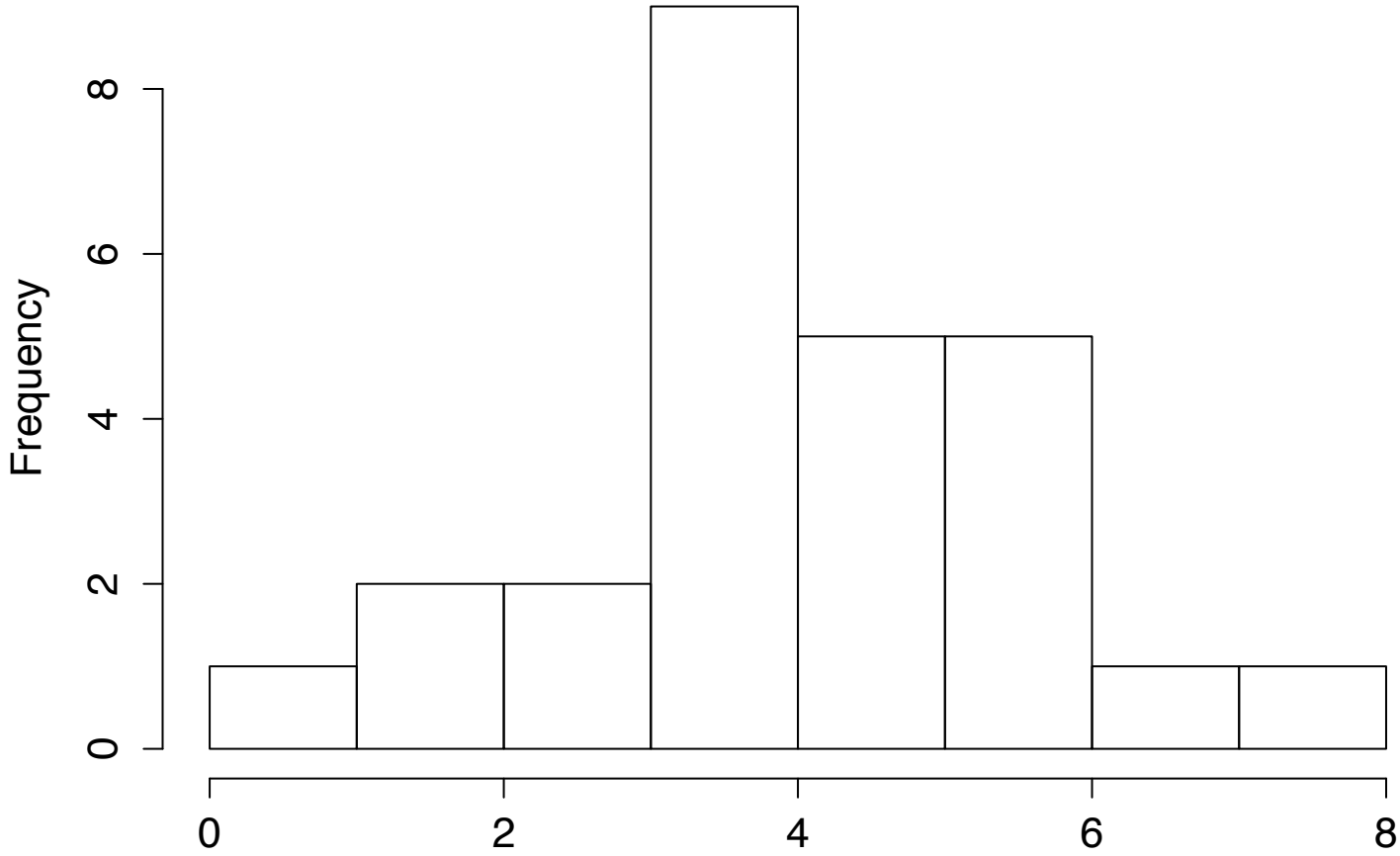
Can transform our data to look like a Normal distribution using for example:

- Logarithm,
- Square root,
- Box-Cox transform.

Transforming variables



Transforming variables (ln)



Continuous probability distributions

Common distributions for modelling continuous random variables:

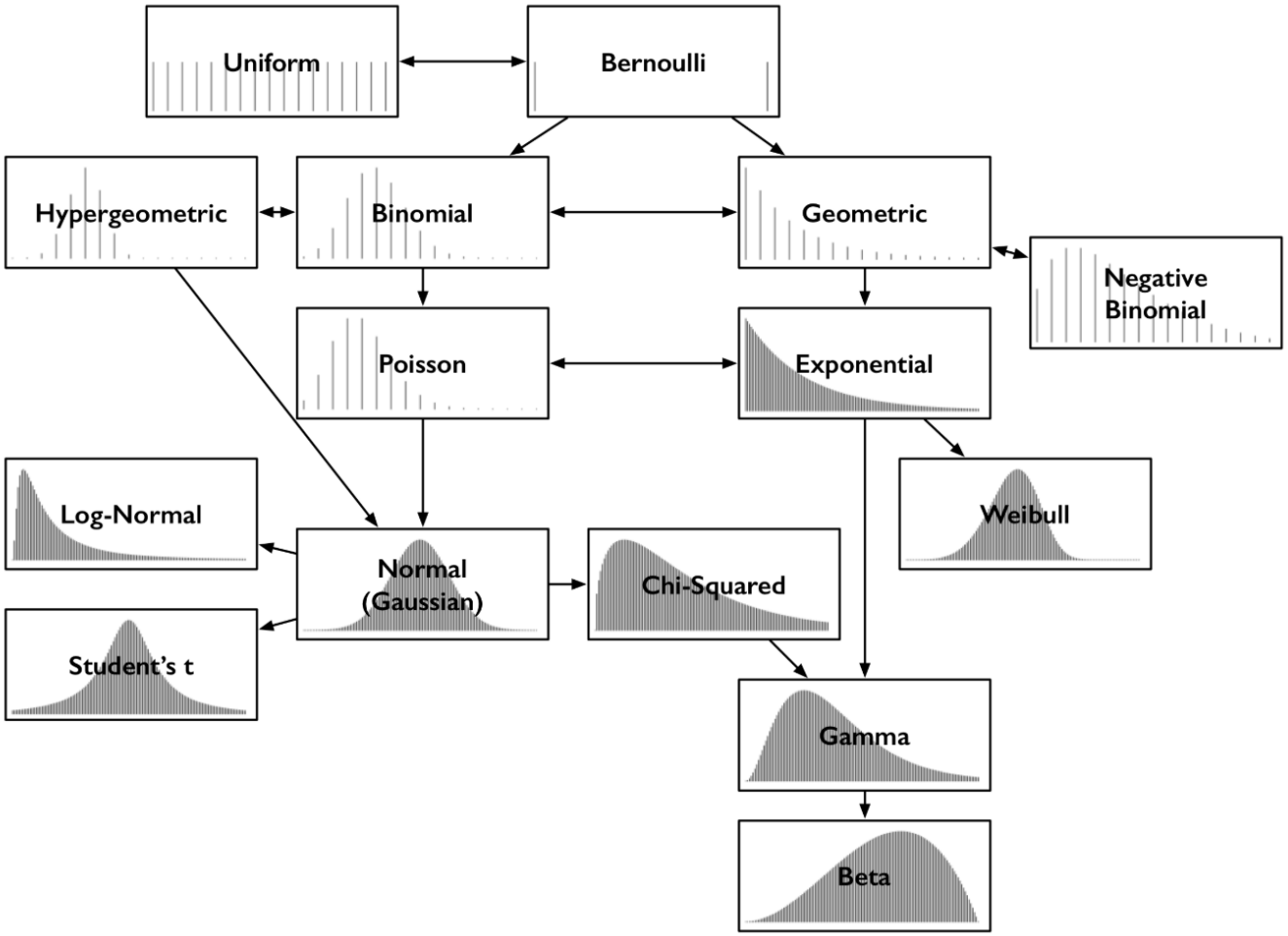
1. t distribution:- similar to Normal distribution but used to model data with smaller sample sizes.

2. Chi-squared distribution:- arises routinely in hypothesis testing and for modelling variance estimates.

3. Exponential distribution:- used to model the time between events in a Poisson process.

4. Etc.

Relationships between distributions



Statistical Inference

Introduction

Usually interested in answering a question they are curious about.

Generally these questions extent to a large population.

A sample of data is collected and used to answer these question(s) about the population.

Introduction

(Good) researchers translate scientific questions about large populations into questions about specific summary statistics, e.g. mean, proportion.

In statistics, this usually requires us to formulate a question about a *population parameter*.

Once this step is completed, data (appropriate for making conclusions about the parameter of interest) should then be collected using a random sample.

Estimation

The estimate of the population parameter of interest is called a *point estimate.*

The accuracy of the point estimate depends on the sample size (*n*) and how well the sample represents the population.

Want to give a measure of how precise this estimate is and an interval (called a *confidence interval*) in which we are confident the population parameter lies.

Typical parameters of interest

Name	Parameter	Statistic
One proportion	p	$\hat{p}$
Two proportions	$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$
One mean	μ	$\bar{y}$
Two means (independent)	$\mu_1 - \mu_2$	$\bar{y}_1 - \bar{y}_2$
Two means (dependent/ paired)	μ_d	$\bar{y}_d$

Sampling distributions

NB! Aim is to make statements about the *population* using a *sample*.

Important to establish the relationship between sample statistics and population parameters.

The specific sample we obtain is just one of many that could have been drawn from the population.

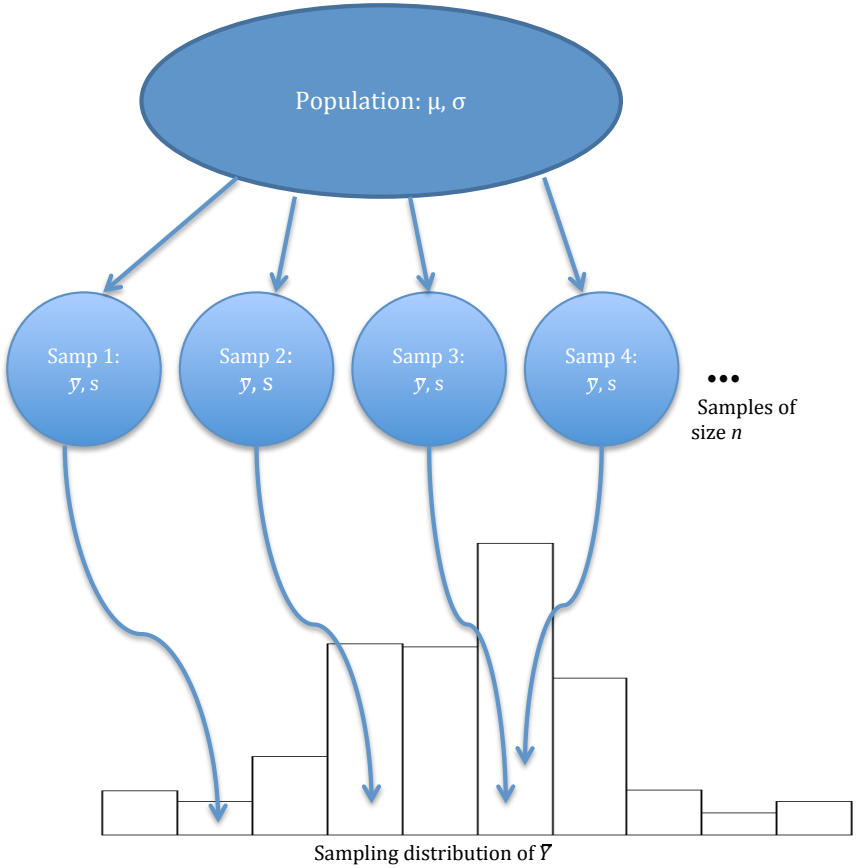
Sampling distributions

If many samples of size *n* are taken from a population and the same statistic (e.g. the mean) is calculated for each sample, the values of these statistics will vary from sample to sample.

Thus the sample statistic is a random variable => can be described by a probability distribution.

Distribution is called the *sampling distribution*.

Sampling distribution of the mean



Central Limit Theorem

The sampling distribution of the mean has some useful properties that are governed by a very important theorem in statistics called the *Central Limit Theorem.*

For large samples, the sampling distribution of the mean has:

- a Normal distribution
- mean = μ
- standard deviation = $\frac{\sigma}{\sqrt{n}} = SE(\bar{y})$

Written as $\bar{Y} \sim N\left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right)$

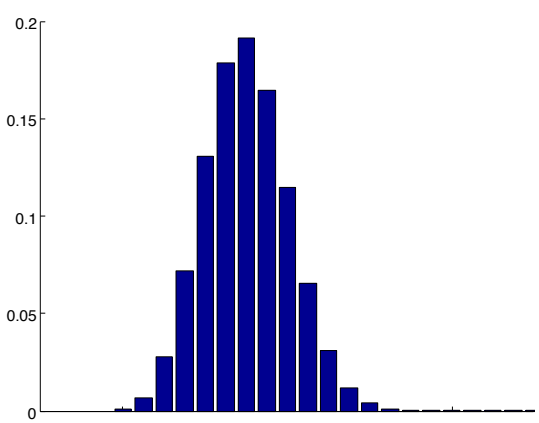
Central Limit Theorem

If the sample size is *sufficiently large* the sampling distribution of the mean will have a Normal distribution, irrespective of the distribution of the data in the population.

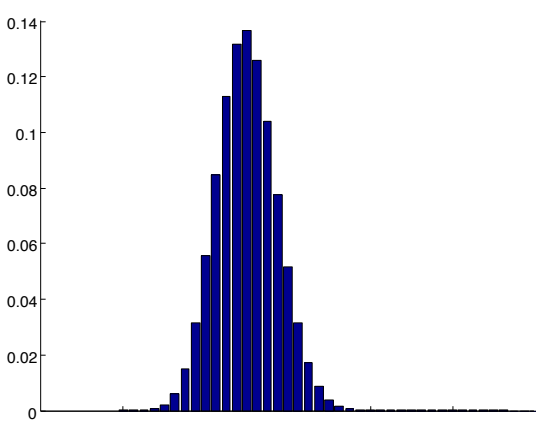
If the data are Normally distributed in the population, the sampling distribution of the mean will *always* have a Normal distribution, *irrespective* of the sample size.

Effect of sample size

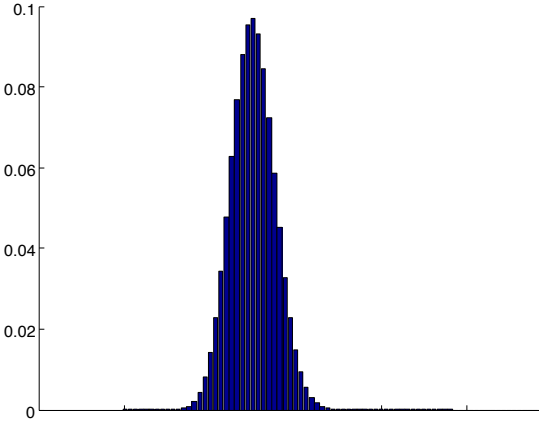
The sampling distribution varies with sample size:



n = 20



n = 40



n = 80

MACSI
Mathematics Applications Consortium for Science & Industry

Statistical inference

We want to understand something about populations using information from a random sample.

