

Topic: Exploratory Data Analysis (EDA)

The Statistical Process

School of Mathematics and Applied Statistics



UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Statistics involves . . .

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

Probability theory provides the foundation for all of the above.

What is Statistics?

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

Video: 6.23mins https://www.youtube.com/watch?v=wG8L_C20Mu8



Notes

-
-
-
-

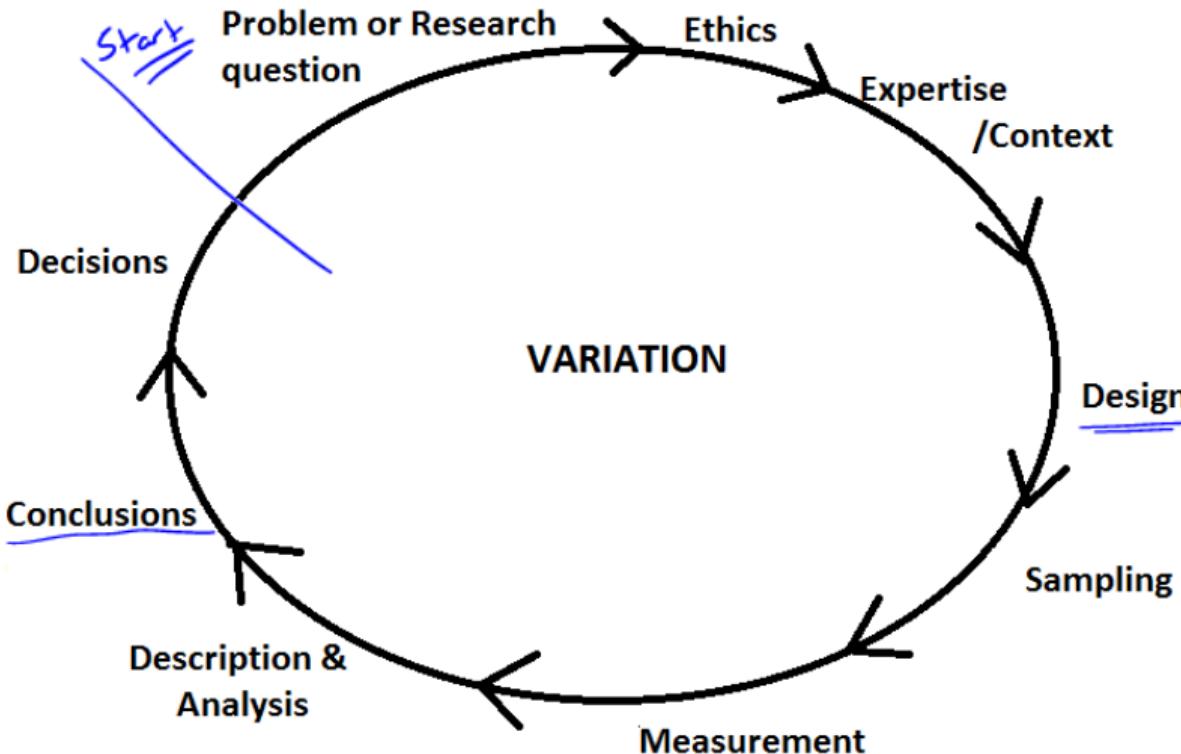
Statistics is ?

Statistics is a study of variability in the world around us

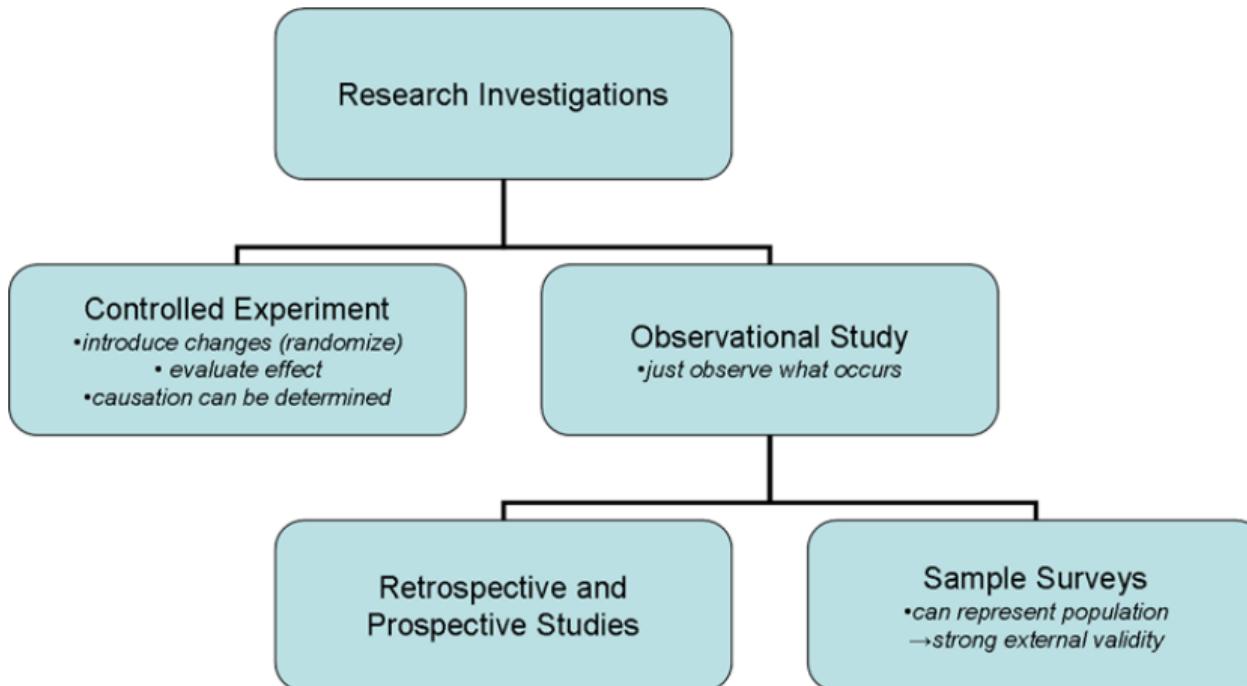
- What **brands** of mobile phone are most popular, least popular?
- How much **time** does the typical uni student spend on FaceBook per day?
- How many **texts** does the typical uni student send per day?
- What proportion of **emails** are spam?

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

Statistical Process



Research Designs



Issue: Quality of Evidence

Data Collection

- **Controlled experiments**

- Investigator introduces a change into a process
- Observes & evaluates effect of changes (variation)
- Evidence of causation if factors properly controlled and/or association
- Gold standard for evidence

- **Observational studies**

- Investigator does not interfere with the process
- Observes variation
- Evidence of association
- Weaker form of evidence

- **Simulations** *

Experimental Design

Simple between group design

- Randomly assign subjects to two groups
- Experimental and Control



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

Many different designs (some better than others)

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

Controlled Experiments - Example

Video Plant Experiment Mark Drollinger (2:34min)

<https://www.youtube.com/watch?v=VhZyXmgIFo>

My Notes:

-
-
-

Observational studies

Retrospective

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

Prospective

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

Observational Studies cont.

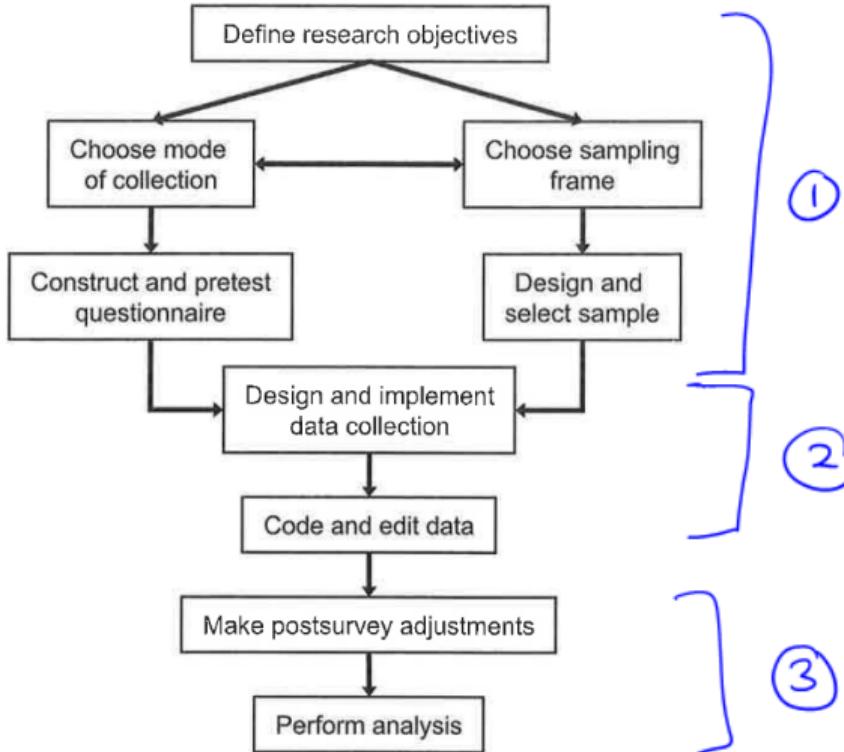
Sample surveys - common type of observational study

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

Three Phases in the Sample Survey Process

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

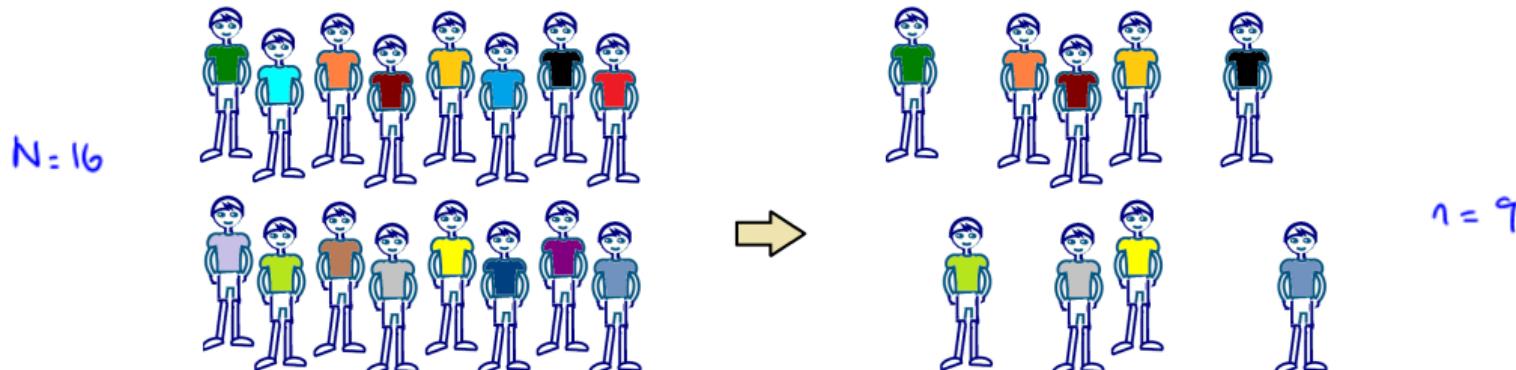
Sample Survey Process



Population vs Sample

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

- Eg people, businesses, hospitals, events etc.



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

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

Populations

Defining the population of interest

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

Eg: Who comprises the population of Doctors in Illawarra?

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

Sampling Frame

The Sampling Frame

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

Problems with lists

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

Survey Modes

Survey modes:

- Mail
- Telephone
- Field interview
- Internet

Sampling Designs

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

My notes:

- ★ • Simple Random Sampling

- Convenience Sampling
- Systematic Sampling
- Cluster Sampling
- Stratified Sampling

In Summary

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

Topic: Exploratory Data Analysis (EDA)

Data Types

School of Mathematics and Applied Statistics



UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Making sense of data

We need to know

- Context
- Units of measurement

How we analyse our data depends upon

- The structure of the data
- The type of measurement of our variables
- The questions being asked

Data set example

units

	A	B	C	D	E	F
Variables	(Units)	(C~)				
1	ID	Gender	EyeColour	LastDigitSnum	Height	Arm_Span
2	1	1	1	5	187	187
3	2	1	1	2	186	188
4	3	1	2	3	175	179.5
5	4	1	2	5	183	177.5
6	5	0	1	8	166	166
7	6	0	1	9	1780	170
8	7	1	1	4	188	184
9	8	1	1	1	190	190
10	9	1	1	0	171	169
11	10	1	1	3	178	170

MATH100_S118 Tutorial 4 Data

Basic Terms

- A **data set** is a collection of observations on one or more variables.
- An **observational unit** is the entity providing the information eg. student (row)
- A **variable** is a characteristic under study that assumes different values for different units eg Height (column)
- The item response for a unit is called an **observation** (cell)

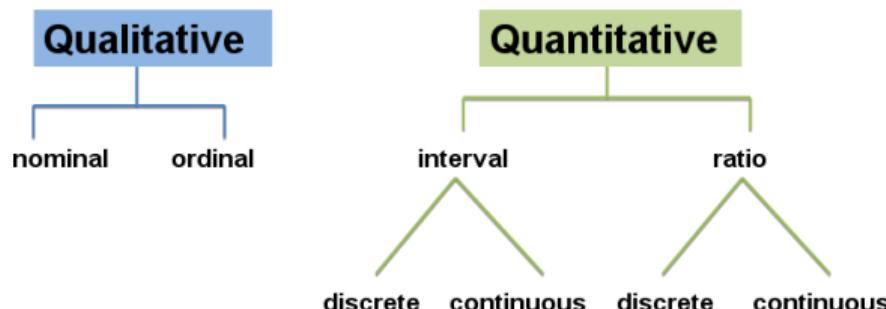
Types of Data

Qualitative Variable

A variable that can be classified into two or more categories is called a qualitative variable.

Quantitative Variable

A variable that can be measured numerically is called a quantitative variable.



Qualitative Data

Qualitative data is classified as **nominal** or **ordinal**.