Since we are trying to infer something about the population but only have a sample => have uncertainty.

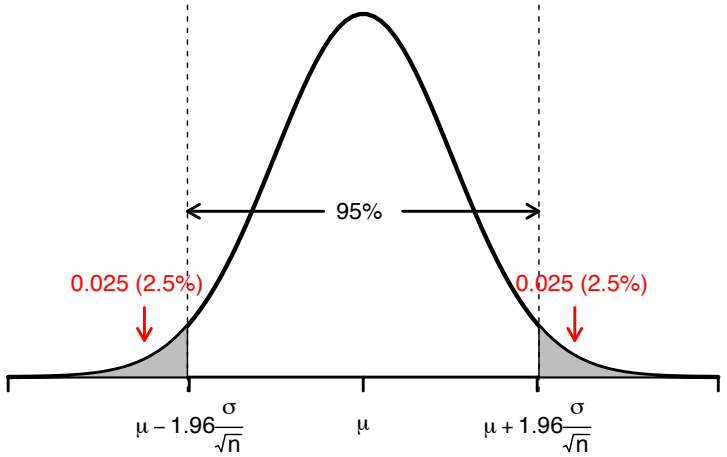
To quantify that uncertainty we use the sampling distribution to construct *confidence intervals*.

Confidence intervals

We have that $\bar{Y} \sim N\left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right)$

Then 95% of sample means will lie in the range:

$$\mu \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$$



Confidence intervals

Usually don't know μ => use $\bar{y}$.

Then a 95% CI for μ is:

$$\bar{y} \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$$

A 90% CI for μ is:

$$\bar{y} \pm 1.645 \times \frac{\sigma}{\sqrt{n}}$$

A 99% CI for μ is:

$$\bar{y} \pm 2.58 \times \frac{\sigma}{\sqrt{n}}$$

General procedure

In general, the limits for a 100(1 - α)% confidence interval can be found by using the relevant $Z_{\alpha/2}$ value from the standard Normal distribution.

The general formula for a CI for μ is given by:

$$\bar{y} \pm Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

Interpreting a CI

Under repeated sampling, 95% (or 90% or 99%) of CIs will contain the true population mean.

If we take 100 samples of the same size from the same population and calculate a 95% CI for each, we expect that 95 of those CIs will contain the true mean.

Width of a CI

The width of a CI is called the *precision.*

The narrower a CI is, the more precise it is.

There are two ways to increase the precision:

- Decrease the confidence level.
- Increase the sample size *n.*

Example

A research firm conducts a survey to determine the mean amount spent on cigarettes per week. A sample of 49 smokers gave an average amount spent per week of €20 and population standard deviation of €5. Construct a 95% CI for the true mean.

$$\bar{y} \pm 1.96 \times \frac{\sigma}{\sqrt{n}} =$$

$$20 \pm 1.96 \times \frac{5}{\sqrt{49}} =$$

[€18.60, €21.40]

CIs when σ is unknown

If σ is unknown, we then use *s* as the point estimate in the CI calculation.

However, we now need to worry about how reliable the estimate *s* is.

s will be more reliable for large samples than small samples.

CIs when σ is unknown

Therefore there will be extra uncertainty around the value of *s* when *n* is small.

Thus the previous CI formula may be inaccurate when *s* is substituted for σ.

We need to take this into account.

MACSI
Mathematics Applications Consortium for Science & Industry

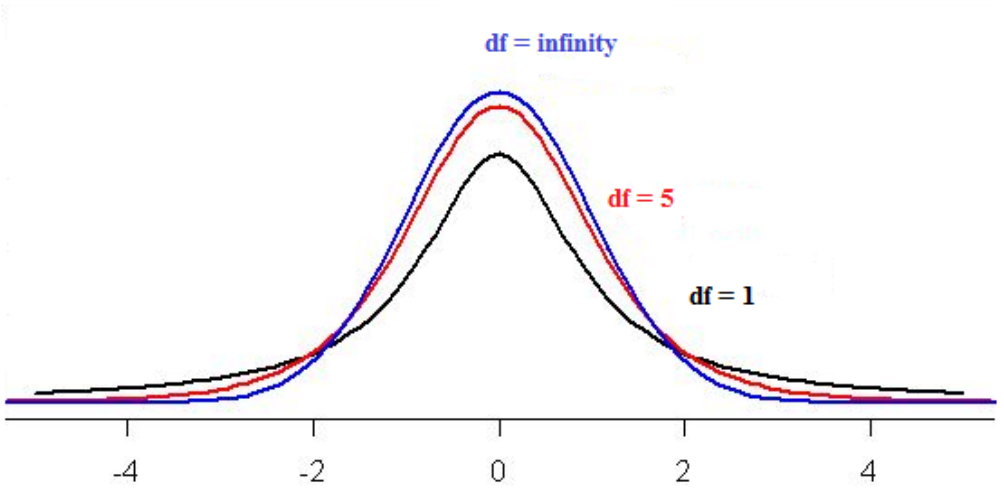
CIs when σ is unknown

When σ is unknown, we replace values from the Normal distribution with values from the *t*-distribution.

The *t*-distribution was first studied by William Sealy Gosset (Student) who worked in the Guinness brewery.

He found that if we change the multiplier in the CI formula, the interpretation of the CI will be preserved for small samples when *s* is used in place of σ.

The *t*-distribution



- Has the same shape as the Normal distribution.
- Flatter and more spread out (fatter tails).
- Shape is characterised by its *degrees-of-freedom = n – 1.*
- As *n* increases, the *t*-distribution looks more and more like a Normal distribution.

CIs when σ is unknown

The general formula for a CI for μ when σ is unknown is then given by:

$$\bar{y} \pm t_{\alpha/2, df=n-1} \times \frac{s}{\sqrt{n}}$$

Assumptions required

One or both of the following must hold for the *t* interval to be valid.

1. Large sample size => ok in general. If there are extreme outliers or skewness, may need alternative approach.

2. Small sample size => data must be drawn from a Normal distribution in the population.

Must check these assumptions using QQ-plots, boxplots, etc.

Example

In the Credit Risk dataset, we want to construct a 90% CI for the Total Bill Amount. We can do this very easily in R using:

```
> t.test(dat$total_bill_amt, conf.level=0.90)

        One Sample t-test

data:  dat$total_bill_amt
t = 99.173, df = 19692, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
 145079.5 149973.4
sample estimates:
mean of x
 147526.4
```

MACSI
Mathematics Applications Consortium for Science & Industry

General procedure for CIs

All confidence intervals have the same general structure:

Estimate ± Value from tables x SE(Estimate)

CIs for typical parameters of interest

Parameter	CI
p	$\hat{p} \pm Z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
$p_1 - p_2$	$(\hat{p}_1 - \hat{p}_2) \pm Z_{\alpha/2} \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$
μ	$\bar{y} \pm t_{\alpha/2} \times \frac{s}{\sqrt{n}}$
$\mu_1 - \mu_2$	$(\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2} \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
μ_d	$\bar{y}_d \pm t_{\alpha/2} \times \frac{s_d}{\sqrt{n_d}}$

Hypothesis testing

Hypothesis testing is a decision-making process used to evaluate claims about population(s) of interest.

You make a claim (hypothesis).

You collect data to test that claim.

Based on the data you collect, is there enough evidence to make you think your claim is wrong? Or do you lack the evidence to dispute it?

Hypothesis testing

The HSE is interested in whether there is a difference between the proportion of people being vaccinated for flu in large versus small nursing homes.

UL wants to test the claim that the average number of sports scholarships for males is greater than that of females.

A process engineer wants to investigate whether there is a difference between the strength of raw materials from two different suppliers.

Setting up a hypothesis test

Each research question is translated into a question about some parameter of interest (e.g. mean, proportion).

Can then view each question as a choice between two competing hypothesis statements.

Hypothesis 1: The average strength of material from supplier A and supplier B is the same.

Hypothesis 2: The average strength of material from supplier A and supplier B is different.

Setting up a hypothesis test

In hypothesis testing these two possible answers are called:

1. The *null hypothesis* – H_0.
2. The *alternative hypothesis* – H_A.

The objective is to use the sample information to decide between them.

Setting up a hypothesis test

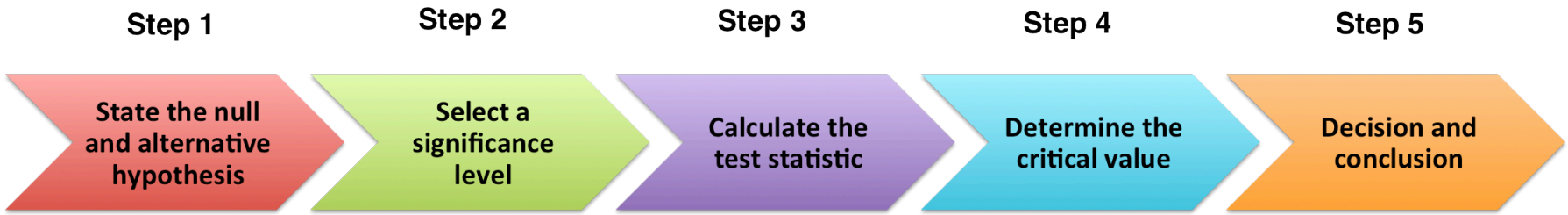
Start by assuming H_0 is true and ask "Do we have sufficient evidence to reject it in favour of H_A?"

If the data are consistent with H_0 we cannot reject H_0.

If the data are incompatible with H_0 this is taken as evidence in favour of H_A.

MACSI
Mathematics Applications Consortium for Science & Industry

Setting up a hypothesis test



MACSI
Mathematics Applications Consortium for Science & Industry

Step 1: Write down H_0 and H_A

Q: Is the average strength of materials from supplier A and supplier B different?

H_0: The average strength is the same.

H_A: The average strength is different.

In mathematical notation we can write:

$$H_0: \mu_1 = \mu_2$$

$$H_A: \mu_1 \neq \mu_2$$

Called a *two-tailed* test.

Step 1: Write down H_0 and H_A

Q: Is the average strength of materials from supplier A higher than supplier B?

H_0: The average strength is the same.

H_A: The average strength is higher.

In mathematical notation we can write:

$$H_0: \mu_1 = \mu_2$$

$$H_A: \mu_1 > \mu_2$$

Called a *one-tailed* test.

Step 1: Write down H_0 and H_A

Q: Is the average strength of materials from supplier A lower than supplier B?

H_0: The average strength is the same.

H_A: The average strength is lower.

In mathematical notation we can write:

$$H_0: \mu_1 = \mu_2$$

$$H_A: \mu_1 < \mu_2$$

Called a *one-tailed* test.

Step 2: Select a significance level α

This is decided on by the researcher before seeing the data.

Values typically used are 0.05, 0.01, 0.001.

Controls the chance of making an error when we reject H_0 (Type I error).

Step 3: Calculate the test statistic

This is the *evidence* from our data.

We use this value to decide whether our data are compatible with H_0 or not.

We use different formulae depending on the question of interest.

Step 4: Calculate the p-value

The p-value is computed by assuming H_0 is true, and the determining the probability of observing a result at least as extreme as the observed test statistic if H_0 is true.

The p-value is then compared with α to decide between H_0 and H_A.

Step 5: Decision and conclusion

If the p-value < α => reject H_0 and the data are more consistent with H_A.

If the p-value > α => not enough evidence to reject H_0.

NB! State conclusion in words and relate back to the original research question!

MACSI
Mathematics Applications Consortium for Science & Industry

Step 5: Decision and conclusion

If the p-value > α, it is tempting to say that you *accept* H_0.

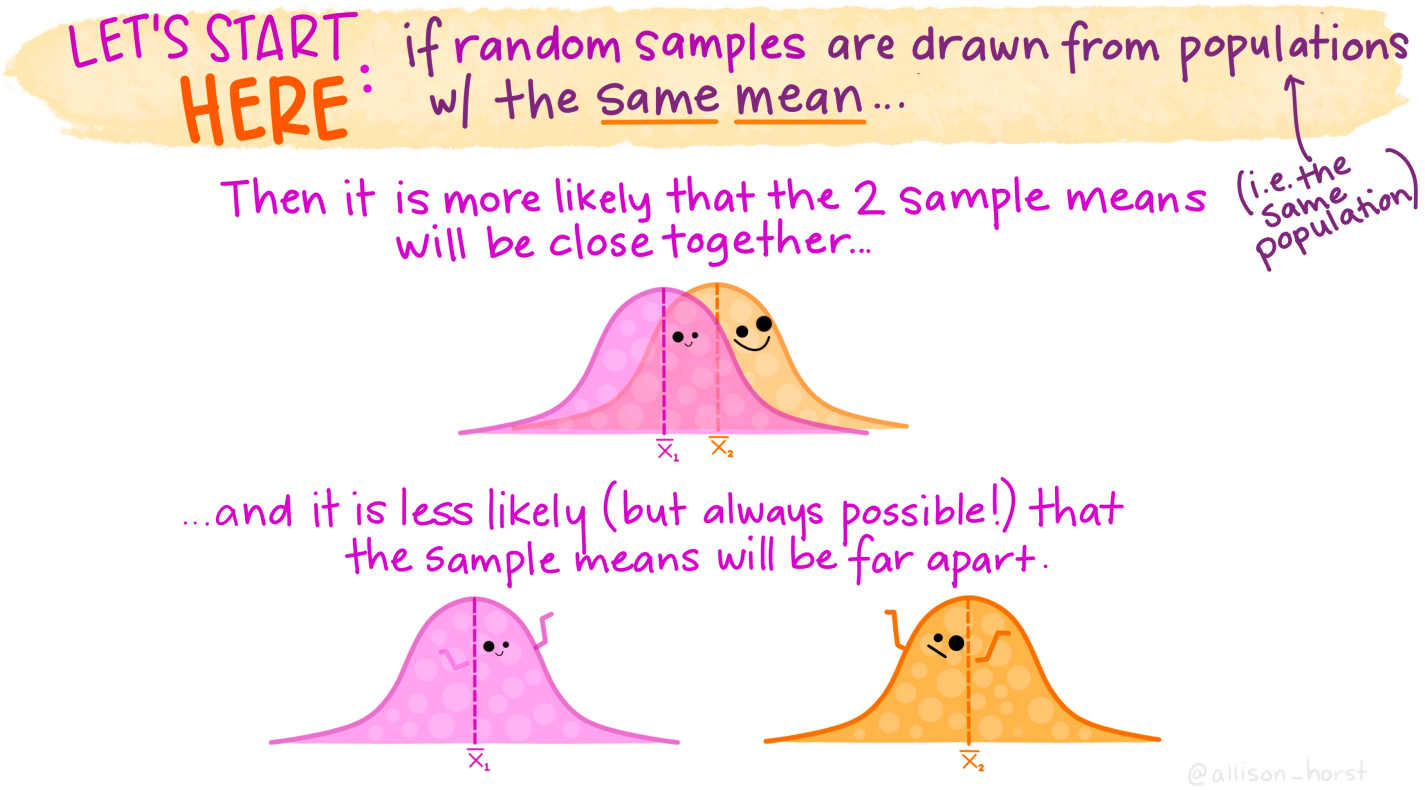
NO!!