Nominal (or Categorical) variable

A qualitative variable that can be classified into two or more categories which have no order.

eg. Own a bicycle: Y, N. *Binary \Rightarrow 2*

eg. Brand of mobile phone

eg. Country of birth

Ordinal variable

A qualitative variable that can be classified into two or more categories which have some order.

eg. Age-group *18-25, 26-40, 41-60*

eg. ATAR band

eg. Qn with Likert scale: Strongly agree - Strongly disagree

Quantitative Data

Quantitative data can be classified on an **interval** or **ratio** scale.

Interval scale

An interval scale is one in which the same difference between two values means the same thing everywhere on the scale, but ratios of differences are not meaningful.

An interval scale may have a zero, but it is not a true zero with respect to the property it is measuring.

eg. Temperature (degrees Celsius)



°F.

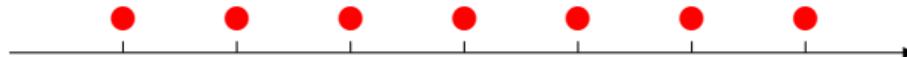
Ratio Scale

On a ratio scale something with twice the value has twice the property.

eg. Height: a child that is 1.4m tall is twice as tall as a child that is 70cm tall.

Discrete or Continuous

- If the possible values are separate points on the number line, a measurement is said to be **discrete**, e.g. no. of emails.



- The possible values of a **continuous** measurement form 1 or more intervals on the number line, e.g. length, weight.



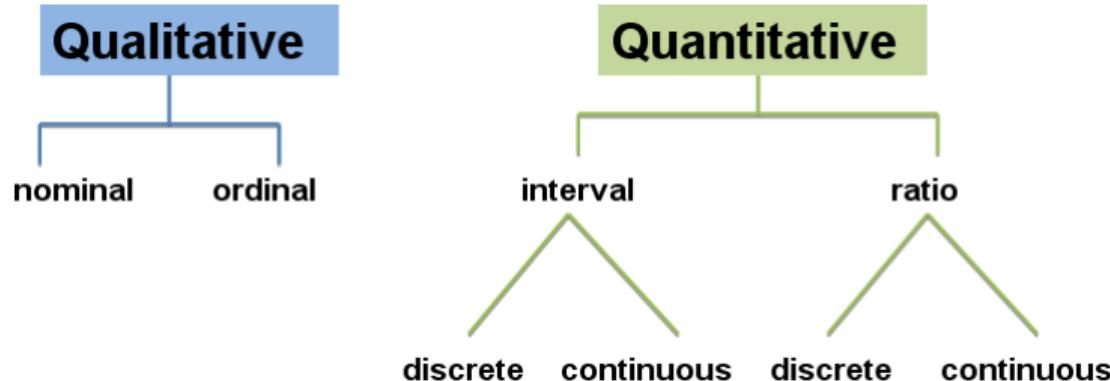
Note that length, weight are on a continuous scale even though it can only take positive values.

Activity

Classify the following variables

- **Handedness** Left / right. nominal \Rightarrow binary 2 categories.
- **Time** in mins spent on FaceBook on a particular day. quantitative / cont. \Rightarrow ratio
- Number of **texts** sent on a particular day. quant. \Rightarrow ratio.
 \Rightarrow discrete.
- **Coffee** cup size ordered: S/M/L. qual. \Rightarrow ordinal. 3 categories.
- **Coffee** type ordered: Capp/Latte/Flat white / ... etc qual. \Rightarrow nominal.

Types of Data - summary



Once we know the type of data we can choose the best way

- to summarise the data; and
- represent the data in tables and graphs.

i.e. How can we summarise and display the data effectively?

Topic: Exploratory Data Analysis (EDA)

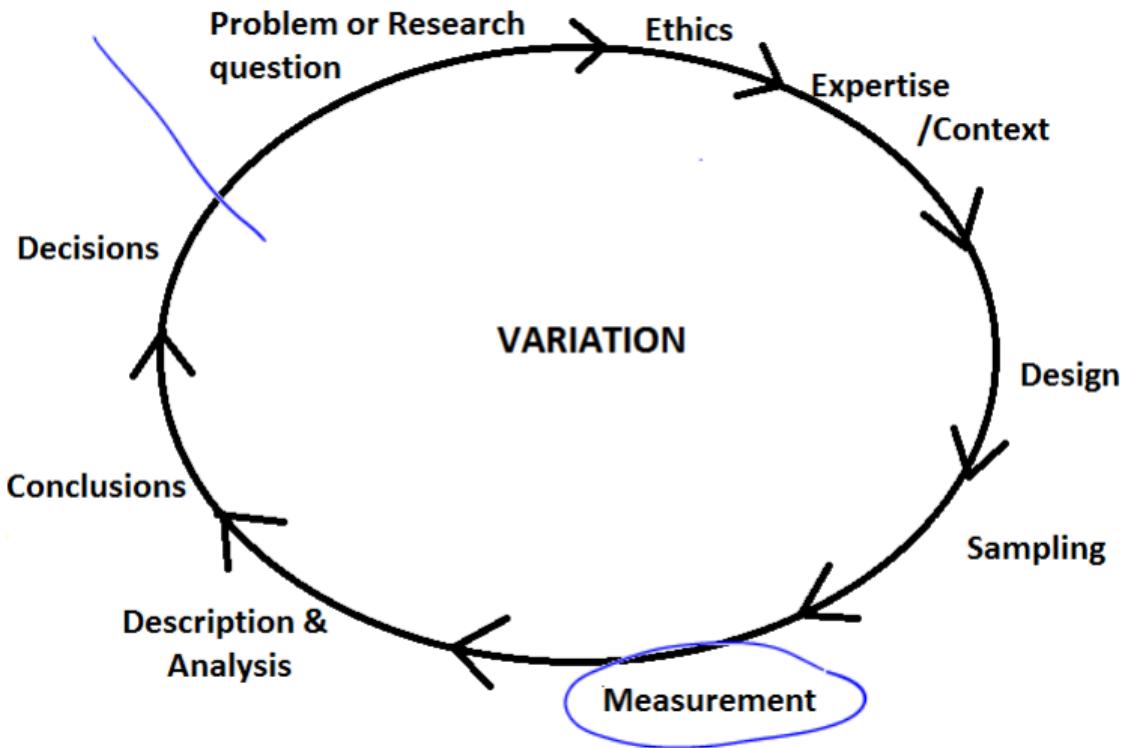
Measurement

School of Mathematics and Applied Statistics



UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Statistical Process or Problem Solving Process

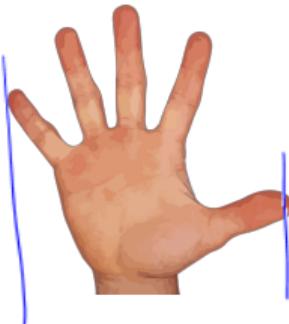


Measurement

May be

- **simple**

- measuring your hand span



One dimension

- **complex**

- measuring the spread/path of a bushfire, heat, etc



Activity: Physical Measurement

Measurement Exercise: Estimate the width (in metres) of the room.

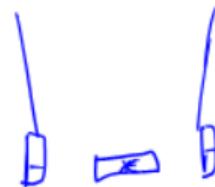
In a previous year, 2 volunteers were asked to estimate the width of the lecture theatre. *You might like to try this for the room you are currently in.*

Person 1 results:

- $< 9m$
- $\approx 9m$.

Person 2 results:

- $\approx 7m$
- $\approx 8m$



Discuss:

- What do you notice?
- How could the estimation be improved?
- What would be an appropriate measurement tool?



Measurement cont.

In our estimation example:

- there needed to be better specifications
- We need to know where to start and where to finish.

Measurement considerations:

- Measurement may be a source of variation, usually called measurement error.
- There is variation between people measuring same quantity
- There is some variation within measurements taken by one person
- Variation between measurers is often greater than within one person's measurements

Units of Measurement

International System of Units

Unit name	Symbol	Quantity
metre	m	length
kilogram	kg	mass
second	s	time
Kelvin	K	thermo dynamic temperature
Ampere	A	electric current
Mole	mol	amount of substance
Candela	cd	luminous intensity

Source: <http://physics.nist.gov/cuu/Units/units.html>

Other types of Measurement

Educational and psychological measurement



- There are many educational / psychological measures or scales eg intelligence (eg. WISC), depression, anxiety, cognitive impairment
- Many employers use psychological measurement for selection purposes
- Needs to be **valid** - does the questionnaire measure what it says it measures?
 - Are the questions culturally / gender/ age appropriate?
- Needs to be **reliable** - would the same measure be obtained if the questionnaire was repeated?
- Standardised tests go through a rigorous procedure of development

Measuring uncertainty - probability //

Measurement by Estimation

- There is variation in estimates
- Individual estimates are often unreliable \Rightarrow repetition.
- However the centre of the distribution may provide a good estimate

This leads us to the different types of **summary statistics**

Topic: Exploratory Data Analysis (EDA)

Measures of Centre

School of Mathematics and Applied Statistics



UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Statistical Process

- Ethics
- Nature of the question to be answered
- Context/Expertise
- Design:
 - Experiment vs. observational study
 - Sampling
 - Measurement
- **Description and analysis**
- Conclusions and decision making

VARIATION



Measures of Centre/Location Statistics

- **Population Mean:**

$$\text{'mu'} \quad \mu = \frac{1}{N} \sum_{i=1}^N x_i = \frac{1}{N} (x_1 + x_2 + \dots + x_N)$$

N = population size

- **Sample Mean:**

$$\text{'xbar'} \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} (x_1 + x_2 + \dots + x_n)$$

n = sample size

- **Trimmed Mean:** Average after eliminating a percentage of the highest and lowest.
Eg. some packages use 5%

- **Median:** Middle score when data values arranged in order from smallest to largest

- **Mode:** Most frequent score

Quantitative Data: Sample Mean

- Consider n quantitative data values $x_1, x_2 \dots, x_n$
- The **mean** is the *average* value:

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

Capital sigma

- Notation:** x_i denotes the *i*th value in the list (with no order), $i = 1, \dots, n$

- In R:

$n=5$

`x <- c(3, 1, 4, 5, 9)`

$$\bar{x} = \frac{1}{5} \times 22$$

`mean(x)`

$$= \frac{22}{5} = 4.4$$

4.4

Sample Median

To find the median, Q_2 , the *middle score*

$$x_{(1)} = \min$$

$$x_{(n)} = \max$$

- 1 Sort the n data values in ascending order: $x_{(1)}, x_{(2)}, \dots, x_{(n)}$
Note $x_{(i)}$ denotes the i th value in the list (with ascending order) $i = 1, \dots, n$.
- 2 Calculate $\frac{n+1}{2}$ \Rightarrow position / rank (of middle position)
- 3 Determine the observation which is in the $(\frac{n+1}{2})$ th position
- 4 Unlike the mean which can be pulled up or down by unusual data values (outliers), the median is only affected slightly by outliers: It is more **robust**.

- 5 In R:

```
x <- c(3,1,4,5,9)
median(x)
```

$$\begin{matrix} \textcircled{1} & \textcircled{2} & \textcircled{3} \\ 1, 3, \textcircled{4}, 5, 9 \end{matrix}$$

$Q_2 = 4.$

$$\frac{n+1}{2} = \frac{5}{2} = 3^{\text{rd}} \text{ value.}$$

Sample Median - Exercise

Determine median using the calculated rank:

- for **odd** n , the median is the *middle* sorted data value.

Ex: determine median of $\{6, 5, 8, 2, 9\}$ $\Rightarrow 2, 5, \boxed{6}, 8, 9$

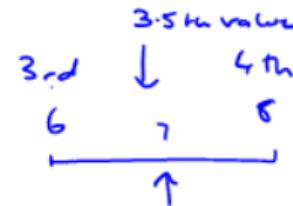
$$n=5 \quad \frac{n+1}{2} = 3\text{rd value}$$

$$\underline{Q_2 = 6.}$$

- for **even** n , the median is the average of the middle 2 data values.