Incorrectly makes it seem that the value specified in H_0 is the true value of the parameter but lots of different scenarios could generate the same H_0.

Can only say we do not have enough evidence to reject H_0.

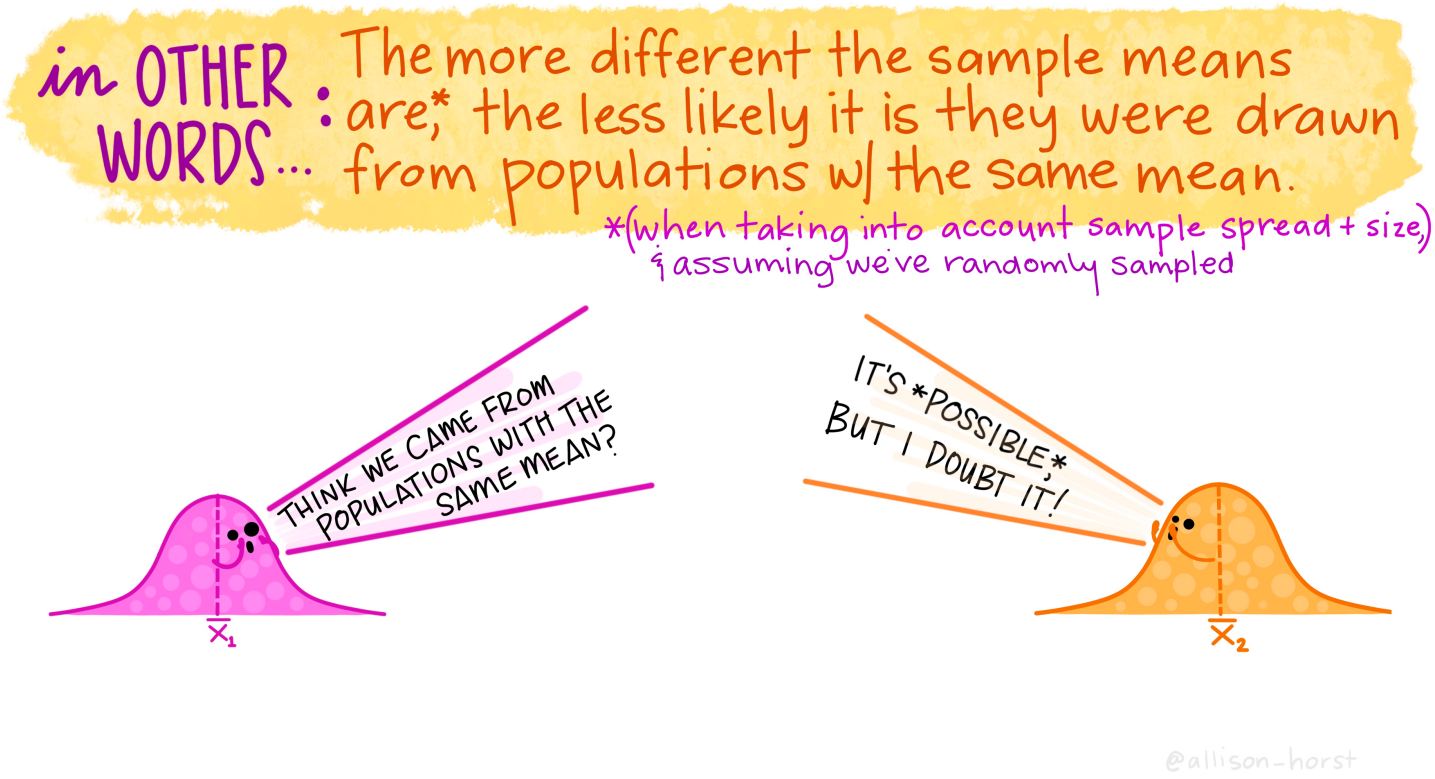
MACSI
Mathematics Applications Consortium for Science & Industry

Comparing two means



Credit: @alison_horst

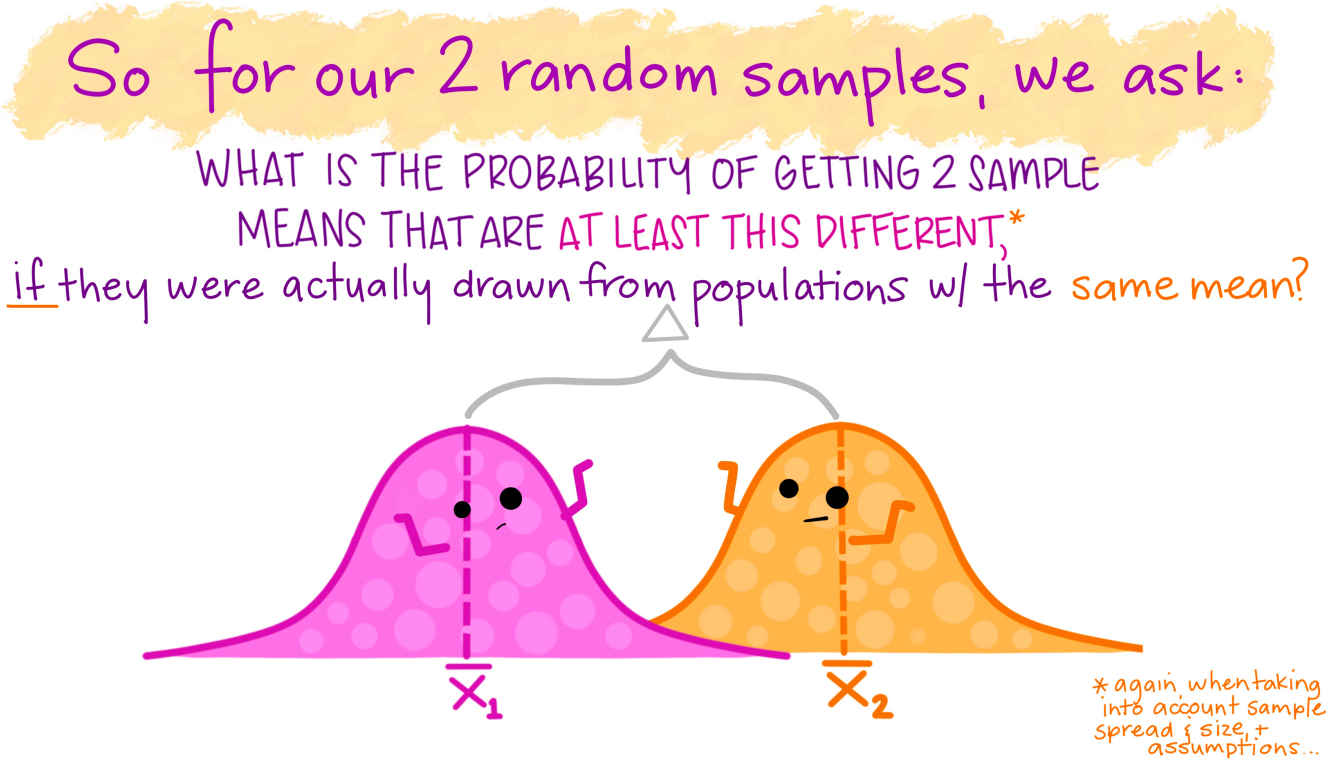
Comparing two means



Credit: @alison_horst

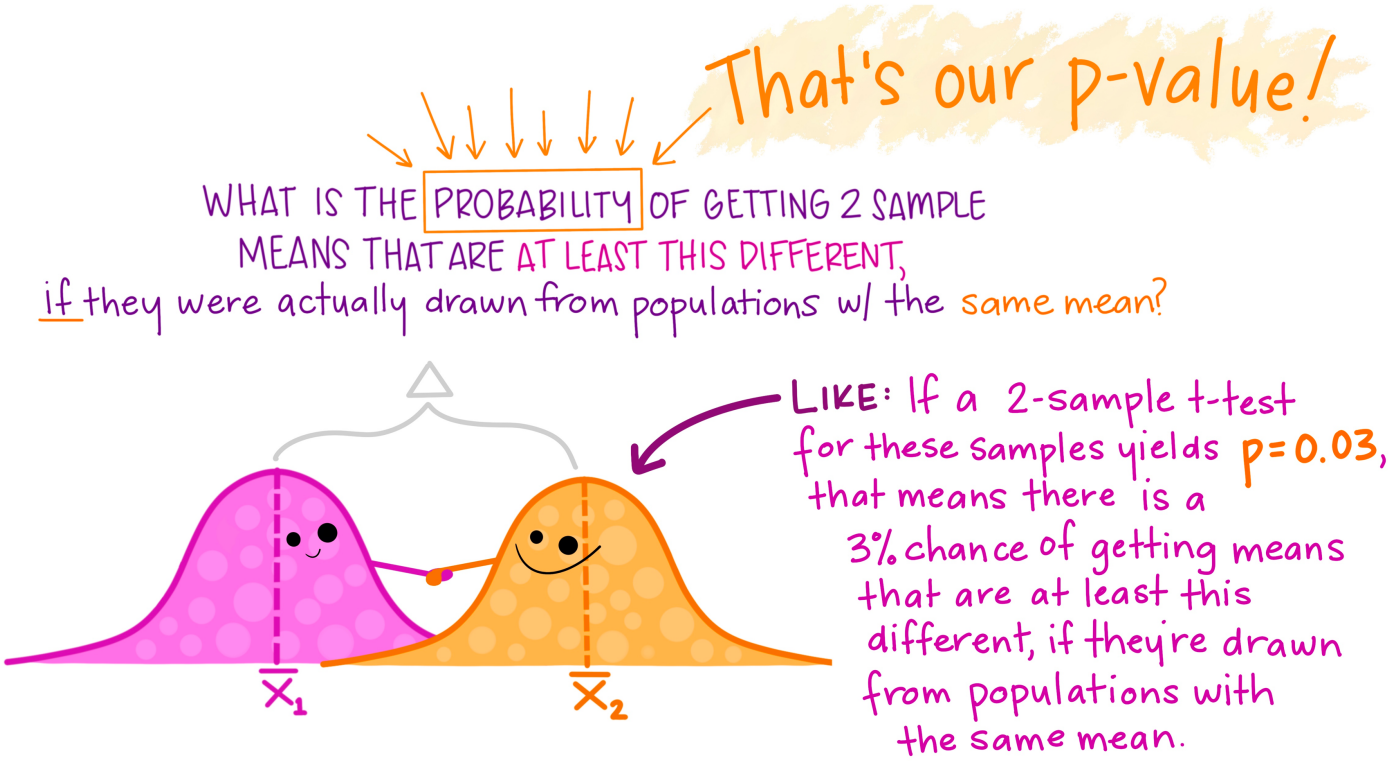
MACSI
Mathematics Applications Consortium for Science & Industry

Comparing two means



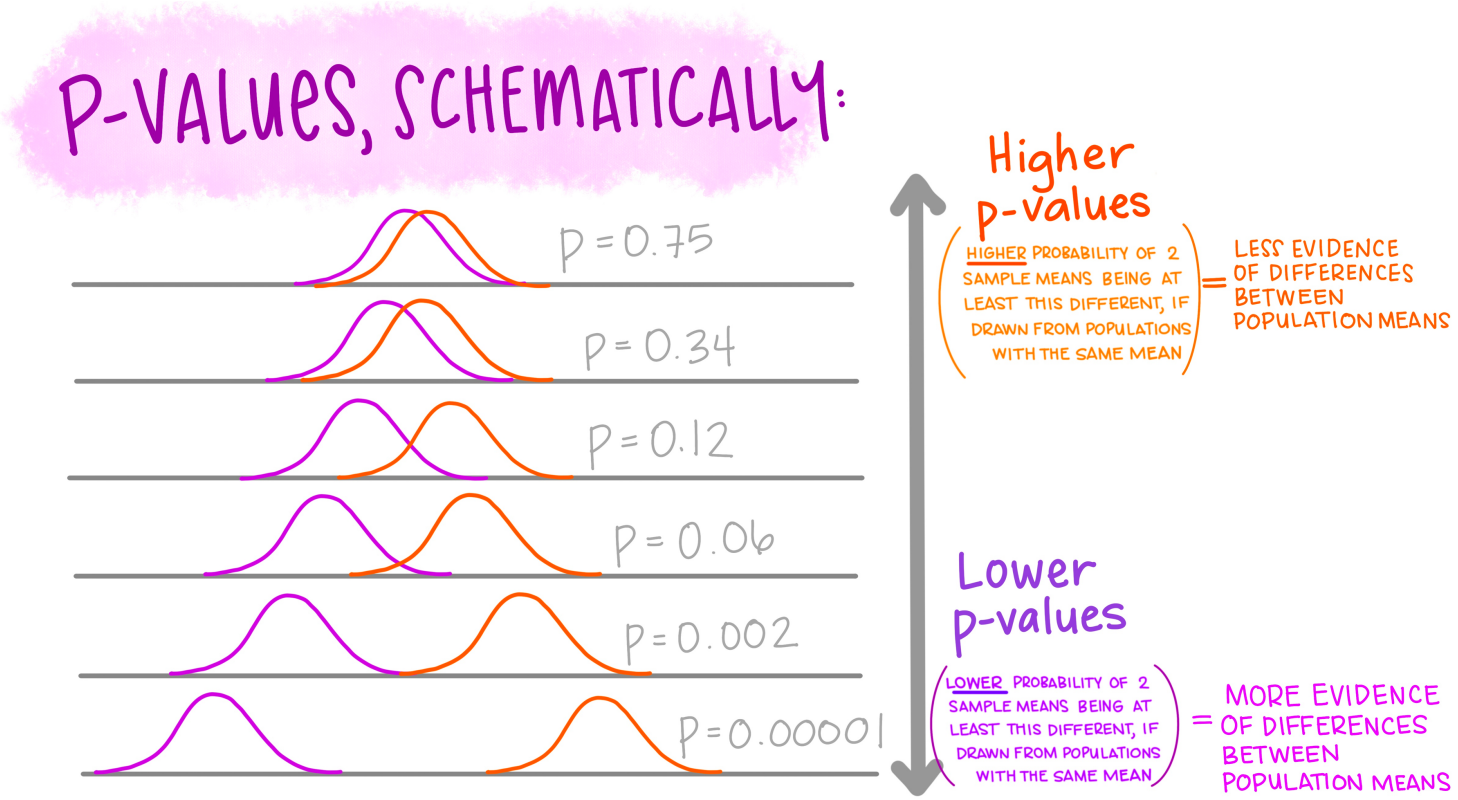
Credit: @alison_horst

Comparing two means



Credit: @alison_horst

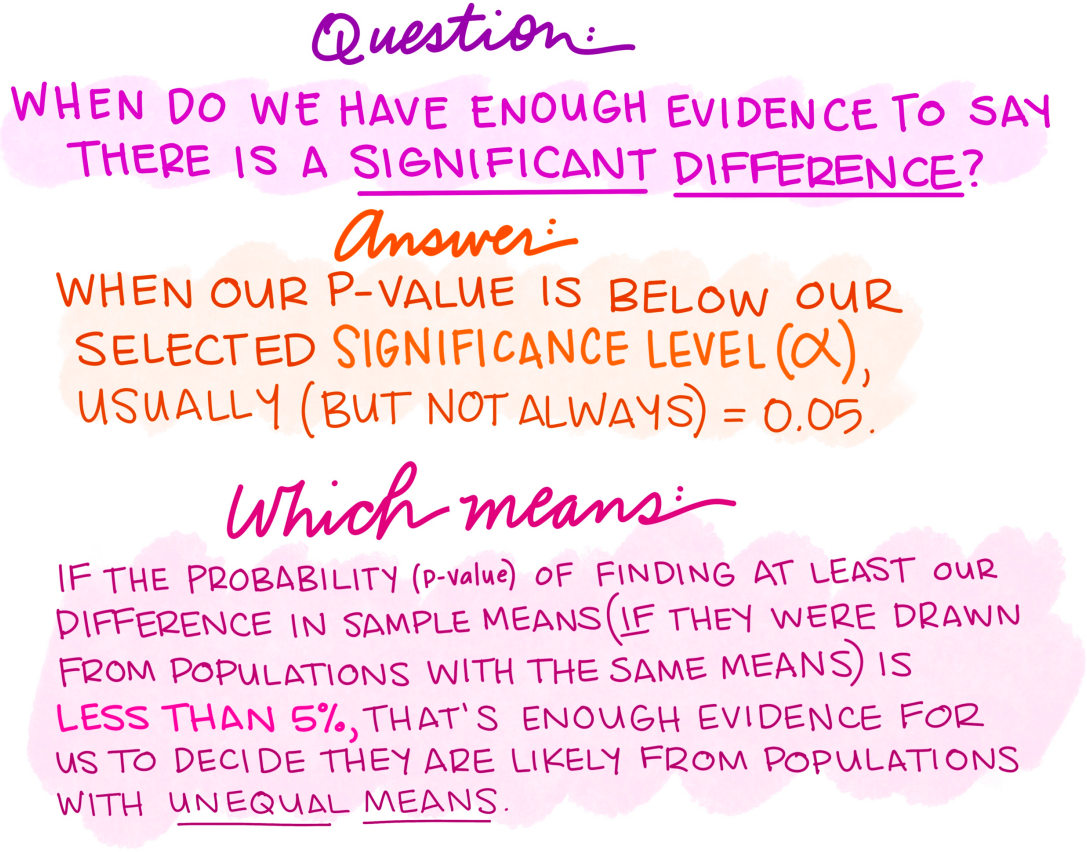
Comparing two means



Credit: @alison_horst

MACSI
Mathematics Applications Consortium for Science & Industry

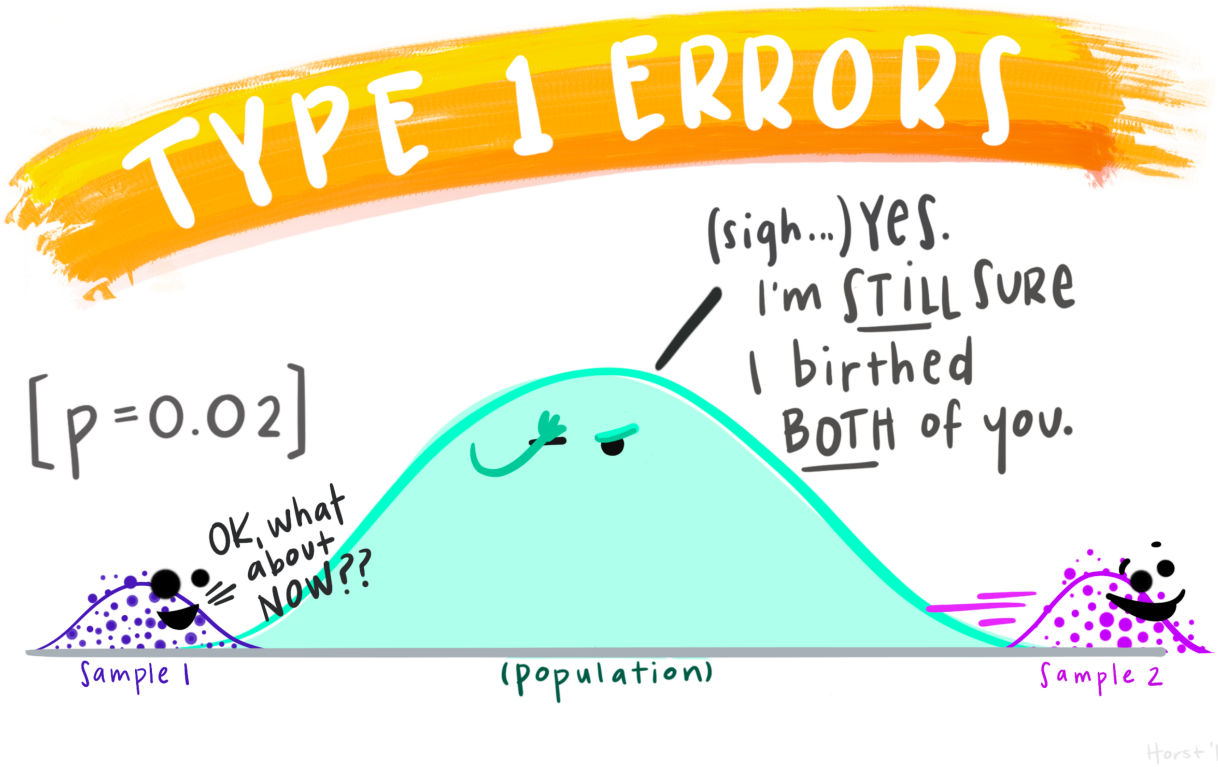
Comparing two means



Credit: @alison_horst

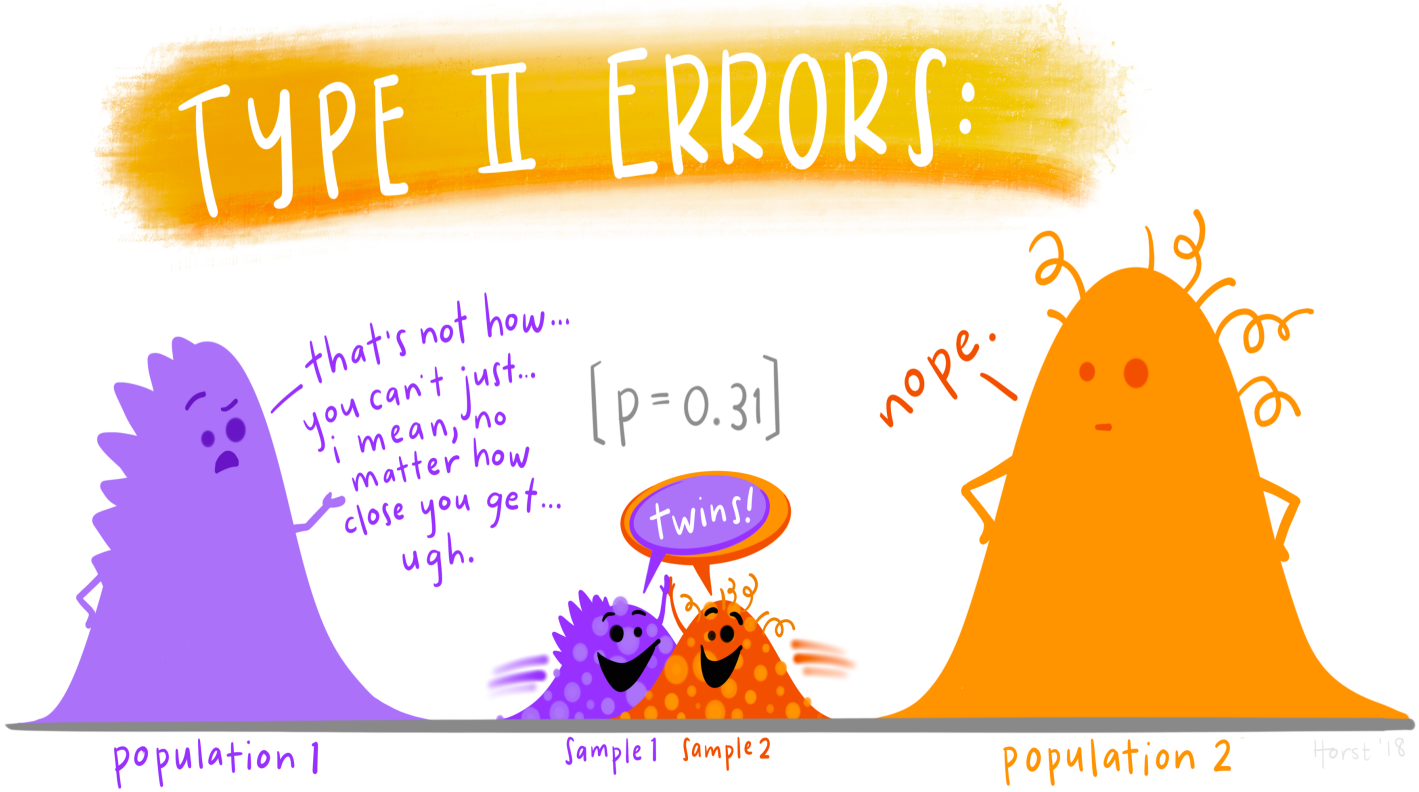
MACSI
Mathematics Applications Consortium for Science & Industry

Type I error



Credit: @alison_horst

Type I error



Credit: @alison_horst

Comparing two independent means

Step 1: Write down H_0 and H_A.

Step 2: Select the significance level α.

Step 3: Calculate the test statistic (assuming H_0 is true).

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Step 4: Calculate the p-value = Pr($T > |t|\,|\,H_0$ is true).

Step 5: Decision and conclusion.

MACSI
Mathematics Applications Consortium for Science & Industry

Assumptions required

One or both of the following must hold for EACH group:

1. Large sample size => ok in general. If there are extreme outliers or skewness, may need alternative approach.

2. Small sample size => data must be drawn from a Normal distribution in the population.

Must check these assumptions using QQ-plots, boxplots, etc.

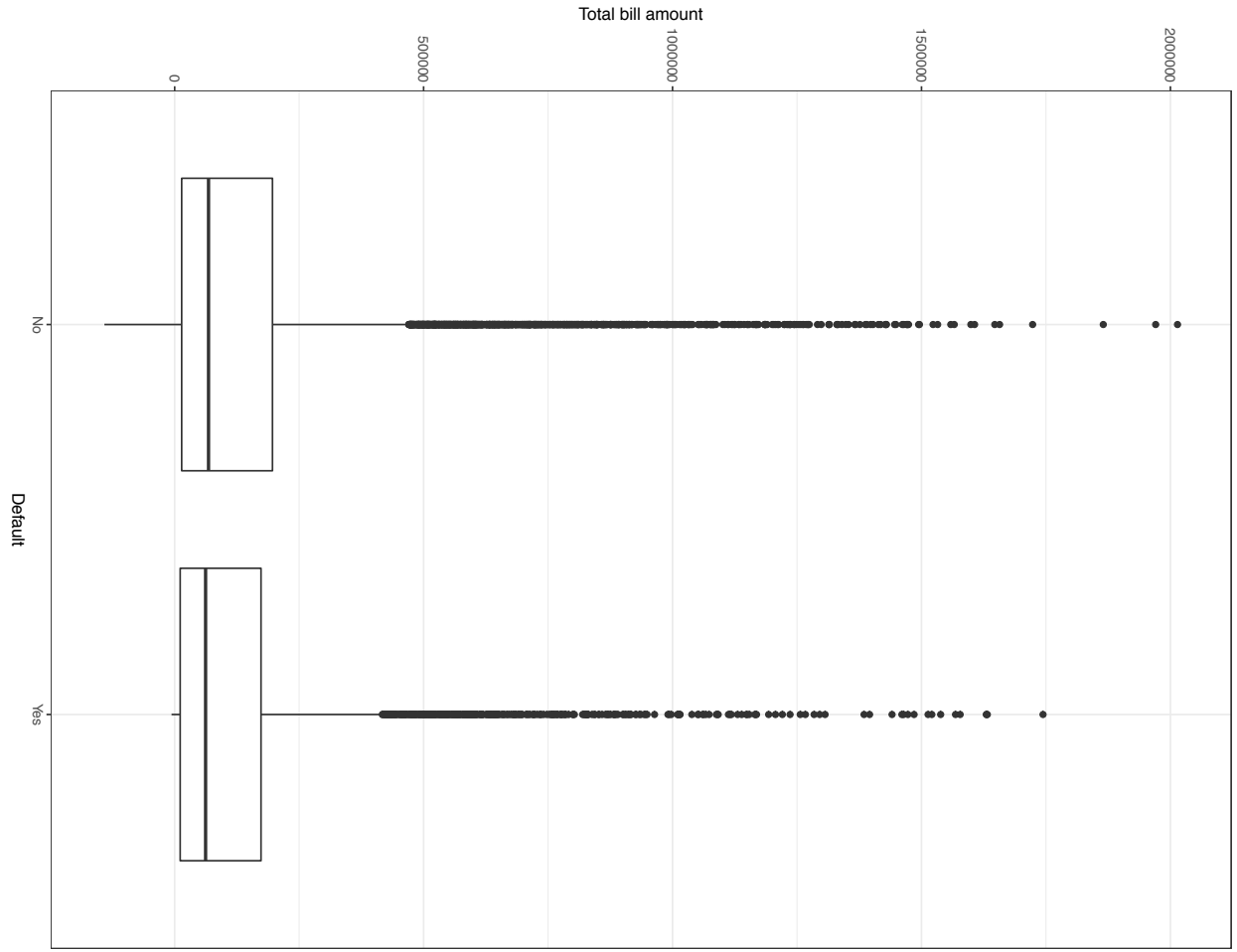
Example

For the Credit Risk dataset, we want to determine if on average the Total Bill Amount differs between customers who default and those who don't. Let group 1 denote those who do not default and group 2 denote those who do default.