Ex: determine median of $\{2, 5, \boxed{6}, 8, 9, 11\}$ $n=6$.

$$\frac{n+1}{2} = \frac{7}{2} = 3.5\text{th value}$$



$$\frac{6+8}{2} = 7 \quad Q_2 = 7.$$

Exercises

Determine the mean, median and mode of the following:

- ① Set 1: 1, 8, 4, 2, 7, 8 in order \Rightarrow 1, 2, 4, 7, 8, 8 $n=6$.

$$\bar{x}_1 = \frac{30}{6} = 5 \quad \text{mode} = 8$$

$$\frac{6+1}{2} = \underline{\underline{3.5^{\text{th}} \text{ value}}}$$

3rd 4th
4 7

$$Q_2 = \frac{4+7}{2} = 5.5$$

- ② Set 2: 1, 8, 4, 2, 7, 8, 40 $n=7$

1 2 4 7 8 8 40

$$\bar{x}_2 = \frac{70}{7} = 10 \quad \text{median}$$

$$\frac{n+1}{2} = \frac{8}{2} = 4^{\text{th}} \text{ value.}$$

mode = 8

$$\underline{\underline{Q_2 = 7.}}$$

$$Q_2 < \bar{x}_2$$

Exercise: Wollongong Monthly Average Temp: Jan 2009 - Jun 2018

Exercise: Find the median temperature from the stem-and-leaf plot $n = \underline{\underline{114}}$:

Wollongong Temperature (Monthly Average)
Stem-and-Leaf Plot

CF	Frequency	Stem & Leaf
1	1.00	16 . 6
12	11.00	17 . 11333357789
24	12.00	18 . 122333556789
28	4.00	19 . 0348
39	11.00	20 . 00244445789
45	6.00	21 . 012499
55	10.00	22 . 0033444689
68	13.00	23 . 1224567777889
	4.00	24 . 1679
	22.00	25 . 0123334455666677889999
	9.00	26 . 001256789
	6.00	27 . 014468
	4.00	28 . 0113
	1.00	29 . 8

Stem width: 1.0 29.8°C
Each leaf: 1 case(s)

$$\frac{n+1}{2} = \frac{115}{2} = 57.5 \text{ th value}$$

57th 58th
 23.2°C 23.2°C
 $Q_2 = 23.2^\circ\text{C}$

Leaf width = $\frac{1}{10}$ the stem width
 $= 0.1$

In R: Wollongong Monthly Average Temp: Jan 2009 - Jun 2018

↓ column name.

```
> median(Temps_Airport$Temp_Wollo)
[1] 23.2 ✓
```



```
> stem(Temps_Airport$Temp_Wollo)
```

The decimal point is at the |

16 | 6

16.6°C

17 | 11333357789

18 | 122333556789

19 | 0348

20 | 00244445789

21 | 012499

22 | 0033444689

23 | 1224567777889

24 | 1679

25 | 0123334455666677889999

26 | 001256789

27 | 014468

28 | 0113

29 | 8

Topic: Exploratory Data Analysis (EDA)

Measures of Variability - Part A

School of Mathematics and Applied Statistics



UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Measuring Variability: Motivating Example

Consider the following data sets: $n=7$

Data set 1: 55, 55, 55, 55, 55, 55, 55

$$\begin{array}{r} \text{Total} \\ 385 \end{array}$$

Data set 2: 47, 51, 54, 55, 56, 59, 63

$$\begin{array}{r} \text{Total} \\ 385 \end{array}$$

Data set 3: 39, 47, 53, 55, 57, 63, 71

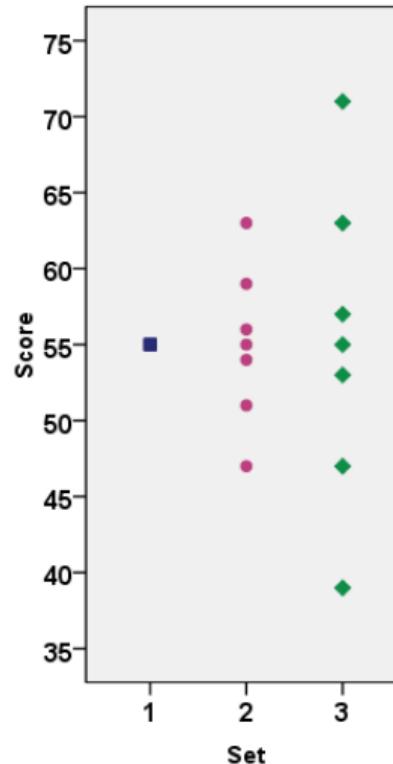
$$\begin{array}{r} \text{Total} \\ 385 \end{array}$$

For each data set

- Median = 55 = Q_2
- Mean = $385/7 = 55$ = \bar{x}

But the spread of the scores

vary.



How do we Measure Variability?

Variability (spread) can be measured by:

- **Variance σ^2 or s^2**

- uses all data values but is inflated by outliers

- **Standard deviation σ or s**

- uses all data values but is inflated by outliers

- **Range** = maximum – minimum = $x_{(n)}$ – $x_{(1)}$

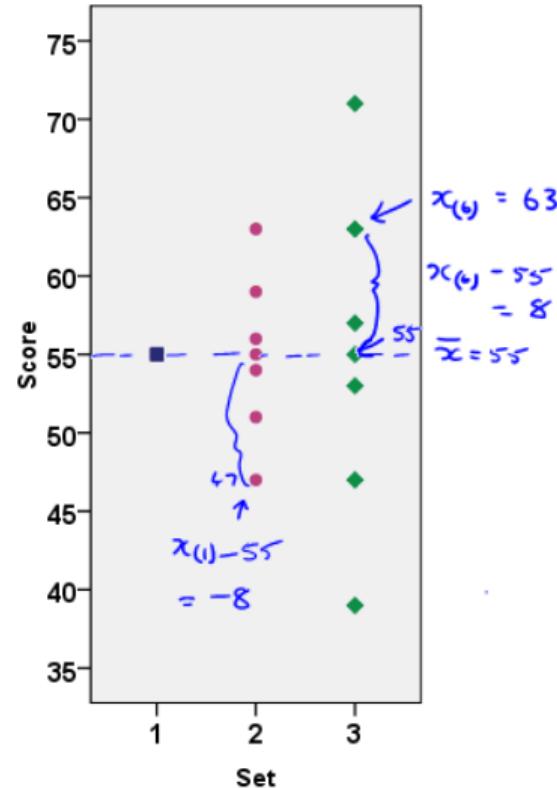
- unreliable measure, depends on extreme values

- **Interquartile range**: $IQR = Q_3 - Q_1$

- spans middle 50% of data,
 - unaffected by outliers, ignores variation in tails

Variance and Standard Deviation

- The **mean** is used as a reference point. $\bar{x} = 55$
- Consider the deviation from the i th point to the mean: $x_i - \bar{x}$
- Deviations may be positive or negative
- Variance** is based on the squared deviations.
- We can also describe the difference in spread using a notion of average distance from the mean.
- This measure of variability is called the **standard deviation**.
- sum of deviations = 0



Variance

Variance is based on the squared distances of individual data points from the mean.

- **Population Variance**

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

↓
 Sigma
 ↓
 σ^2

↓
 pop. mean.
 ↓
 mu.

- **Sample Variance**

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

↓
 Sample mean
 ↓

- measurement is in squared units
- is never negative, and
- is only zero when all data values are identical
- s^2 is an unbiased estimator of σ^2 . ✓✓

Σ

Standard Deviation

Standard deviation σ or s

- is the square root of the variance
- is measured in same units of measurement as data
- Population standard deviation**

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}$$

$$\sigma = \sqrt{\sigma^2}$$

Sample standard deviation

$$\downarrow s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

$$s = \sqrt{s^2}$$

- On calculator**, enter data then use: STAT mode

Pop
sd.

σ_n or $x\sigma_n$
 σ_x

or

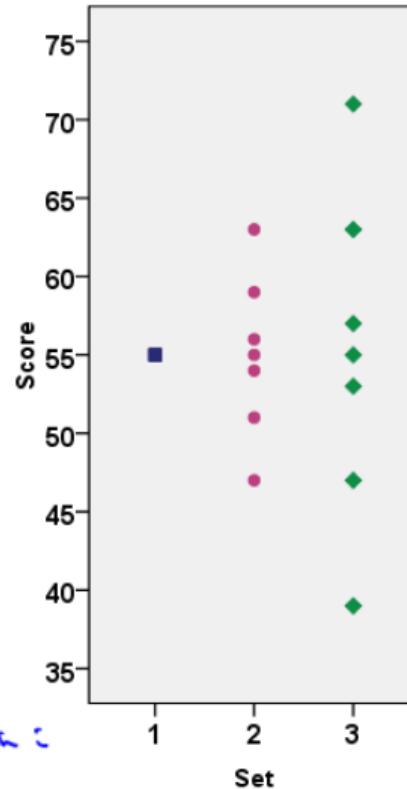
σ_{n-1} or $x\sigma_{n-1}$ or s_{n-1} or s key
 s_x

Sample sd.

Example

- Data set 1: 55, 55, 55, 55, 55, 55, 55
- Data set 2: 47, 51, 54, 55, 56, 59, 63
- Data set 3: 39, 47, 53, 55, 57, 63, 71

	Median	Mean	SD
Set 1	55	55	
Set 2	55	55	
Set 3	55	55	



Discuss: What is the SD for Set 1? Why?

$$s_1 = 0$$

$$(s)$$

$$x_i - \bar{x} = 0$$

$$x_i = 55 \text{ for each } i$$

$$\bar{x} = 55$$

Activity: Calculate Variance and SD

Exercise: What is the sample variance and standard deviation for Set 2 and Set 3?

Calculate s^2 and s :

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad s = \sqrt{s^2}$$

Data Set 2

$$\bar{x} = 55$$

47, 51, 54, 55, 55, 56, 59, 63

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
47	-8	64
51	-4	16
54	-1	1
55	0	0
55	1	1
56	4	16
59	8	64
63	0	0
$\sum_{i=1}^n (x_i - \bar{x})^2 =$		$\sum_{i=1}^n (x_i - \bar{x})^2 = 162$

Data Set 3

39, 47, 53, 55, 57, 63, 71

$$s^2_2 = \frac{1}{7-1} \times 162$$

$$= 27$$

$$s = \sqrt{27}$$

$$= 5.196\ldots$$

Activity: Calculate Variance and SD cont.

Set 3:

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
39	-16	256
47	-8	64
53	-2	4
55	0	0
57	2	4
63	8	64
71	16	256
		<u>648</u>

$$= \sum_{i=1}^7 (x_i - \bar{x})^2$$

$$s_3^2 = \frac{1}{7-1} \times 648$$

$$= \underline{\underline{108}}$$

$$s_3 = \sqrt{108}$$

$$= 10.392$$

s	Set 1	Set 2	Set 3
	0	5.196	<u>10.392</u>

In R: Calculate mean, variance and sd

```
> Set2 <- c(47, 51, 54, 55, 56, 59, 63)
```

```
> mean(Set2)
```

```
[1] 55 ✓
```

```
> var(Set2)
```

```
[1] 27 ✓
```

```
> sd(Set2)
```

```
[1] 5.196152 ✓
```

```
> Set3 <- c(39, 47, 53, 55, 57, 63, 71)
```

```
> mean(Set3)
```

```
[1] 55
```

```
> var(Set3)
```

```
[1] 108
```

```
> sd(Set3)
```

```
[1] 10.3923 ✓
```

Topic: Exploratory Data Analysis (EDA)

Measures of Variability - Part B

Quartiles, IQR & Box Plots

School of Mathematics and Applied Statistics



UNIVERSITY
OF WOLLONGONG
AUSTRALIA

How do we Measure Variability?

Variability (spread) can be measured by:

- ✓ • **Variance σ^2 or s^2**
 - uses all data values but is inflated by outliers
- ✓ • **Standard deviation σ or s**
 - uses all data values but is inflated by outliers
- ✓ • **Range** = maximum – minimum = $x_{(n)} - x_{(1)}$
 - unreliable measure, depends on extreme values
- * • **Interquartile range:** $IQR = Q_3 - Q_1$
 - spans middle 50% of data,
 - unaffected by outliers, ignores variation in tails

Interquartile Range (IQR)

- IQR = Upper quartile (Q_3) - lower quartile (Q_1)
- or IQR = 75^{th} - 25^{th} percentile
- There are different ways of calculating quartiles:
we will use the repeated median method
 - Q_1 = median of lower half of sorted data.
 - Q_3 = median of upper half of sorted data.
 - For n even, split the data into two halves - find median of lower half to get Q_1 ; find median of upper half to get Q_3 .
 - For n odd, leave Q_2 in both halves to find Q_1 and Q_3 .

↑
median

Five-number summaries

Data for a quantitative variable can be summarised by giving the following five numbers:

- | | |
|-----------------------|-----------------|
| ① the minimum value, | $x_{(1)}$ (min) |
| ② the lower quartile, | Q_1 (or LQ) |
| ③ the median, | Q_2 |
| ④ the upper quartile, | Q_3 (or UQ) |
| ⑤ the maximum value. | $x_{(n)}$ (max) |

The ordered set $(x_{(1)}, Q_1, Q_2, Q_3, x_{(n)})$ is the five-number summary of the data.

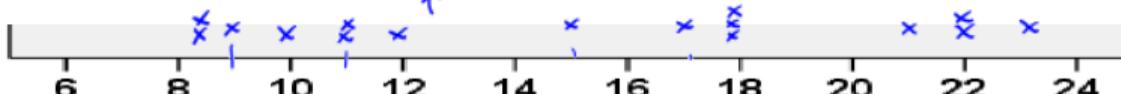
They can be used to construct box plots.

Exercise: Handspan Set 1

$n=16$
 Draw a dot plot and calculate the 5 number summary for handspan (right) for 16 students in a tutorial class.

8.5, 8.5, 9, 10, 11, 11, 12, 15, 17, 18, 18, 18, 21, 22, 22, 23 (cm)

Dot Plot:



For Q_2

$$\frac{n+1}{2} = \frac{17}{2} = 8.5^{\text{th}} \text{ value}$$

$$Q_2 = 16.$$

Q_3

4.5th from 17.

$$\underbrace{18 \quad 21}_{\downarrow}$$

$$\frac{18+21}{2} = \frac{39}{2} = 19.5$$

$$Q_3 = 19.5$$

For Q_1

$$\frac{n+1}{2} = \frac{9}{2} = 4.5^{\text{th}} \text{ value}$$

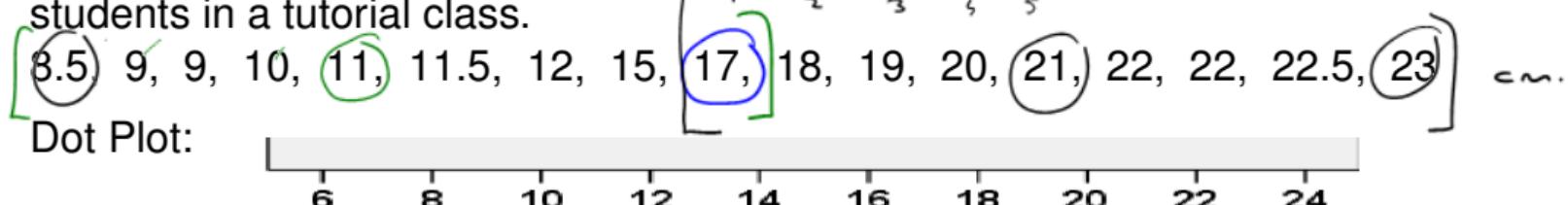
$$Q_1 = 10.5$$

5 no. summary is

$$(8.5, 10.5, 16, 19.5, 23) \text{ cm.}$$

Exercise: Handspan Set 2

$n=17$.
 Draw a dot plot and calculate the 5 number summary for handspan (right) for 17 students in a tutorial class.