$\bar{y}_1$ = 149264.9 s_1 = 208227.8 n_1 = 15306

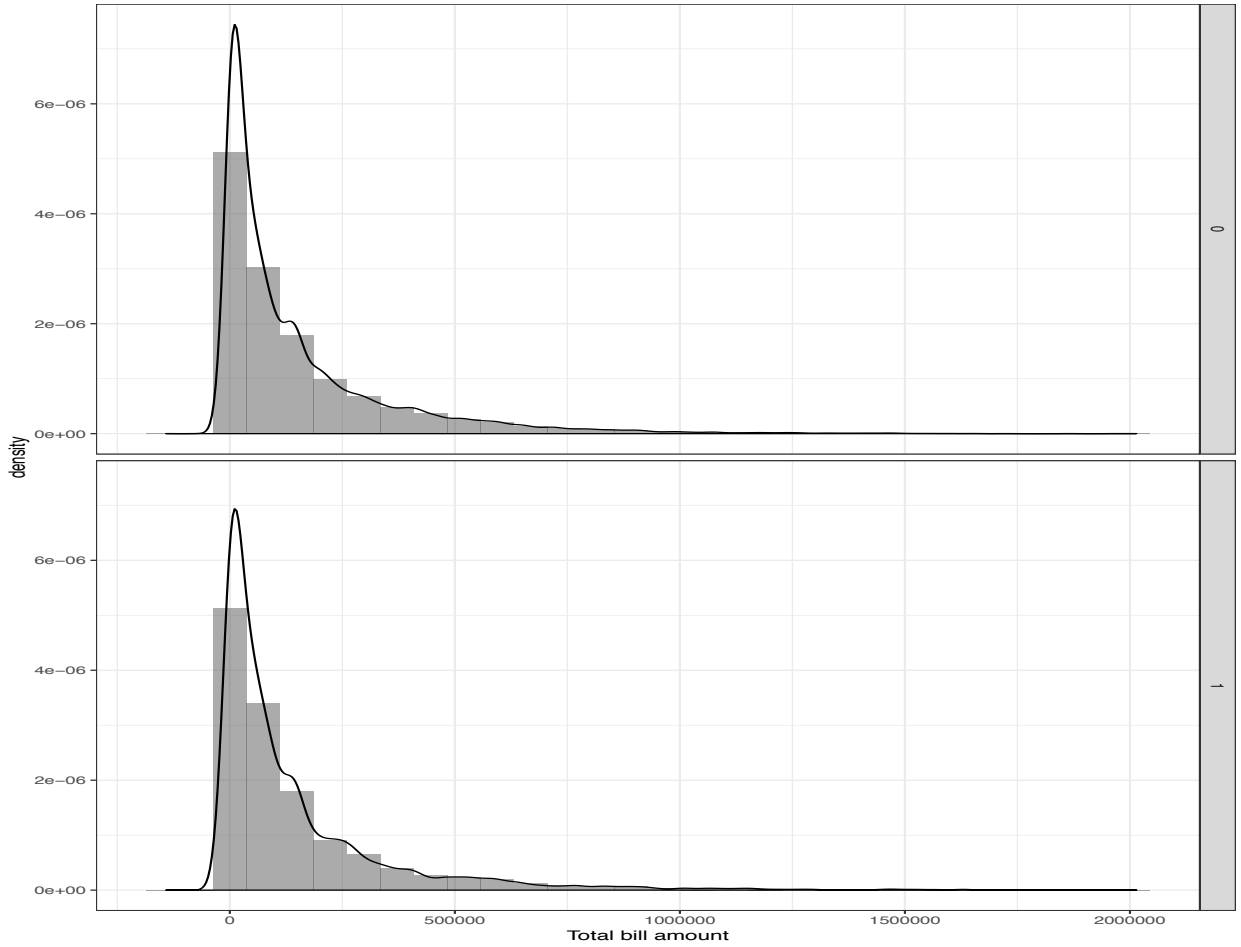
$\bar{y}_2$ = 141461.0 s_2 = 210486.8 n_2 = 4387

Example



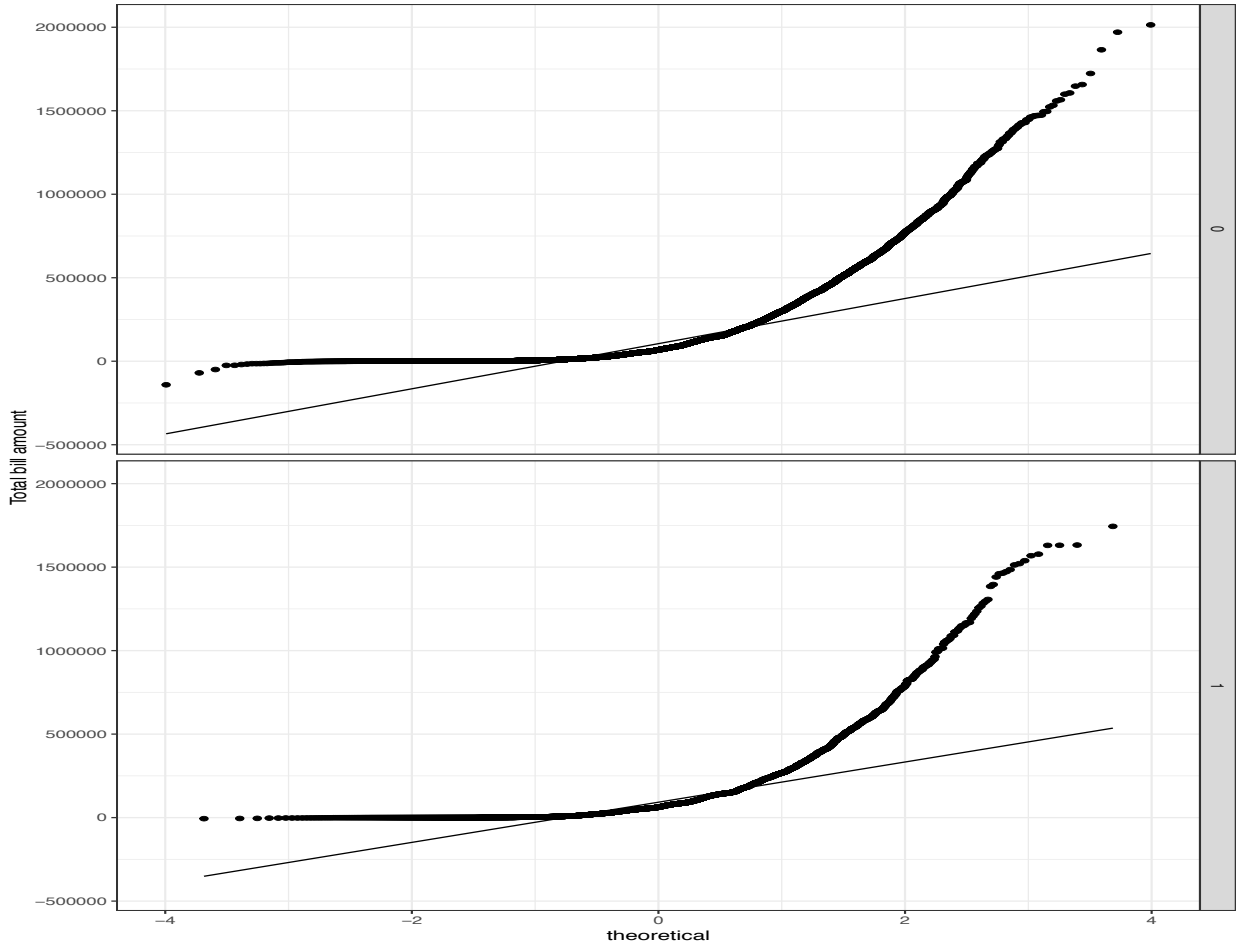
MACSI
Mathematics Applications Consortium for Science & Industry

Example



MACSI
Mathematics Applications Consortium for Science & Industry

Example



MACSI
Mathematics Applications Consortium for Science & Industry

Example

Step 1: $H_0: \mu_1 = \mu_2$

$H_A: \mu_1 \neq \mu_2$

Step 2: Set $\alpha = 0.05$.

Step 3: Calculate the test statistic (assuming H_0 is true).

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(149264.9 - 141461.0) - (0)}{\sqrt{\frac{208227.8^2}{15306} + \frac{210486.8^2}{4387}}} = 2.17$$

Step 4: Calculate the p-value = 0.03

Step 5: Since the p-value < α => enough evidence to reject H_0. There is evidence of a statistically significant difference between the average total bill amount in the two groups.

Example

```
> t.test(dat$total_bill_amt~dat$default)

        Welch Two Sample t-test

data:  dat$total_bill_amt by dat$default
t = 2.1701, df = 7033.1, p-value = 0.03003
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
   754.441 14853.293
sample estimates:
mean in group 0 mean in group 1
       149264.9        141461.0
```

Comparing paired data

Samples are paired when one member of each sample is linked in some way with a member of the second sample.

- The weights of individuals undertaking a fitness programme were recorded before and after they completed the programme.
- Two measurement processes were used to measure the weight of tablets on a production line.

The main effect of pairing is that the members of a pair are more alike than non-paired members.

Comparing two paired means

Calculate the column of differences $d_i = y_{i1} - y_{i2}$.

Step 1: Write down H_0 and H_A.

Step 2: Select the significance level α.

Step 3: Calculate the test statistic (assuming H_0 is true).

$$t = \frac{\bar{y}_d - \mu_d}{\frac{s_d}{\sqrt{n_d}}}$$

Step 4: Calculate the p-value = Pr($T > |t|$ | H_0 is true).

Step 5: Decision and conclusion.

Assumptions required

One or both of the following must hold for the *differences:*

1. Large sample size => ok in general. If there are extreme outliers or skewness, may need alternative approach.

2. Small sample size => *differences* must be drawn from a Normal distribution in the population.

Must check these assumptions using QQ-plots, boxplots, etc.

Example

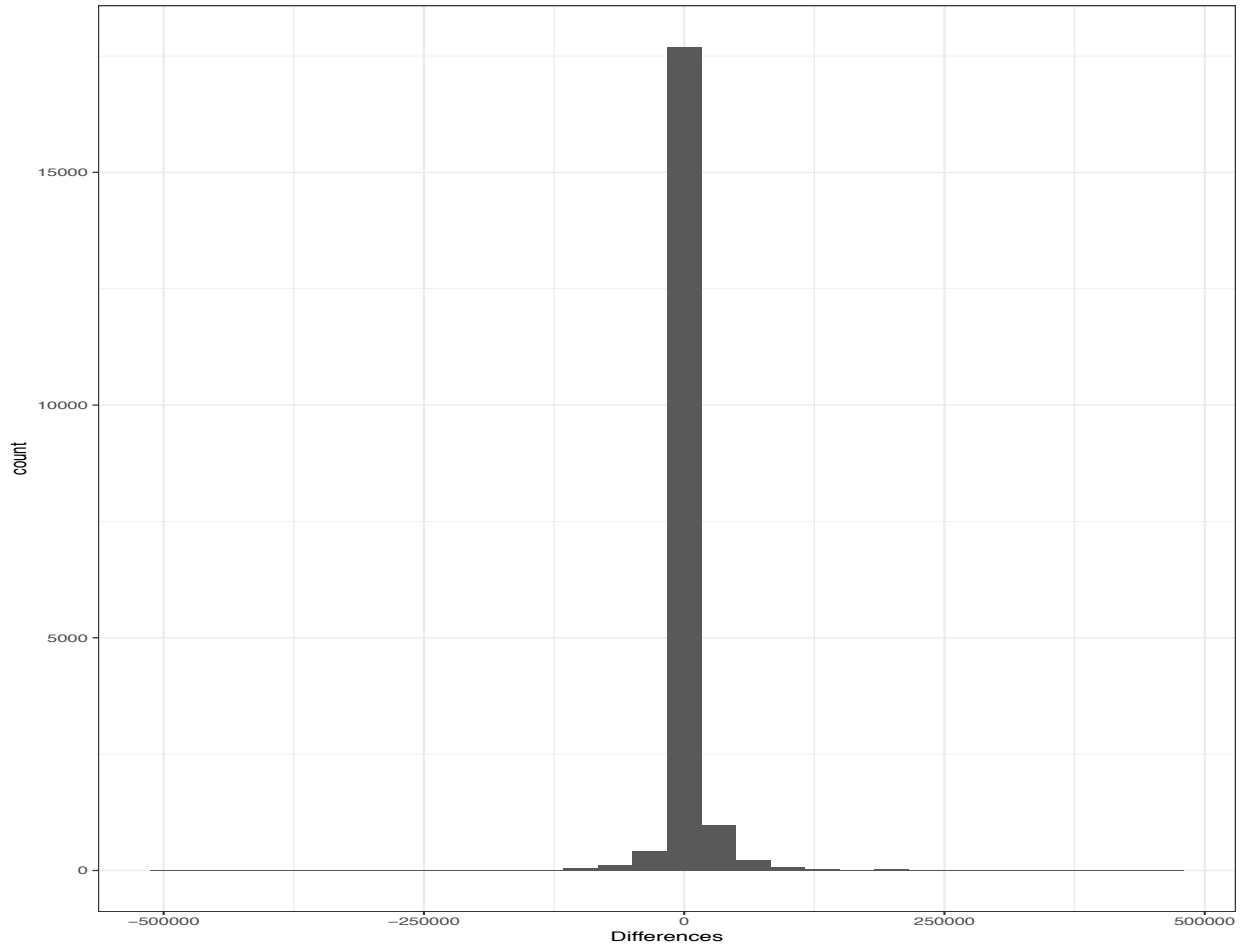
For the Credit Risk dataset, we want to determine if the average bill amount for bill 1 and 2 are different. This is paired data as each individual is measured twice. First calculate the column of differences:

Bill Amt 1	Bill Amt 2	Diffs (d_i)
2682	1725	957
8617	5670	2947
64400	57069	7331
367965	412023	-44058
11285	14096	-2811
0	0	0
…	…	…

$\bar{y}_d$ = 1913.3
s_d = 22974.19
n_d = 19693

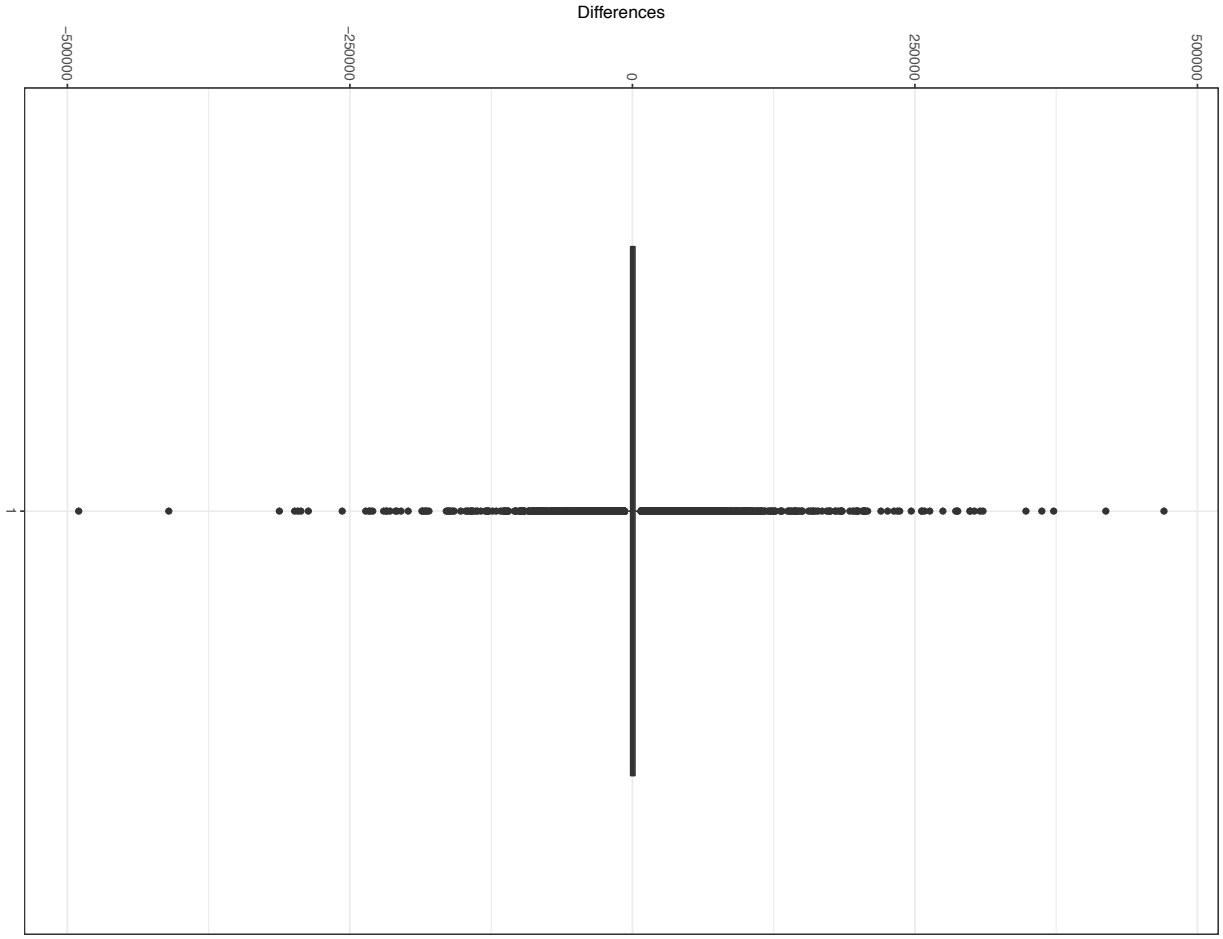
MACSI
Mathematics Applications Consortium for Science & Industry

Example

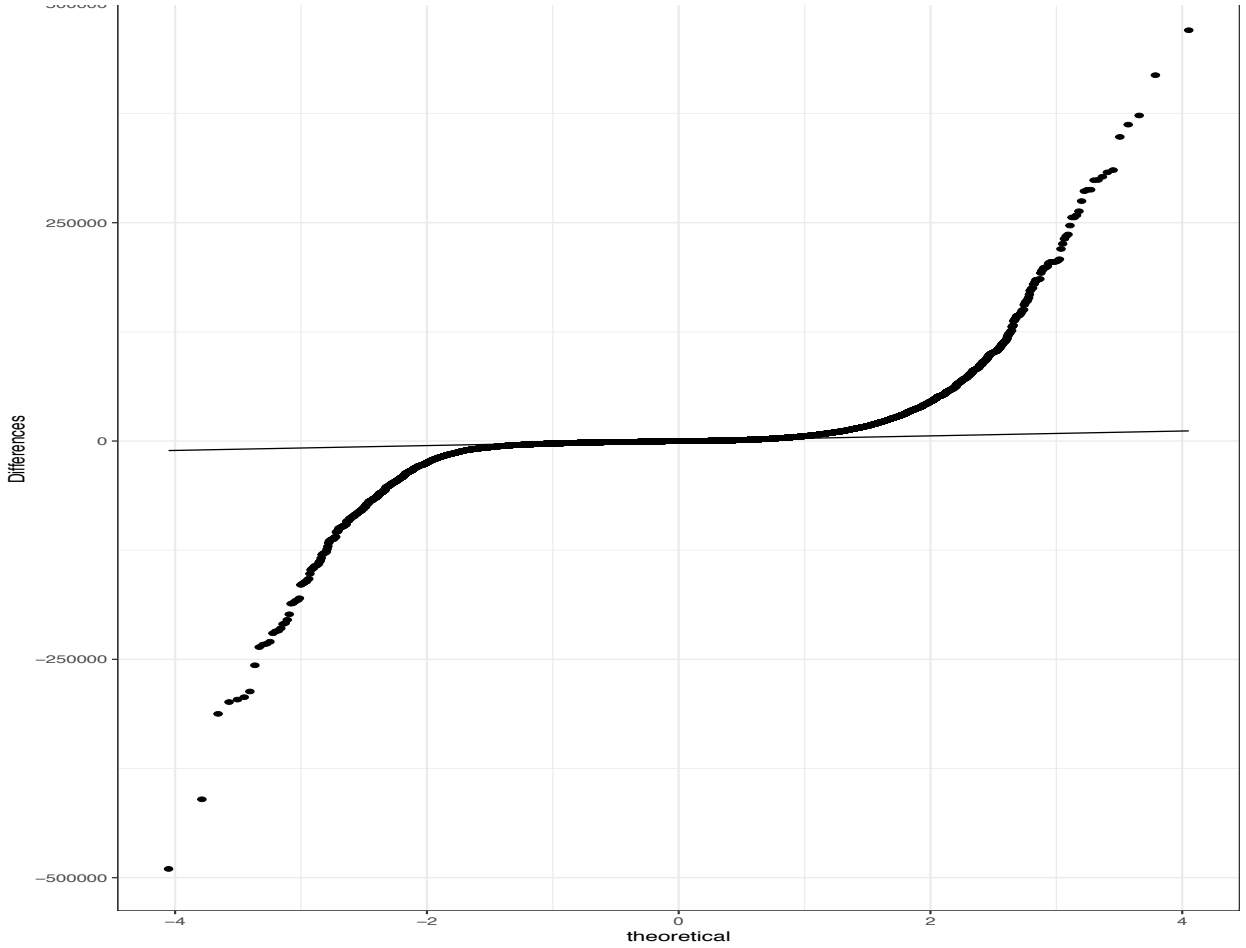


MACSI
Mathematics Applications Consortium for Science & Industry

Example



Example



MACSI
Mathematics Applications Consortium for Science & Industry

Example

Step 1: $H_0: \mu_d = 0$

$H_A: \mu_d \neq 0$

Step 2: Set $\alpha = 0.05$.

Step 3: Calculate the test statistic (assuming H_0 is true).

$$t = \frac{\bar{y}_d - \mu_d}{\frac{s_d}{\sqrt{n_d}}} = \frac{1913.3 - 0}{\frac{22974.19}{\sqrt{19693}}} = 11.69$$

Step 4: Calculate the p-value < 0.001

Step 5: Since the p-value < α => enough evidence to reject H_0. There is evidence of a statistically significant difference between the average two bill amounts.

Example

```
> t.test(dat$bill_amt1, dat$bill_amt2, paired=TRUE)

        Paired t-test

data:  dat$bill_amt1 and dat$bill_amt2
t = 11.687, df = 19692, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1592.409 2234.193
sample estimates:
mean of the differences
                1913.301
```

Issues with p-values

Significance level is arbitrary. What does p-value = 0.049 or 0.051 mean?

Creates an artificial dichotomy where a significant result is judged as real and important, whereas non-significant means there is no effect.

The p-value depends on *n* and the size of the difference. If *n* is large enough => any difference (no matter how small) will be statistically significant.

Issues with p-values

Calculating a p-value assuming H_0 is true makes little sense as H_0 is almost always false.

A significance test gives no information about the size of the difference, the uncertainty around it or the strength of evidence against H_0.

If H_0 is not rejected, it can be for many reasons other than H_0 is true. Absence of evidence ≠ evidence of absence.

Useful resources

The American Statistical Association's statement on p-values (http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108).

The Nature comment on retiring p-values (https://www.nature.com/articles/d41586-019-00857-9).

Frank Harrell's website. NB the Datamethods discussion board (https://www.fharrell.com/).

MACSI
Mathematics Applications Consortium for Science & Industry

Solution?

Use/report confidence intervals.

Calculate effect sizes.

Use statistical model selection criteria.

Effect sizes

Effect size statistics provide an indication of the magnitude of the differences between groups, not just whether the differences could have occurred by chance.

Effect sizes for independent groups *t*-test

1. **Cohen's d:**

$$d = \frac{M_1 - M_2}{SD_{pooled}}$$

(0.2 = small, 0.5 = medium, 0.8 = large)

2. **Eta²:**

$$Eta^2 = \frac{t^2}{t^2 + (n_1 + n_2 - 2)}$$

(0.01 = small, 0.06 = medium, 0.14 = large)

MACSI
Mathematics Applications Consortium for Science & Industry

Example

1. **Cohen's d:**

$$d = \frac{M_1 - M_2}{SD_{pooled}} = \frac{149264.9 - 141461}{208733} = 0.037$$

2. **Eta^2:**

$$Eta^2 = \frac{t^2}{t^2 + (n_1 + n_2 - 2)} = \frac{2.17^2}{2.17^2 + (15306 + 4387 - 2)}$$

$$= 0.0002$$

MACSI
Mathematics Applications Consortium for Science & Industry

Effect sizes for paired *t*-test

1. **Cohen's d:**

$$d = \frac{Mean\ difference}{SD_{diff}}$$

(0.2 = small, 0.5 = medium, 0.8 = large)

2. **Eta^2:**

$$Eta^2 = \frac{t^2}{t^2 + (n_d - 1)}$$

(0.01 = small, 0.06 = medium, 0.14 = large)

MACSI
Mathematics Applications Consortium for Science & Industry

Example

1. **Cohen's d:**

$$d = \frac{Mean\ difference}{SD_{diff}} = \frac{1913.3}{22974.19} = 0.08$$

2. **Eta^2:**

$$Eta^2 = \frac{t^2}{t^2+(n_d-1)} = \frac{11.69^2}{11.69^2+(19693-1)} = 0.007$$

Comparing two independent proportions

Step 1: Write down H_0 and H_A.

Step 2: Select the significance level α.

Step 3: Calculate the test statistic (assuming H_0 is true).

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_c(1-p_c)}{n_1} + \frac{p_c(1-p_c)}{n_2}}}, \quad p_c = \frac{Y_1 + Y_2}{n_1 + n_2}$$

Step 4: Calculate the p-value = Pr($Z > |z|$ | H_0 is true).

Step 5: Decision and conclusion.

Example

For the Credit Risk dataset, we want to determine if the proportion of male customers who default is different to the proportion of female customers who default. Let group 1 denote male customers who do not default and group 2 denote female customers who do not default.

$Y_1 = 1919$ $n_1 = 7817$ $\hat{p}_1 = 1919/7817 = 0.245$

$Y_2 = 2468$ $n_2 = 11876$ $\hat{p}_2 = 2468/11876 = 0.208$

Example

Step 1: $H_0: p_1 = p_2$

$H_A: p_1 \neq p_2$

Step 2: Set $\alpha = 0.05$.

Step 3: Calculate the test statistic (assuming H_0 is true).

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_c(1-p_c)}{n_1} + \frac{p_c(1-p_c)}{n_2}}}$$

$$= \frac{(0.245 - 0.208) - (0)}{\sqrt{\frac{0.223(1-0.223)}{7817} + \frac{0.223(1-0.223)}{11876}}} = 6.10$$

Step 4: Calculate the p-value < 0.001

Step 5: Since the p-value < α => enough evidence to reject H_0. There is evidence of a statistically significant difference between the proportion of males and females who default.

Example

```
> prop.test(x=c(1919,2468), n=c(7817,11876), correct=FALSE)

        2-sample test for equality of proportions without continuity correction

data:  c(1919, 2468) out of c(7817, 11876)
X-squared = 38.649, df = 1, p-value = 5.074e-10
alternative hypothesis: two.sided
95 percent confidence interval:
 0.02566507 0.04968797
sample estimates:
   prop 1    prop 2
0.2454906 0.2078141
```

Correlation and regression

Introduction

Often want to describe the relationship between two *quantitative* variables, *X* and *Y*.

For example:

- Does the amount spent per month by a company on training its sales team affects it monthly sales?
- Does the number of hours spent studying for an exam influence the exam score?
- Can we predict the asking price of a car based on its age?

Introduction

Linear regression and *correlation* are two commonly used methods.

Typically denote one variable *Y* and the other *X*.

Y is called the dependent (response) variable, while *X* is called the independent (explanatory) variable.

Have *n* pairs of values (X_i , Y_i) for $i = 1, \ldots, n$.

Scatterplots

The first step in any correlation analysis is to create a *scatterplot*.

Use the *X* axis to represent the *X* variable and the *Y* axis to represent the *Y* variable.

Each (X_i , Y_i) pair is plotted as a point.

Scatterplots

Helps us to determine if a relationship exists between and *Y* and if so the nature of this relationship (linear, curvi-linear, etc.).

Is a means of identifying *outliers* which can have a major impact on our analysis.

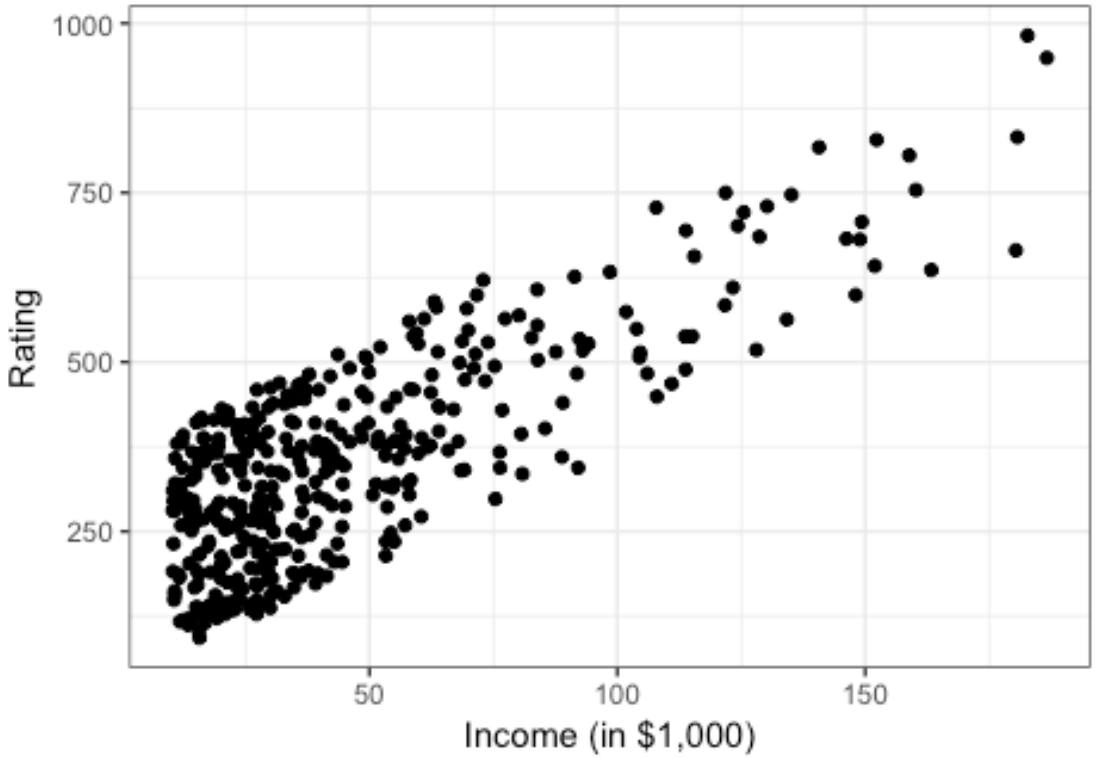
Outliers should always be identified and investigated further.