Dot Plot:

$$\text{Median } Q_2$$

$$\frac{n+1}{2} = \frac{18}{2} = 9^{\text{th}} \text{ value}$$

$$Q_2 = 17.$$

$$\text{For } Q_3 \quad Q_3 = 21. \quad (\text{incl } 17).$$

5. no. summary is

$$(8.5, 11, 17, 21, 23) \text{ cm.}$$

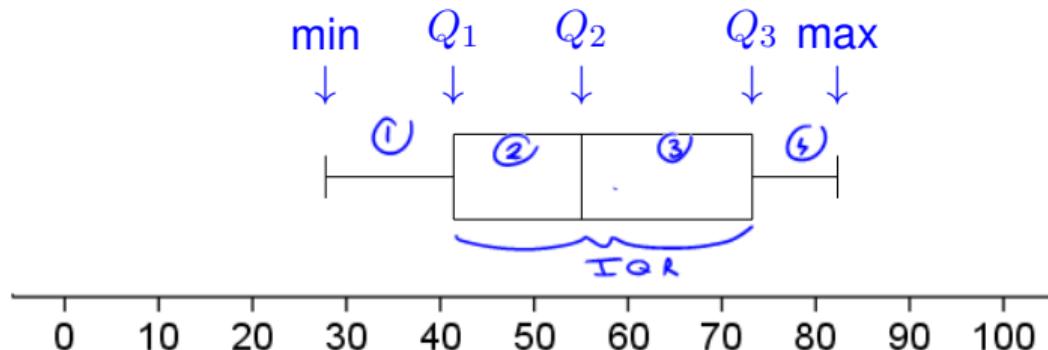
$$\text{For } Q_1$$

$$\frac{n+1}{2} = \frac{9+1}{2} = 5^{\text{th}} \text{ value}$$

$$Q_1 = 11$$

Basic box plots

The most basic **box plot** is a box-and-whisker diagram drawn alongside a scale to indicate the **five-number summary**.



The **interquartile range** ($=Q_3 - Q_1$) is the length of the central box.

Question: What proportion of points lie within each section of the box plot? 25%

Box plots

- Centre and spread can be seen at a glance.
- Width of box (if drawn horizontally) or Height of box (if drawn vertically) is IQR.
- Shows whether the distribution is
 - approximately symmetric (equal whiskers, crossbar in the middle of the box) or
 - not symmetric so it is skewed.
- Can plot boxplots side-by-side on same scale when comparing 2 or more groups.
- Outliers can be plotted separately as dots.

Quartiles in R

R code: In R, the repeated median method is implemented in the : **fivenum**

- For n even: ($n=16$)

```
x <- c(8.5, 8.5, 9, 10, 11, 11, 12, 15, 17, 18, 18, 18, 21, 22, 22, 23)
fivenum(x)
[1] 8.5 10.5 16.0 19.5 23.0
     min   Q1    Q2    Q3    max
           ↑      ↑      ↑      ↑
           4th element  2nd element
```

✓ agree

```
IQRx <- fivenum(x) [4] - fivenum(x) [2]
```

```
IQRx
```

```
[1] 9
```

✓

$$\begin{aligned} IQR &= Q_3 - Q_1 \\ &= 19.5 - 10.5 \\ &= 9. \end{aligned}$$

Quartiles in R cont.

R code: In R, the repeated median method is implemented in the : `fivenum` function

For n odd: ($n=17$)

```
y<-c(8.5, 9, 9, 10, 11, 11.5, 12, 15, 17, 18, 19, 20, 21, 22, 22, 22.5, 23)
fivenum(y)
```

```
[1] 8.5 11.0 17.0 21.0 23.0      ✓
    min   Q1   Q2   Q3   max
```

$$\text{IQR} = 21 - 11 \\ = 10$$

```
IQRy <- fivenum(y)[4] - fivenum(y)[2]
```

```
IQRy
```

```
[1] 10      ✓
```

Quartiles in R - a different method

For your information, another widely used definition is to use the ranks where:

Q_1 is the $\frac{(n+3)}{4}^{th}$ observation; Q_3 is the $\frac{(3n+1)}{4}^{th}$ observation.

In R, this method is implemented in the **quantile** function

```
quantile(x)
```

0%	25%	50%	75%	100%
8.50	10.75	16.00	18.75	23.00

```
quantile(y)
```

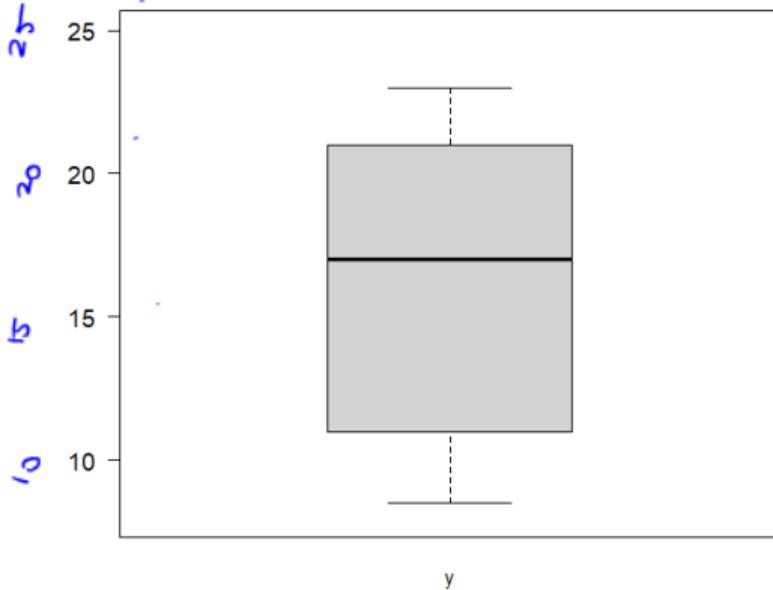
0%	25%	50%	75%	100%
8.5	11.0	17.0	21.0	23.0

In R: Box Plots

R code:

`boxplot(y)` draws a single boxplot of data y .

```
boxplot(y, xlab="y",  
        ylim =c(8, 25), las=1) )
```

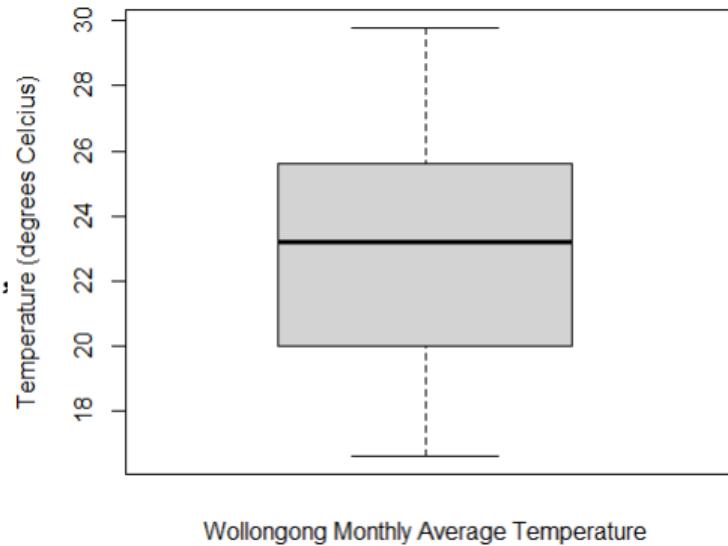


In R: Box Plots

R code:

Add options for labels: $n = 114$

```
boxplot(Temps_Airport$Temp_Wollo,  
ylab="Temperature (degrees Celcius)",  
xlab="Wollongong Monthly Average  
Temperature")
```



Topic: Exploratory Data Analysis (EDA)

Measures of Variability - Part C

Identifying Outliers & Errors

School of Mathematics and Applied Statistics



UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Outliers on box plots

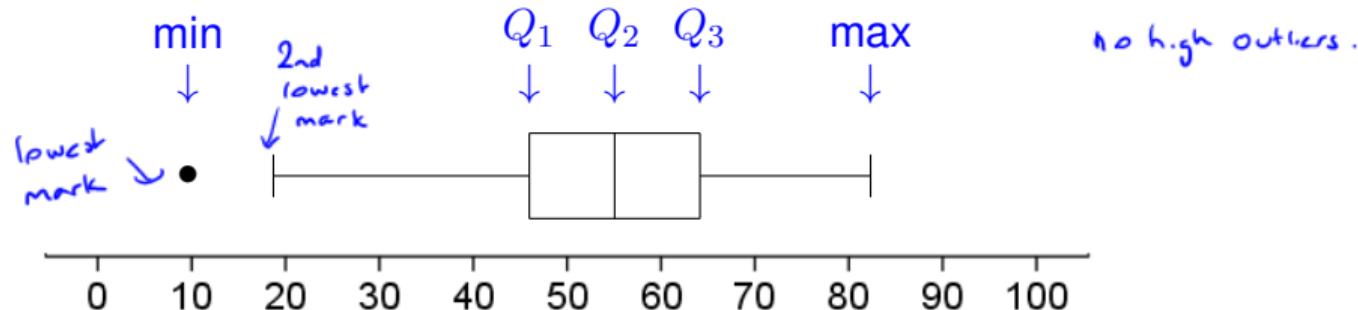
Outliers

A data point is identified as an **outlier**

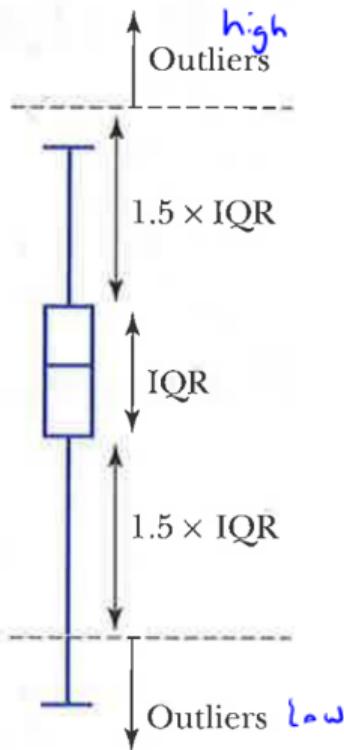
if it is more than $1.5 \times \text{IQR}$ beyond the upper or lower quartiles

It is marked separately with a **dot** (usually) or a cross.

The **whiskers** are then drawn only as far as the most extreme points which are not outliers.



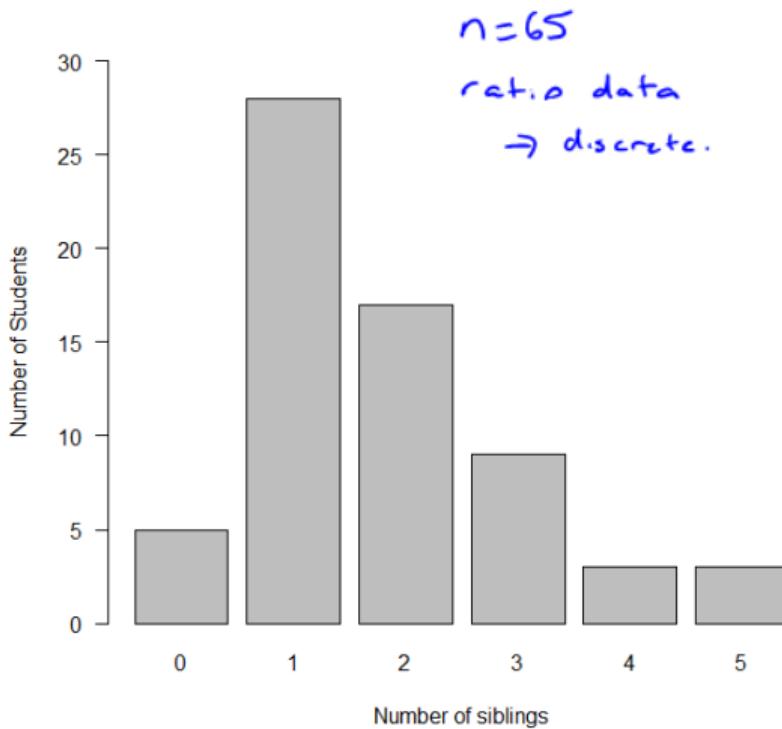
Identifying Outliers



Steps to identifying outliers:

- 1 Sort the n data values in ascending order or create a stem-and-leaf plot
- 2 Determine the five number summary
- 3 Calculate $IQR = Q_3 - Q_1$ ←
- 4 Calculate bounds for low/high outliers
 - Low: bound is $Q_1 - 1.5 \times IQR$
 - High: bound is $Q_3 + 1.5 \times IQR$
- 5 Check for data values outside these bounds:
 - Are there any $x_{(i)} < Q_1 - 1.5 \times IQR$
⇒ low outliers
 - Are there any $x_{(i)} > Q_3 + 1.5 \times IQR$
⇒ high outliers

Example 1: Number of Siblings



No. of Siblings	0	1	2	3	4	5
Frequency	5	28	17	9	3	3
Cum. Freq.	5	33	50	59	62	65

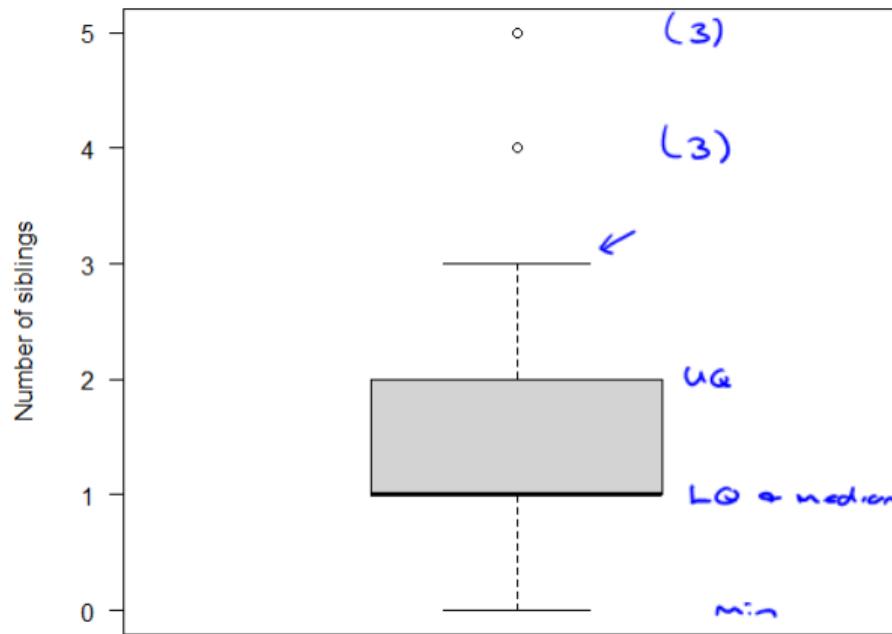
- Range = $x_{(n)} - x_{(1)} = 5 - 0 = 5$
- Median
 $Q_2 = 1$
 $\frac{n+1}{2} = \frac{66}{2} = 33\text{rd value}$.
- Quartiles:
 $Q_1 = 1$
 $\frac{n+1}{4} = \frac{34}{4} = 17\text{th value}$
- $Q_3 = 2$

Example 1: Number of Siblings cont.

- 5-number summary: $(0, Q_1, 1, Q_3, 5)$
- Interquartile Range:
 - $IQR = Q_3 - Q_1 = 2 - 1 = 1$
- Calculate bounds:
 - Low bound is: $Q_1 - 1.5 \times IQR = 1 - 1.5 \times 1 = -0.5$ N/A.
 - High bound is: $Q_3 + 1.5 \times IQR = 2 + 1.5 \times 1 = 3.5$
- Identify outliers:
 - No low outliers
 - High outliers : 4, 4, 4, 5, 5, 5.

Example 1: Number of Siblings cont.

Box plot for Number of Siblings for 65 Students



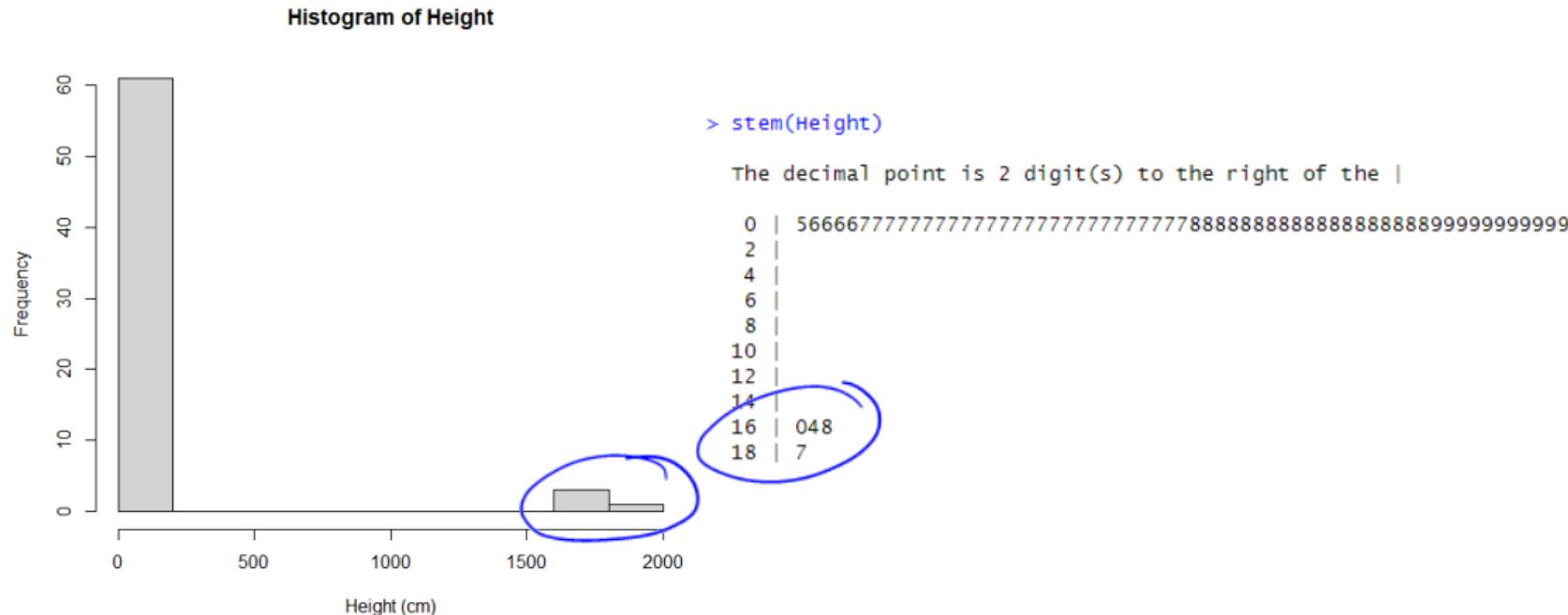
Identifying Errors

Checks

- Outliers are unusual values in the data set
- Check if they are
 - errors- correct if possible - go back to source
 - valid observations - do not usually discard - can check the affect on analysis

Example 2: Height (cm)

The heights in cm were measured for 65 students in a Maths subject.
Check the data by creating a stem & plot and/or a histogram.



Example 2: Height (cm) cont.

Check data: Identify any errors:

```
Heightsort<-sort(Height)
print(Heightsort[55:65])
[1] 187.0 188.0 188.5 190.0 190.0 194.0 194.0 1700.0
[9] 1740.0 1780.0 1866.0                                     mm
```

What should we do?

correct the obs from mm into cm.

Example 2: Height (cm) cont.

Correct those observations that were recorded in mm to be in cm:

```
Heighthv2 <- c(Heightsort[1:61], Heightsort[62:65]/10)
```

```
print(Heighthv2)
```

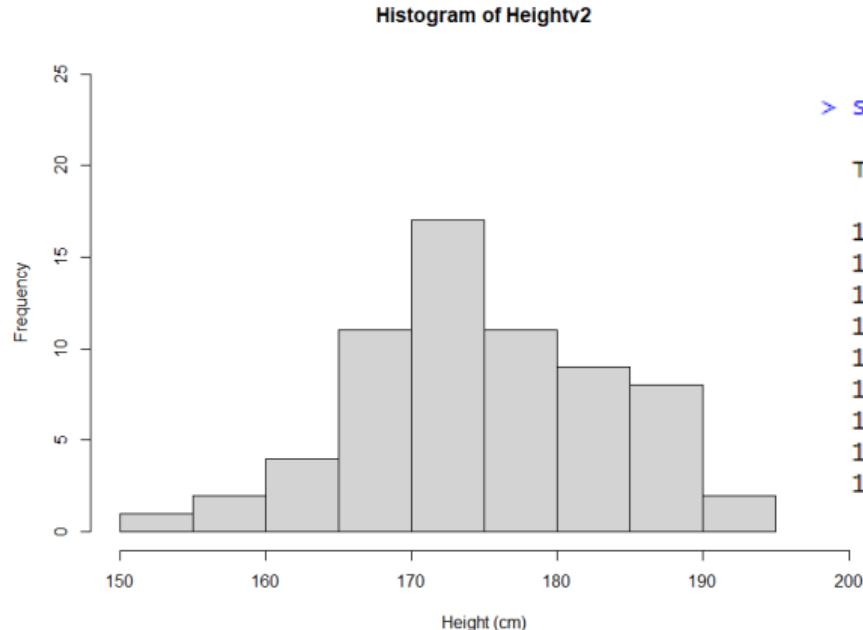
```
[1] 151.0 157.0 157.0 161.4 163.2 165.0 165.0 166.0 166.0 167.0 168.0  
[12] 168.0 168.1 169.0 170.0 170.0 170.0 170.5 171.0 171.0 171.0 171.0  
[23] 171.0 171.0 171.0 172.0 172.0 172.0 172.5 173.0 174.0 175.0 175.0  
[34] 176.0 176.8 177.0 177.0 177.0 177.0 178.0 178.0 179.0 180.0 180.5  
[45] 182.0 182.0 182.0 182.0 183.0 184.2 185.0 185.0 186.0 186.2 187.0  
[56] 188.0 188.5 190.0 190.0 194.0 194.0 170.0 174.0 178.0 186.6
```

cm

now corrected.

Example 2: Height (cm)- corrected

Redo the plots:



```
> stem(Heightv2)
```

The decimal point is 1 digit(s) to the right of the |

15 1	151.
15 77	
16 13	
16 556678889	
17 000011111112223344	
17 556777778889	
18 01222234	
18 55667789	
19 0044	

190 190 194 194 cm

Example 2: Height (cm)- corrected

```
fivenum(Heightv2)  
[1] 151 170 174 182 194  
min Q1 Q2 Q3 max.
```

```
IQRht <- fivenum(Heightv2) [4] - fivenum(Heightv2) [2]
```

IQRht

```
[1] 12
```

```
rangeHt <- fivenum(Heightv2) [5] - fivenum(Heightv2) [1]
```

rangeHt

```
[1] 43
```

Example 2: Height - corrected

- Range = $x_{(n)} - x_{(1)} = \underline{194} - \underline{151} = 43 \text{ cm}$

- Median

$$Q_2 = \underline{174} \text{ cm}$$

confirm the R values
using hand calc's.

- Quartiles:

$$Q_1 = \underline{170} \text{ cm}$$

$$Q_3 = \underline{182} \text{ cm}$$

Example 2: Height - corrected

- 5-number summary: $(\min, 151, 170, 174, 182, 194) \text{ cm.}$

- Interquartile Range:

- $IQR = Q_3 - Q_1 = 182 - 170 = 12 \text{ cm}$

- Calculate bounds:

- Low bound is: $170 - 1.5 \times 12 = 152 \text{ cm.}$

- High bound is: $182 + 18 = 200 \text{ cm} \Rightarrow x$

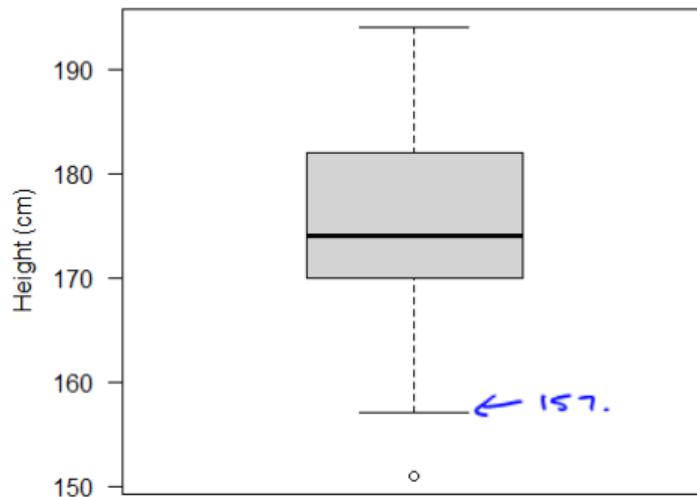
- Identify outliers:

- $1 \text{ low outlier} = 151 \text{ cm}$

- No high outliers.

Example 2: Height - corrected cont.

Box plot for Height for 65 Students



Topic: Exploratory Data Analysis (EDA)

Linear Transformations

School of Mathematics and Applied Statistics



UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Linear Transformations

Ethics

Nature of the question to be answered

(Transforms – data from different perspectives)

Context/ Expertise

Design: Experiments Vs Observation

Sampling

Measurement

Description and Analysis

Conclusions & Decision Making

VARIATION

Activity: Transformation of Measurement Units

A sample of data was collected from students: an excerpt is shown

Height	Shoe Size	Sex
153	8	f
6'	11	m
180cm X	15	m
170cm X	9.5	m

Discuss:

- What do you notice?

$$6' = 6 \text{ feet.}$$

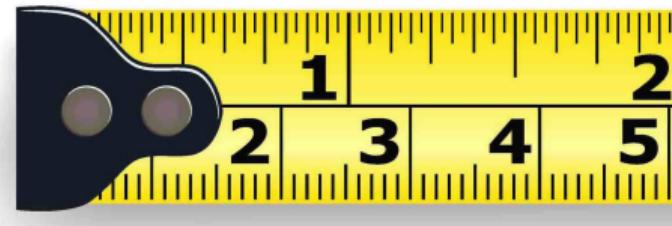
- How could you fix the problem?

convert 6' (feet) into cm.

- What do you need to know to fix the problem?

convert feet \rightarrow inches.
inches \rightarrow cm

Activity: Transforming Height



We know: 1 inch = 2.54 cm and 1 foot = 12 inches

$$6 \text{ feet} = \underline{6 \times 12 = 72} \text{ inches}$$

$$= \underline{72 \times 2.54 = 182.9} \text{ cm}$$

This is an example of a linear transformation or a rescaling of measurements

Transforming Data: Mean

Discuss: What happens to the mean if each data value x_i is rescaled by a **linear transformation** $a + bx_i$?