Example

Using the credit data, we want to predict a person's credit rating using their income.

Person	Income (*X*)	Credit Rating (*Y*)
1	14.81	283
2	106.03	483
3	104.59	514
4	148.92	681
5	55.88	357
…	…	…

Example



Correlation

Two quantitative variables, X and Y are said to be correlated if they display a strong *linear* relationship.

Once we have established that a linear relationship exists, we can estimate the strength of that relationship.

Measured using the (Pearson) correlation coefficient.

Correlation

Sample correlation coefficient = r, and is a measure of the strength of the linear association between X and Y.

Possible values are $-1 \leq r \leq 1$.

Properties of r

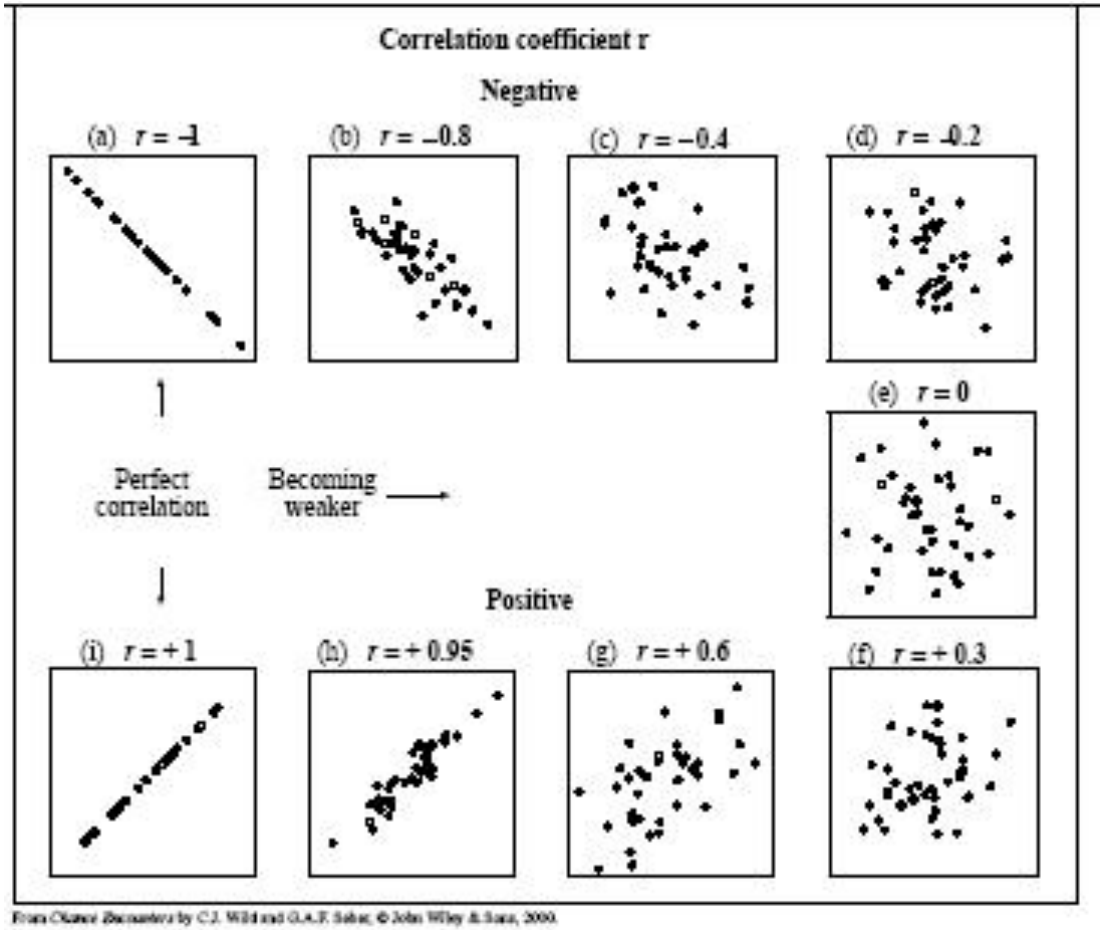
1. Look at the magnitude:

$|r| = 1$, a perfect linear relationship
$|r| \approx 1$, a strong linear relationship
$|r| \approx 0$, a weak linear relationship
$|r| = 0$, no relationship/no linear relationship

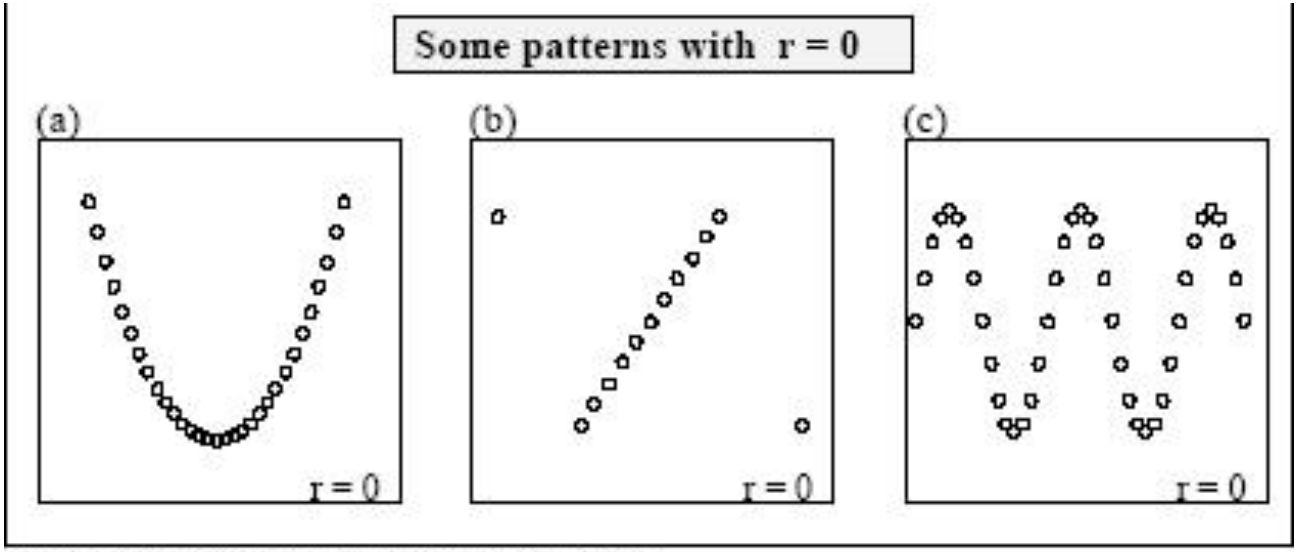
2. Look at the sign:

$r > 0$, a positive linear relationship
$r < 0$, a negative linear relationship

Properties of *r*



Properties of *r*



From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.

Calculating *r*

Calculated as:

$$r = \frac{SS_{XY}}{\sqrt{SS_{XX} \times SS_{YY}}}$$

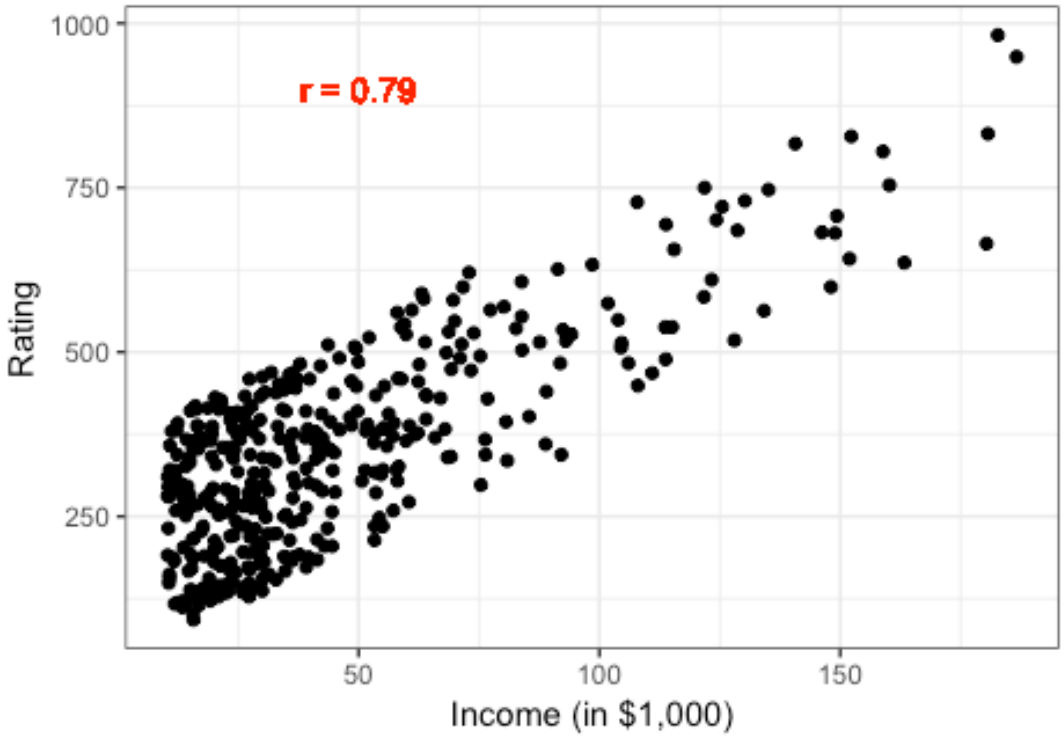
where

$$SS_{XY} = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}$$

$$SS_{XX} = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}$$

$$SS_{YY} = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}$$

Example



MACSI
Mathematics Applications Consortium for Science & Industry

Hypothesis test

Can do a hypothesis test for the correlation coefficient:

$H_0 : \rho = 0$ (No linear relationship)

$H_A : \rho \neq 0$ (There is a linear relationship)

Most software will output p-values for the above test.

This test is sensitive to sample size.

Very large samples will result in significant results even if the correlation is very weak.

Effect size

Need to consider the magnitude of the correlation coefficient (Cohen's recommendations for social sciences).

Effect	*r*
Small	0.10
Medium	0.30
Large	0.50

WARNING!

CORRELATION DOES NOT IMPLY CAUSATION.

A relationship between *X* and *Y*, does not guarantee a *causal* relationship.

Strong correlation => tempting to assume that an increase or decrease in one variable causes a change in the other variable.

Called *spurious relationships*.

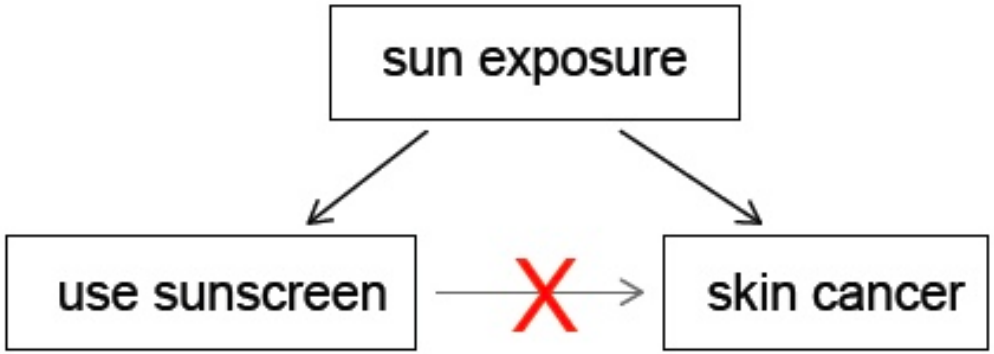
Lurking variables

The relationship between *X* and *Y* might be explained by another variable *Z*.

Called a *lurking variable*.

Directly affects both *X* and *Y* and makes it appear that *X* and *Y* are directly related to each other.

Lurking variables



Confounding variables

Observed relationship could be occurring because both *X* and *Z* are influencing *Y*.

Influence of *X* on *Y* and influence of *Z* on *Y* cannot be separated.

Say that *X* and *Z* are *confounded*.

If two variables have a strong correlation, can only conclude that there is a relationship between those two variables, not that a change in one causes a change in the other.

MACSI
Mathematics Applications Consortium for Science & Industry

Linear regression

Regression model allows us to describe the trend in the data and make predictions, i.e. allows us to predict what value of *Y* we would expect to see for a given value of *X.*

Gives us an equation that uses one variable to help explain the variation in the values of another variable.

Simple linear regression (SLR) has one response variable *Y* and one explanatory variable *X* and assumes the relationship between the two variables is linear.

Linear regression model

If the true relationship in the population is:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Need to make the following assumptions:

- A linear relationship between *X* and *Y.*
- Normality of errors/residuals, .
- Residuals have constant variance.
- Residuals are independent.