Example:

- Values of x : $\{1, 2, 3\}$, then $\bar{x} = 2$
- Let $y_i = a + bx_i$ with $a = 10$ and $b = 3$

then values of y :

$$\{10 + 3 \times 1, 10 + 3 \times 2, 10 + 3 \times 3\} = \{13, 16, 19\},$$

and $\bar{y} = \frac{13 + 16 + 19}{3} = \frac{48}{3} = 16.$

$$\begin{aligned}\bar{y} &= a + b\bar{x} \\ &= 10 + 3 \times 2 \\ &= 10 + 6 = \underline{\underline{16}}\end{aligned}$$

So, the mean is rescaled in the same way as the values of x

Rescaling Data: Prove mathematically for mean

If each data value $x_i, i = 1 \dots n$ is **rescaled by a linear transformation** such that $y_i = a + bx_i$, show that $\bar{y} = a + b\bar{x}$

$$\begin{aligned}
 y_i &= a + bx_i \\
 \text{start } \bar{y} &= \frac{1}{n} \sum_{i=1}^n y_i \\
 &= \frac{1}{n} \sum_{i=1}^n (a + bx_i) \\
 &= \frac{1}{n} [na + b \sum_{i=1}^n x_i] \\
 &= a + b \boxed{\frac{1}{n} \sum_{i=1}^n x_i} \\
 \bar{y} &= a + b \bar{x} \quad \checkmark
 \end{aligned}$$

Rescaling Data: Properties of standard deviation

- What happens to the standard deviation when **a constant is added to or subtracted from** each data value?
- Eg. the standard deviation of $\{2, 4, 6\}$ compared with the standard deviation of $\{1, 3, 5\}$ or $\{8, 10, 12\}$.



Rescaling Data: Properties of standard deviation

- What happens to the standard deviation when **a constant is added to or subtracted from** each data value?
- Eg. the standard deviation of $\{2, 4, 6\}$ compared with the standard deviation of $\{1, 3, 5\}$ or $\{8, 10, 12\}$.



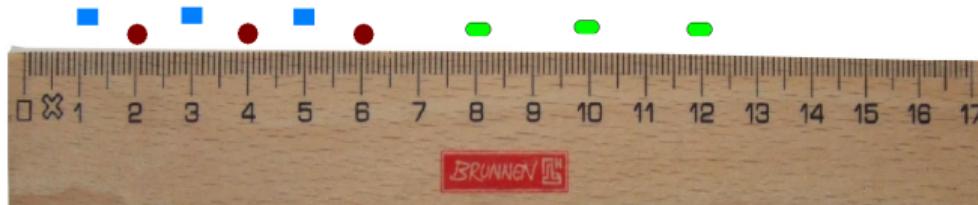
Rescaling Data: Properties of standard deviation

- What happens to the standard deviation when **a constant is added to or subtracted from** each data value?
- Eg. the standard deviation of $\{2, 4, 6\}$ compared with the standard deviation of $\{1, 3, 5\}$ or $\{8, 10, 12\}$.



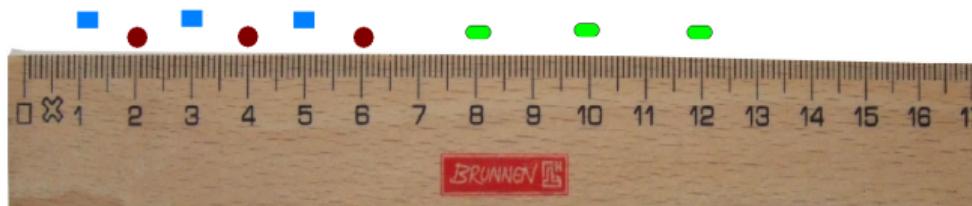
Rescaling Data: Properties of standard deviation

- What happens to the standard deviation when **a constant is added to or subtracted from** each data value?
- Eg. the standard deviation of $\{2, 4, 6\}$ compared with the standard deviation of $\{1, 3, 5\}$ or $\{8, 10, 12\}$.



Rescaling Data: Properties of standard deviation

- What happens to the standard deviation when **a constant is added to or subtracted from** each data value?
- Eg. the standard deviation of $\{2, 4, 6\}$ compared with the standard deviation of $\{1, 3, 5\}$ or $\{8, 10, 12\}$.



- The standard deviation of $\{2, 4, 6\}$ is the same as the standard deviation of $\{1, 3, 5\}$ or $\{8, 10, 12\}$.
- Standard deviation is unaffected when a constant is added to or subtracted from each data value.
- Prove mathematically...

Rescaling Data: Properties of standard deviation

- What happens to the standard deviation when each data value is multiplied by a constant c ?
- Eg. if data values $\{1, 3, 5\}$ are multiplied by 3: $\{3, 9, 15\}$.



Rescaling Data: Properties of standard deviation

- What happens to the standard deviation when each data value is multiplied by a constant c ?
- Eg. if data values $\{1, 3, 5\}$ are multiplied by 3: $\{3, 9, 15\}$.



Rescaling Data: Properties of standard deviation

- What happens to the standard deviation when each data value is multiplied by a constant c ?
- Eg. if data values $\{1, 3, 5\}$ are multiplied by 3: $\{3, 9, 15\}$.



Rescaling Data: Properties of standard deviation

- What happens to the standard deviation when each data value is multiplied by a constant c ?
- Eg. if data values $\{1, 3, 5\}$ are multiplied by 3: $\{3, 9, 15\}$.



- For example, the standard deviation of $\{1, 3, 5\}$ is 2, the standard deviation of $\{3, 9, 15\}$ is 6.

Rescaling Data: Properties of standard deviation

- What happens to the standard deviation when each data value is multiplied by a constant c ?
- Eg. if data values $\{1, 3, 5\}$ are multiplied by 3: $\{3, 9, 15\}$.



- For example, the standard deviation of $\{1, 3, 5\}$ is 2, the standard deviation of $\{3, 9, 15\}$ is 6.
- In fact, the standard deviation is multiplied by the constant $|c|$

$$s_{\text{new}} = |c| \times s_{\text{old}}$$

- Prove mathematically...

Rescaling Data: Prove mathematically for sd

If each data value $x_i, i = 1 \dots n$ is **rescaled by a linear transformation** such that $y_i = a + bx_i$, show that $s_y = \sqrt{b^2 s_x^2} = |b| s_x$

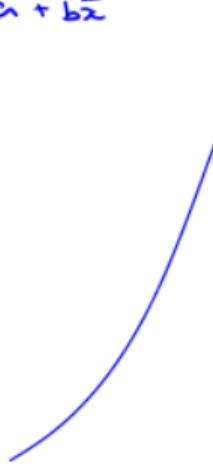
$$s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \quad \bar{y} = a + b\bar{x}$$

$$= \frac{1}{n-1} \sum_{i=1}^n [a + bx_i - (\cancel{a} + b\bar{x})]^2$$

$$= \frac{1}{n-1} \sum_{i=1}^n [b(x_i - \bar{x})]^2$$

$$= b^2 \times \boxed{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

$$s_y^2 = b^2 \times s_x^2$$



$$s_y = \sqrt{b^2 s_x^2}$$

$$= |b| \times s_x$$

\checkmark
as
Required.

$$s_x \geq 0$$

$$s_y \geq 0$$

Topic: Exploratory Data Analysis (EDA)

Nonlinear Transformations

School of Mathematics and Applied Statistics

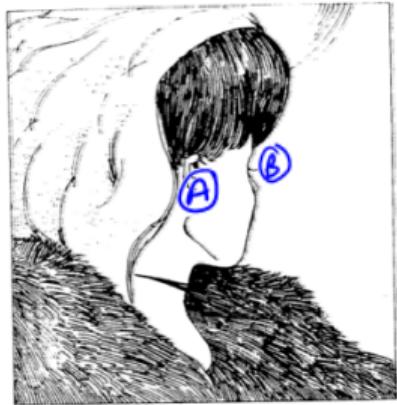


UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Why Transform Data?

- Convert units (a linear transformation) $y_c = a + bx_i$
- To see data from different perspectives
- Spread out dense clusters
- Contract gaps between values in one tail
- Reduce asymmetry and make numerical summaries representative of data
- To make curved lines straight
- To fulfill assumptions underlying statistical tests about data

Different Perspectives: What do you see?

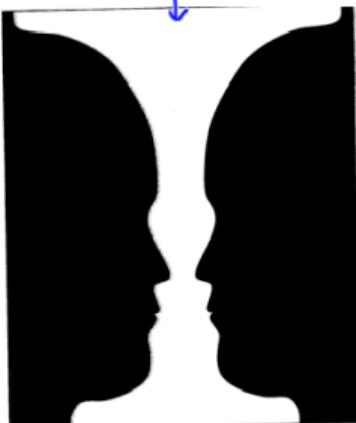


Weiten, 1989, p125

Different Perspectives: What do you see?



Weiten, 1989, p125



Weiten, 1989, p123

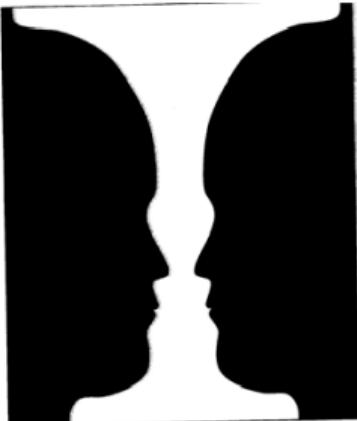
profiles

Transformations allow us to see data from different perspectives

Different Perspectives: What do you see?



Weiten, 1989, p125



Weiten, 1989, p123



Premonition, 2007 <http://www.finerminds.com/metaphysical/premonition/>

Transformations allow us to see data from different perspectives

Nonlinear Transformations

Transform each data point x_i by taking the square root i.e. $y_i = \sqrt{x_i}$



Discuss: What is the effect of taking the square root for these data points?

It _____ a tail of high values

In reverse:

Discuss: What is the effect of squaring each value? Consider $x_i = y_i^2$

It _____ the upper tail values

Nonlinear Transformations

Transform each data point x_i by taking the square root i.e. $y_i = \sqrt{x_i}$



Discuss: What is the effect of taking the square root for these data points?

It contracts a tail of high values
or pulls in

In reverse:

Discuss: What is the effect of squaring each value? Consider $x_i = y_i^2$

It stretches the upper tail values \Rightarrow *larger spread.*

Example: Brain Weights of mammals

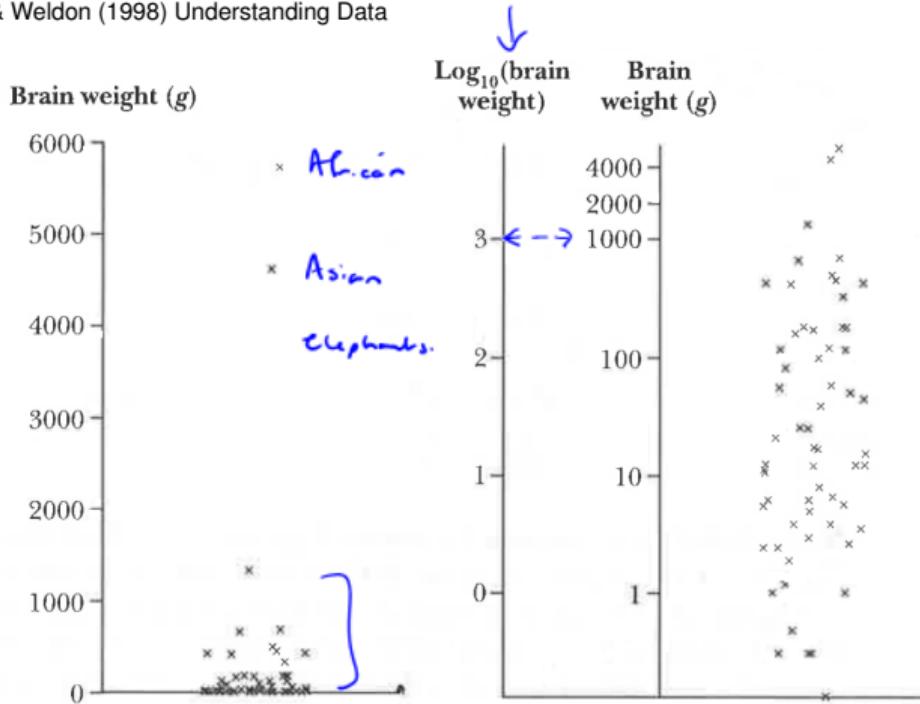
The average brain weights
of 62 species of mammals

Species	Brain weight (g)	\log_{10} (brain wt)	Species	Brain weight (g)	\log_{10} (brain wt)
Arctic fox	44.50	1.648	Human	1320	3.121
Owl monkey	15.50	1.190	African elephant	5712	3.757
Mountain beaver	8.100	0.908	Water opossum	3.900	0.591
Cow	423.0	2.626	Rhesus monkey	179.0	2.253
Gray wolf	119.5	2.077	Kangaroo	56.00	1.748
Goat	115.0	2.061	Yellow-bellied marmot	17.00	1.230
Roe deer	98.20	1.992	Golden hamster	1.000	0.000
Guinea pig	5.500	0.740	Mouse	0.400	-0.398
Vervet	58.00	1.763	Little brown bat	0.250	-0.602
Chinchilla	6.400	0.806	Slow loris	12.50	1.097
Ground squirrel	4.000	0.602	Okapi	490.0	2.690
Arctic ground squirrel	5.700	0.756	Rabbit	12.10	1.083
African giant pouched rat	6.600	0.820	Sheep	175.0	2.243
Lesser short-tailed shrew	0.140	-0.854	Jaguar	157.0	2.196
Star-nosed mole	1.000	0.000	Chimpanzee	440.0	2.643
Nine-banded armadillo	10.80	1.033	Baboon	179.5	2.254
Tree hyrax	12.30	1.090	Desert hedgehog	2.400	0.380
N. American opossum	6.300	0.799	Giant armadillo	81.00	1.908
Asian elephant	4603	3.663	Rock hyrax (<i>P. habess.</i>)	21.00	1.322
Big brown bat	0.300	-0.523	Raccoon	39.20	1.593
Donkey	419.0	2.622	Rat	1.900	0.279
Horse	655.0	2.816	E. American mole	1.200	0.079
European hedgehog	3.500	0.544	Mole rat	3.000	0.477
Patas monkey	115.0	2.061	Musk shrew	0.330	-0.481
Cat	25.60	1.408	Pig	180.0	2.255
Galago	5.000	0.699	Echidna	25.00	1.398

Example: Brain Weights of mammals

The average brain weights of 62 species of mammals

Source: p.39 Griffiths, Stirling & Weldon (1998) Understanding Data

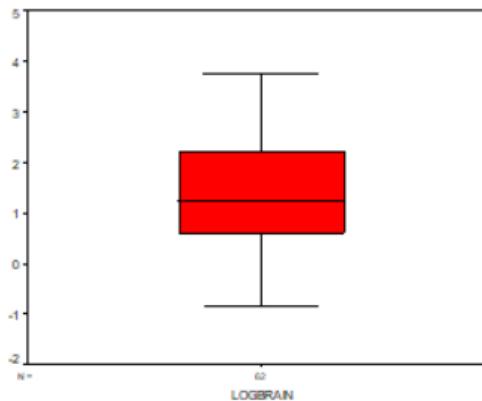
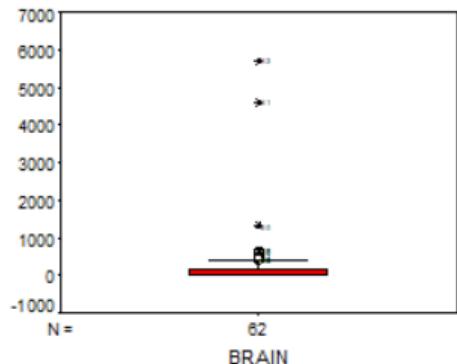


Example: Brain Weights of mammals

The average brain weights of 62 species of mammals:

Original scale (g) and

after applying a log (base 10) transformation:



Discuss: What effect has the transformation had?