Linear regression model

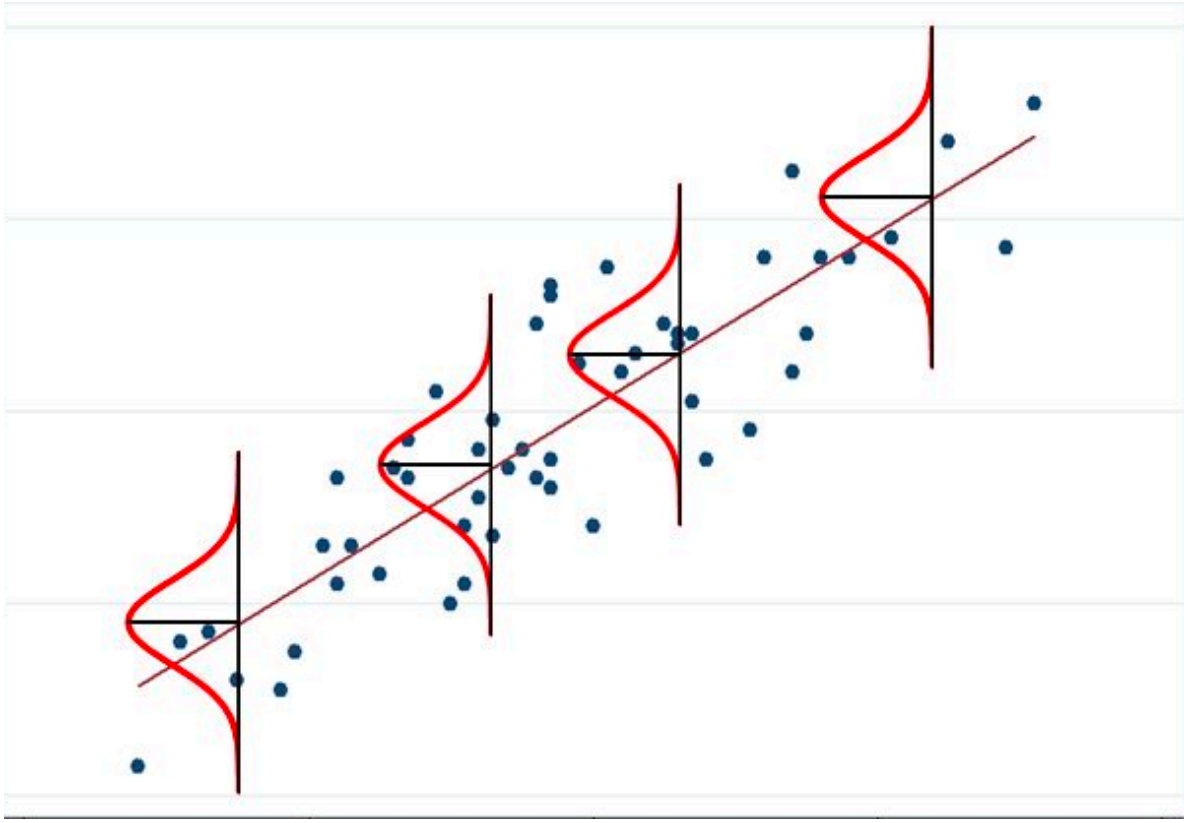
These assumptions can be summarised as follows:

$$\varepsilon_i \sim N(0, \sigma^2) \text{ and are i.i.d.}$$

Equivalently:

$$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2) \text{ and are i.i.d.}$$

Linear regression model



Linear regression model

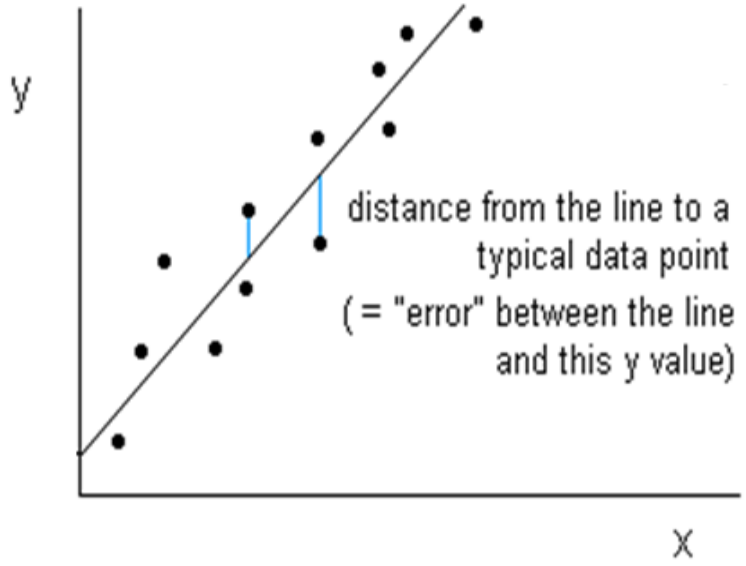
The estimated regression line is:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Find the estimated regression coefficients from the data using the *least squares* principle.

Can use the estimated regression line to make predictions.

Least squares



Can't fit a straight line through every point.

Relationship exists but not perfect.

Line predicts $\hat{y}_i$ but we actually observe y_i.

There is some error associated with the fitted line = $\hat{\varepsilon}_i = y_i - \hat{y}_i$

Least squares

The closeness of points to line is measured by the residual sum of squares:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

To find the best estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ from the data, minimise the *SSE*.

Called the *least squares* principle.

Estimating the regression coefficients

The (least squares) regression coefficients are given by:

$$\hat{\beta}_1 = \frac{SS_{XY}}{SS_{XX}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Residuals are estimated as:

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

Interpreting the regression coefficients

The coefficients have specific interpretations.

The intercept ($\hat{\beta}_0$) is the average value of *Y* when *X* = 0.

The slope ($\hat{\beta}_1$) is the average change in *Y* for every one unit increase in *X.*

Making predictions

Can use the estimated regression equation to make predictions.

Fill in specified values for *X* and calculate the predicted *Y* value.

NB! These predicted values are only accurate for *X* values within the range of *X* values used to estimate the regression coefficients!

Example

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 197.8411     7.7089   25.66   <2e-16 ***
Income        3.4742     0.1345   25.83   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 94.71 on 398 degrees of freedom
Multiple R-squared:  0.6263,    Adjusted R-squared:  0.6253
F-statistic:   667 on 1 and 398 DF,  p-value: < 2.2e-16
```

The regression equation is:

$$\hat{y}_i = 197.84 + 3.47x_i$$

Example

Interpreting the coefficients:

$\hat{\beta}_0$: When Income (X) = \$0 the average Credit Rating is 197.84.

$\hat{\beta}_1$: For every unit increase (additional \$1,000) in Income the average increase in Credit Rating is 3.47 units.

Example

Making predictions:

The predicted Credit Rating for an individual with an Income of $62,000 is:

$$197.84+3.47*(62) = 412.98$$

Residual analysis

Remember SLR assumes $\varepsilon_i \sim N(0, \sigma^2)$ and are i.i.d.

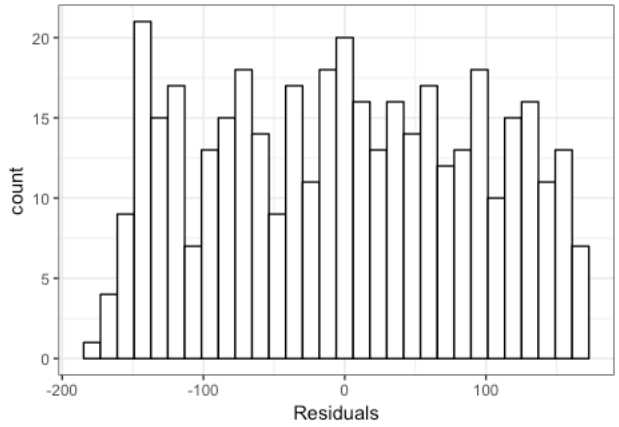
Check these assumptions using plots of the residuals ε_i .

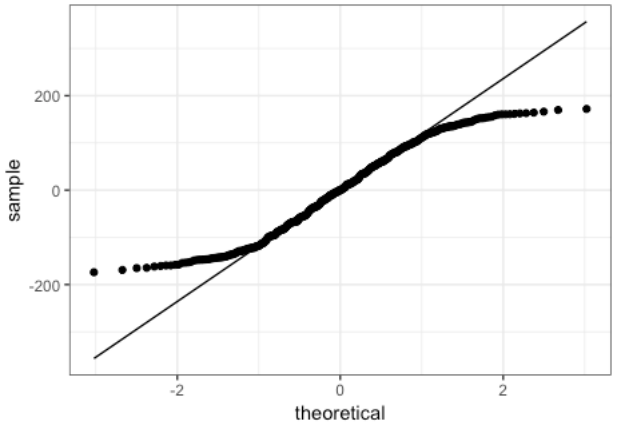
1. A Normal probability plot/histogram checks that the ε_i values are Normally distributed.

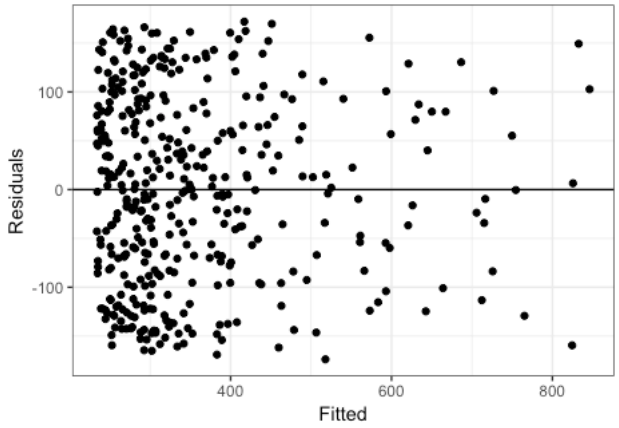
2. A plot of the ε_i values versus the $\hat{y}_i$ values checks the constant variance assumption.

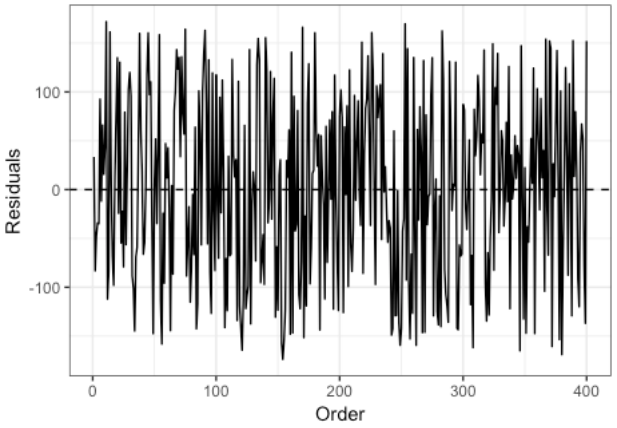
3. A plot of the ε_i values vs the order the observations checks the independence assumption.

Residual analysis









MACSI
Mathematics Applications Consortium for Science & Industry

Inference for the regression coefficients

Data are a sample.

$\hat{\beta}_1$ is a sample statistic used to estimate β_1.

Thus $\hat{\beta}_1$ has a sampling distribution with mean β_1 and standard error:

$$SE(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{SS_{xx}}} \quad \text{where} \quad \hat{\sigma}^2 = \frac{\text{SSE}}{n-2}$$

Inference for the regression coefficients

A confidence interval for β_1 is given by:

$$\hat{\beta}_1 \pm t_{\alpha/2, df=n-2} \times SE(\hat{\beta}_1)$$

A hypothesis test for β_1 is given by:

$H_0 : \beta_1 = 0$ (No linear relationship between X and Y)

$H_A : \beta_1 \neq 0$ (There is a linear relationship between X and Y)

Example

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 197.8411     7.7089   25.66   <2e-16 ***
Income        3.4742     0.1345   25.83   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 94.71 on 398 degrees of freedom
Multiple R-squared:  0.6263,    Adjusted R-squared:  0.6253
F-statistic:   667 on 1 and 398 DF,  p-value: < 2.2e-16
```

The p-value < 0.001 => reject H_0. There is evidence of a statistically significant relationship between Income and Credit Rating.

Example

A 95% confidence interval for β_1 is given by:

[3.21, 3.74]

For each $1,000 increase in Income, there will be an average increase in Credit Rating of between 3.21 and 3.74 units.

Coefficient of Determination

The *coefficient of determination,* R^2, is the proportion of the total variation in Y that is explained by X.

Is a measure of how well our model fits and gives a measure of the regression equation's ability to make predictions.

The closer R^2 is to 100%, the better the model describes the data.

$$R^2 = \frac{SS_{XY}^2}{SS_{XX} \times SS_{YY}} \times 100$$

Example

From the R output:

$$R^2 = 62.63\%.$$

This is a reasonably good model.

Approx. 62% of the variability in Credit Rating can be explained by Income.

Outliers and influential points

Unusual observations can be classified as:

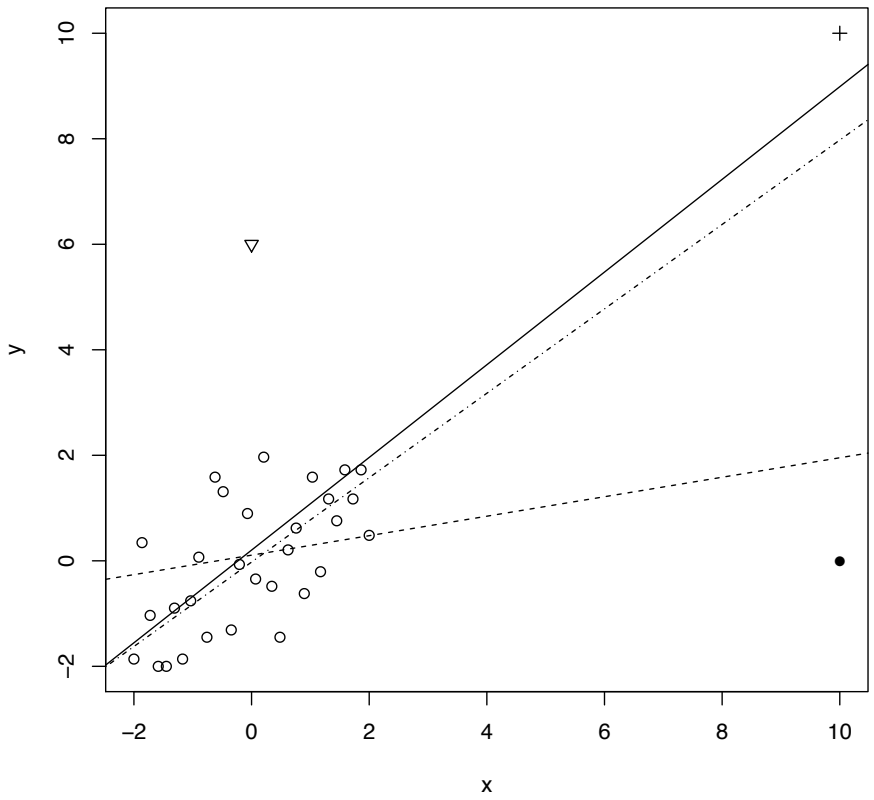
1. Outliers;
2. influential points.

Outliers tend to be unusual values in *Y.*

Influential (leverage) points tend to be unusual values in *X*.

Both can affect the fit of the regression model.

Outliers and influential points



Solid line: Fit excluding dot.

Dotted line: Fit excluding dot and cross.

Dashed line: Fit excluding cross.

Triangle: Outlier.

Cross and dot: High leverage points. Cross not an outlier (close to line).

Dot: Influential point (affects fit).

Outliers and influential points

Check for data-entry error first.

Examine physical context - why did it happen? May be a discovery of interest.

Exclude, and then include, the point in the analysis and compare the results. Outliers should always be reported.

Outliers and influential points

Consider robust regression if outliers cannot be reasonably identified as mistakes but are naturally occurring.

NEVER exclude outliers in an automatic manner! Why?

Outliers and influential points

NASA launched *Nimbus 7* satellite to record atmospheric information.

After years of operation the British Antarctic Survey observed a large decrease in atmospheric ozone over the Antarctic.

On closer examination of the data, it was found that the data processing programme automatically discarded extremely low observations (assumed to be mistakes).

The discovery of the Antarctic ozone hole was thus delayed by several years! Major implications for the planet?