Power Transformations of Data

p		Effect on whole positive values
2	square	Pushes out upper tail and contracts lower tail
1	identity	leaves unchanged
1/2	square root	Contracts upper tail of whole numbers, pulls in outliers & spreads out lower tail
1/3	Cube root	Contracts upper tail of whole numbers
-1	reciprocal	Order of magnitude is reversed
-1/2	Reciprocal square root	Order of magnitude is reversed
-2	reciprocal square	Order of magnitude is reversed

$$\ln\left(\frac{x}{1-x}\right)$$

Source: p.40 Griffiths, Stirling & Weldon (1998) Understanding Data

Transformations - Reporting

It is important to:

- Always state clearly that a transformation has been applied
- Data values should usually be transformed back when results are reported.
- Example: Brain weights

Median of approximately 1.2 is in log base 10 units

So to get the median in the original units we have $10^{1.2} = 15.85 \text{ g}$

Example: Richter scale

An example of a **nonlinear transformation** is the Richter scale

- transforms the measured intensity of earthquakes to a logarithmic scale

Video: How does the Richter Scale work? 4.56mins

<https://www.youtube.com/watch?v=NaNw9LHq9dc>

Notes: 25 April 2015 Nepal Earthquake

- _____ on Richter scale
- _____ on Modified Mercalli scale

This complicated cocktail of factors have to be considered to help determine how earthquakes are experienced & categorized.

Z score transformation

This **transforms** the variable X into variable Z such that

$$z_i = \frac{x_i - \bar{x}}{s_x} = \frac{x_i - \bar{x}}{\frac{s_x}{s_x}} = \left(\frac{-\bar{x}}{s_x} \right) + \left(\frac{1}{s_x} x_i \right)$$

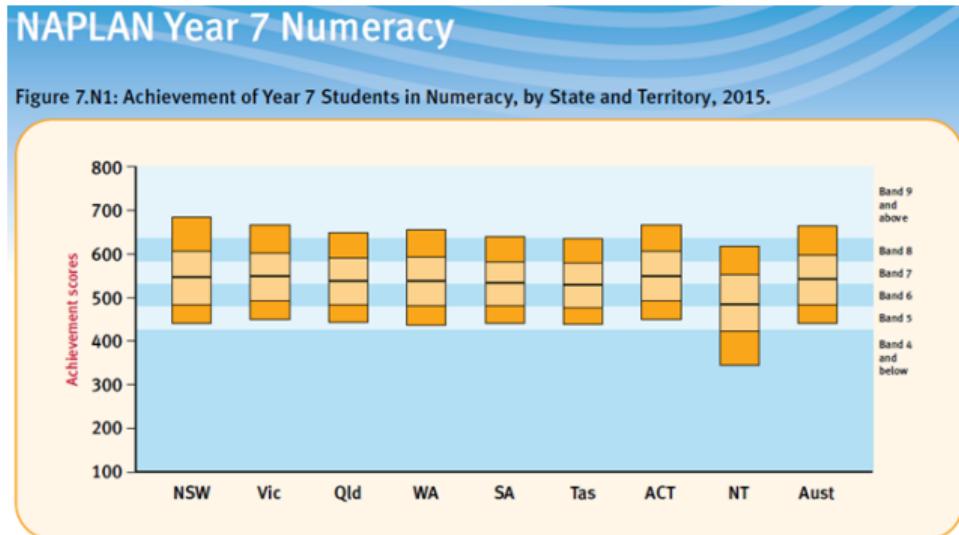
so that the transformed variable Z has

Mean=0 and Standard deviation $s_z = 1$

When two sets of scores are transformed in this manner, each with their own mean and standard deviation we can **compare the standardised scores.**

This is especially useful if the data follow an approximately normal distribution.

Activity: Standardised Scores



Leah sat the Year 7 NAPLAN test in Qld and Kate sat it in ACT and Diana sat it in NT.
Each achieved a mark of 590.

Relative to the state they live in, who performed better in Numeracy?

Activity: Standardised Scores

	NSW	Vic	Qld	WA	SA	Tas	ACT	NT	Aust
Mean scale score / (S.D.)	546.7 (74.4)	548.4 (66.1)	538.9 (62.9)	538.3 (67.3)	532.7 (60.7)	528.8 (60.3)	549.4 (65.7)	484.7 (81.2)	542.5 (68.6)

Leah (Qld)

$$\begin{aligned} z_L &= \frac{x_L - \bar{x}_Q}{s_Q} \\ &= \frac{590 - 538.9}{62.9} \\ &= 0.8124. \end{aligned}$$

(2)

Kate (ACT)

$$z_k = \frac{590 - 549.4}{65.7} = 0.6180$$

(3)

Diana (NT)

$$z_d = \frac{590 - 484.7}{81.2} = 1.2968.$$

(1)

Overview: Specific transformations

Linear transformations

- Add and subtract constants
- Multiply and divide by constants
- Z scores to standardise data

Nonlinear transformations

- Square root (cube root, ...)
- Square (cube, ...)
- Logarithm
- Reciprocal
- Logit
- Helpful when statistical tests require assumptions about population from which data has been drawn such as Normal distribution.

Topic: Exploratory Data Analysis (EDA)

Presentation of Univariate Data

Part A: In Tables

School of Mathematics and Applied Statistics



UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Where in the Statistical Process

- Ethics
- Nature of the question to be answered
- Context/Expertise
- Design:
 - Experiment vs. observational study
 - Sampling
 - Measurement
- **Description and analysis**
 - Reporting
- Conclusions and decision making

VARIATION



Presentation of Data

The aim is to turn data into meaningful information covering all major aspects of the data, with precision AND to communicate that information (paragraphs).

Data can be presented

- as a list as *raw data*
 - in a table in *tabular form*
 - in a graph *graphically*

Appropriate presentation is important.

Presentation of Data in a Table

A frequency table:

- is a useful way to present categorical or discrete data
- lists all possible values along with number of observations (frequency or count) for each value
- Relative frequency = frequency / total is often included (possibly as %)
- Cumulative Frequency

Frequency Table: Qualitative Data

Categorical data: Travelling to School

	Tally	Frequency	Relative Frequency	
Bike		9	$9/30=0.30$	30%
Bus		8	$8/30=0.27$	27%
Car		10	$10/30= 0.33$	33%
Walked		3	$3/30 =0.10$	10%
Total		30	1.0	

Frequency Table: Quantitative Discrete Data

Discrete data: Number of Siblings for 65 students:

No. of Siblings	Frequency	Relative Frequency	Cumulative Frequency
0	5	$5/65 = 0.077$	5
1	28	$28/65 = 0.431$	33
2	17	$17/65 = 0.262$	50 $\rightarrow 50/65 = 77\%$
3	9	$9/65 = 0.138$	59
4	3	$3/65 = 0.046$	62
5	3	$3/65 = 0.046$	65 ✓
Total	65	1.000	X

Grouping Quantitative Data

For discrete (with many different values) or continuous data, it is often necessary to group the observations into **classes**.

The larger the chosen class width, the smaller the number of classes.

Eg. Marks in a mathematics test

Marks in test	Frequency	Relative Frequency	Cumulative Frequency
30 up to 39	1	$1/50 = 0.02$	1
40 up to 49	7	$7/50 = 0.14$	8 ←
50 up to 59	10	$10/50 = 0.20$	18
60 up to 69	14	$14/50 = 0.28$	32
70 up to 79	10	$10/50 = 0.20$	42
80 up to 89	6	$6/50 = 0.12$	48
90 up to 100	2	$2/50 = 0.04$	50
Total	50	1.00	

Note: once the data have been grouped, the raw data are no longer visible.

Topic: Exploratory Data Analysis (EDA)

Presentation of Univariate Data

Part B: In Graphs

School of Mathematics and Applied Statistics



UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Presentation of Data in Graphs

Graphs

- represent the data *visually*
- help in understanding the nature or *distribution of the data*
- are used to illustrate *relationships* between variables

Presentation of Data in Graphs

Graphs

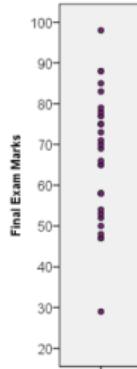
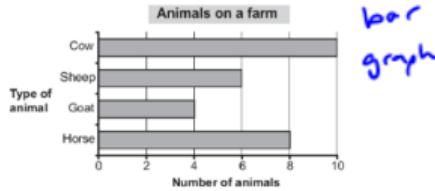
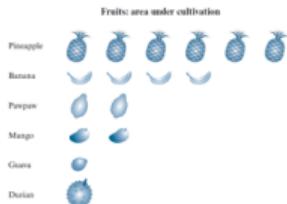
- represent the data *visually*
- help in understanding the nature or *distribution of the data*
- are used to illustrate *relationships* between variables

There are many different types of graphs, some are

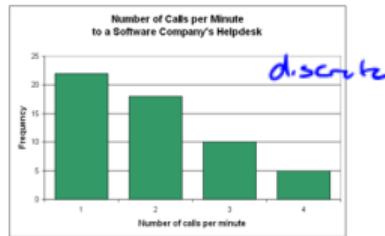
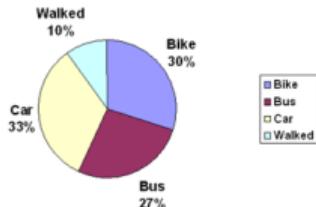
- Pictograms
- Pie graphs
- Bar / column graphs
- Dot plots
- Histograms
- Stem-and-leaf plots

The type of data variable will determine the types of graphs that are suitable.

Appropriate Graphics - Examples



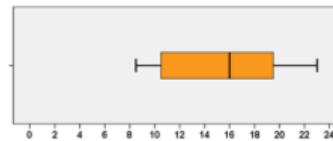
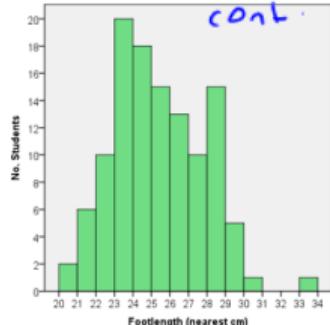
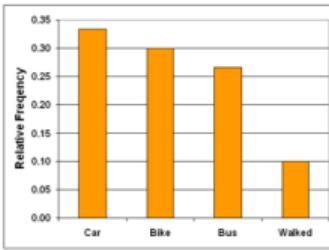
dot plots



Examination result Stem-and-Leaf Plot

Frequency	Stem & Leaf
1.00	2 . 9
.00	3 .
3.00	4 . 778
7.00	5 . 0234888
4.00	6 . 5569
10.00	7 . 0135557789
4.00	8 . 3588
1.00	9 . 8

Stem width: 10
Each leaf: 1 case(s)



boxplot

Different plots for different data types . . .

For one qualitative variable:

- pictograms
- pie charts
- bar graphs

Different plots for different data types . . .

For **one qualitative** variable:

- pictograms
- pie charts
- bar graphs

For **one quantitative** variable:

- dot plots
- bar graphs
(a small number of discrete values)
- histograms
(grouped discrete or continuous data)
- stem-and-leaf plots

Different plots for different data types . . .

For one qualitative variable:

- pictograms
- pie charts
- bar graphs

For two qualitative variables:

- stacked bar graphs
- clustered bar graphs

For one quantitative variable:

- dot plots
- bar graphs
(a small number of discrete values)
- histograms
(grouped discrete or continuous data)
- stem-and-leaf plots

Different plots for different data types . . .

For one qualitative variable:

- pictograms
- pie charts
- bar graphs

For one quantitative variable:

- dot plots
- bar graphs
(a small number of discrete values)
- histograms
(grouped discrete or continuous data)
- stem-and-leaf plots

For two qualitative variables:

- stacked bar graphs
- clustered bar graphs

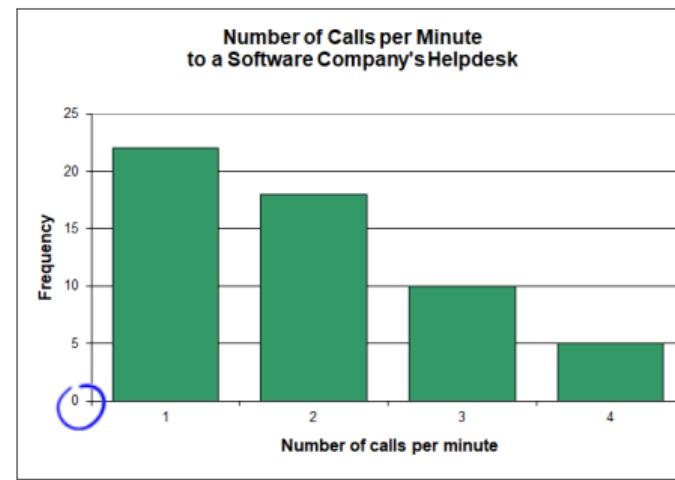
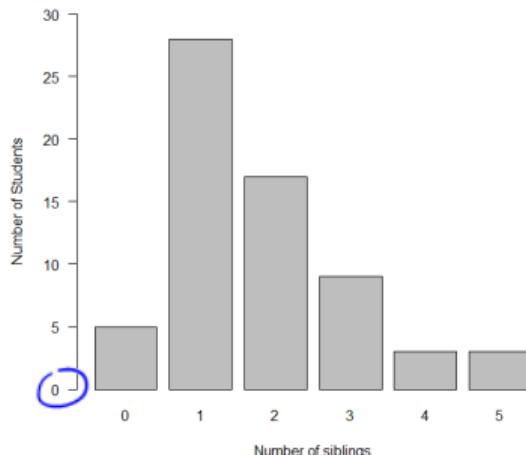
For two quantitative variable/s:

- scatterplots
- line plots (against time)

Bar chart structure

In a bar chart

- the bars are separated - they **do not touch**
- the width of the bars should be the **same** for each category
- the **height (or length)** of each bar represents a quantity, whereas its width means nothing
- the frequency scale **MUST** start at zero

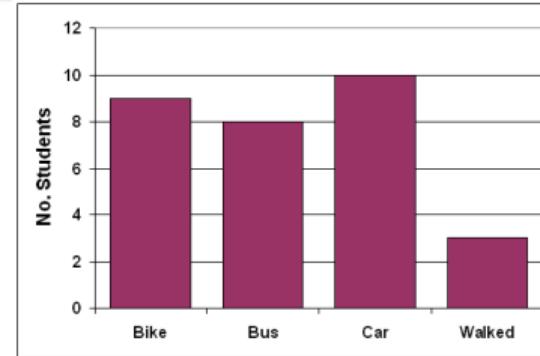


Bar Charts: Examples

Example: Bar Chart for qualitative data: order of bars can be rearranged

Vertical scale

-can show frequencies



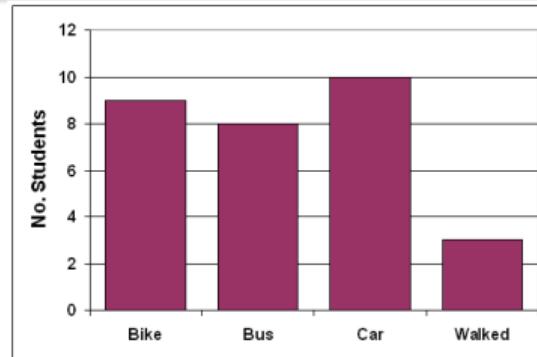
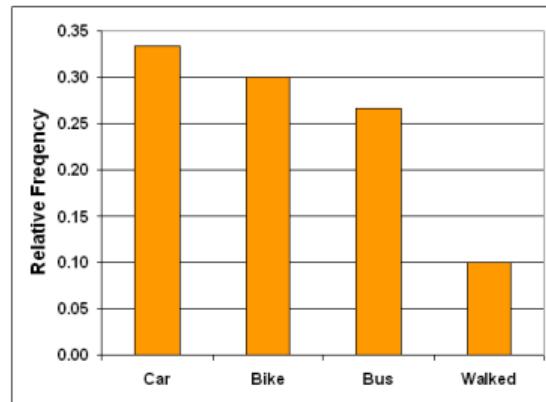
Bar Charts: Examples

Example: Bar Chart for qualitative data: order of bars can be rearranged

Vertical scale

-can show frequencies

- or relative frequencies



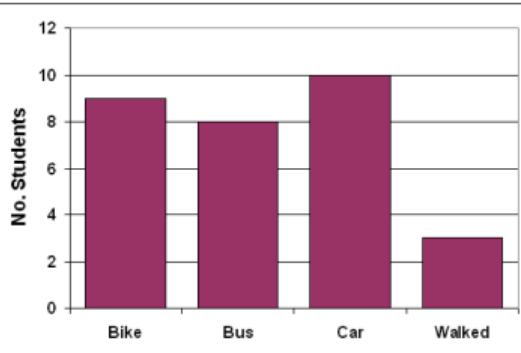
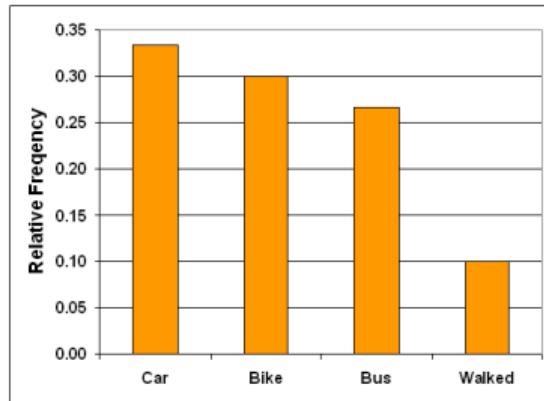
Bar Charts: Examples

Example: Bar Chart for qualitative data: order of bars can be rearranged

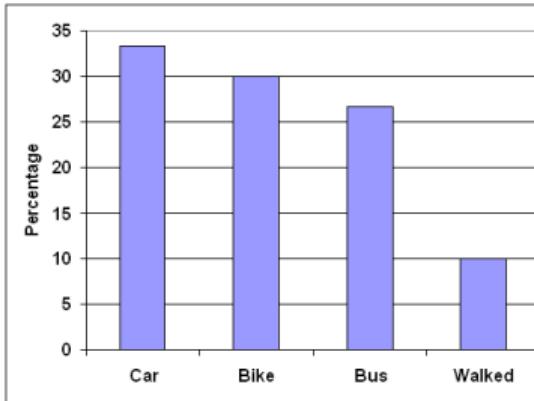
Vertical scale

-can show frequencies

- or relative frequencies



- or percentages



In R: Bar Charts

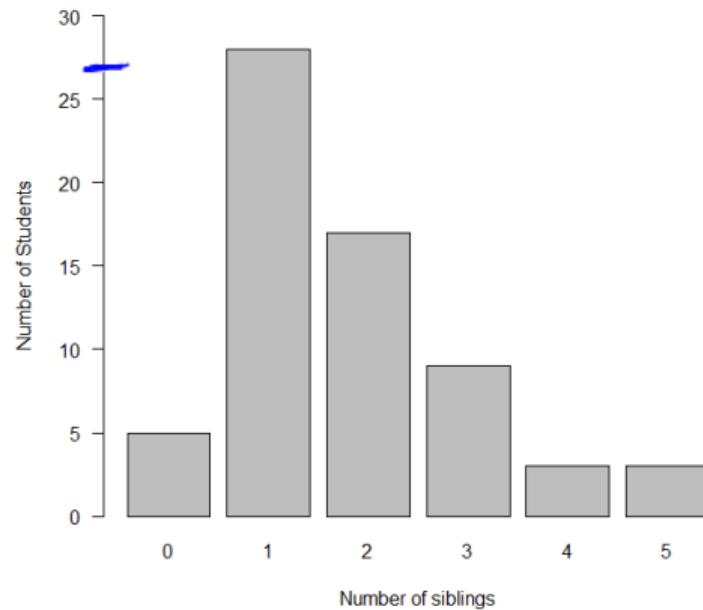
R code:

```
Siblings <- c(M100data$Siblings)  
Siblingfreq <- table(Siblings)  
Siblingfreq
```

* Siblings

0	1	2	3	4	5
5	28	17	9	3	3

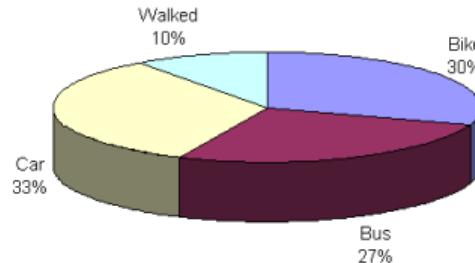
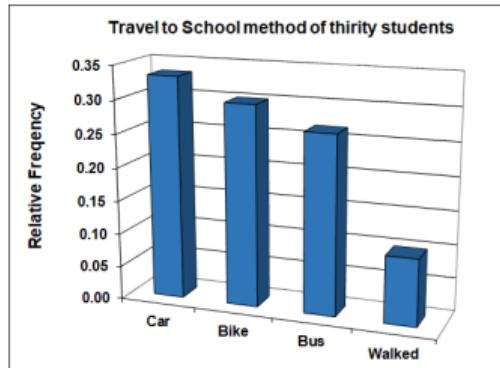
```
barplot(Siblingfreq,  
xlab = "Number of siblings",  
ylab="Number of Students",  
las=1, ylim =c(0, 30))
```



Inappropriate Graphics - Examples

Use the fewest dimensions possible:

- using 3D is volume which can distract
- don't use unless necessary - i.e. not for univariate data



- avoid pie charts
- a simple bar chart is often more effective

Histograms- Quantitative Data

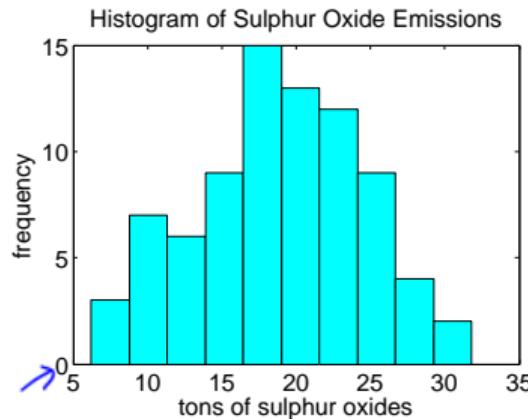
Histograms are used to represent

- **continuous:** interval or ratio data
- **discrete:** grouped data
(too many unique values for a bar chart)

Histograms- Quantitative Data

Histograms are used to represent

- **continuous:** interval or ratio data
- **discrete:** grouped data
(too many unique values for a bar chart)



- Real number scale on horizontal axis, **no gaps** between bars (**bins**).
- Observations are **grouped** into bins (classes), not necessarily of constant width.
- Vertical scale must **start at zero**
- Frequency (count) or rel. freq. is represented by **area** of bar.

Area of histogram bars

- Area = height \times width, so vertical axis of histogram should ideally display **density** (relative frequency \div width).
- For **constant bin width**, area is proportional to height, so vertical axis can display frequency if preferred.
- For **non-constant bin width**, vertical axis **must** display **density**.

Histogram with constant bin width

Example: Sulphur emission data

Histograms: describing shape

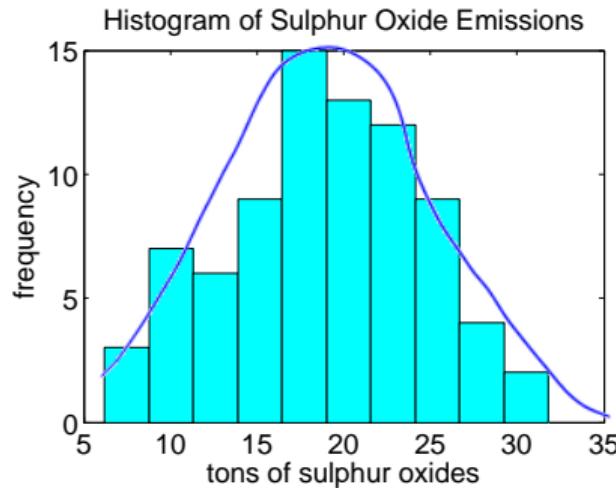
- Reasonably symmetric
- Unimodal - single hump

Histogram with constant bin width

Example: Sulphur emission data

Histograms: describing shape

- Reasonably symmetric
- Unimodal - single hump



Histogram with Non-constant bin width

If 6 out of 80 observations satisfy $6 \leq x < 10$, then for first bar

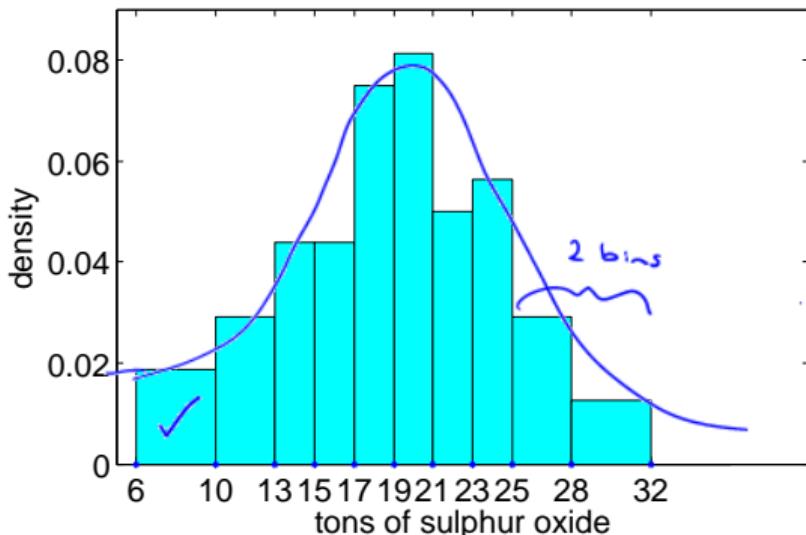
$$\text{density} = \frac{\frac{\text{rel. freq.}}{\text{width}}}{\frac{6/80}{(10-6)}} = \frac{6/80}{(10-6)} = \frac{6^3}{80} \times \frac{1}{4_2}$$

$$= \frac{3}{160}.$$

$$= \underline{0.01875}$$

Total area of bars = 1

Histogram of sulphur emissions



Purpose of histogram

To display **overall shape and interesting features**, including

- Outliers?
- Long or short tails?
- Symmetry or skewness?
- Bell shape, U shape, uniform, ... ?
- Unimodal/bimodal (1 or 2 humps)?

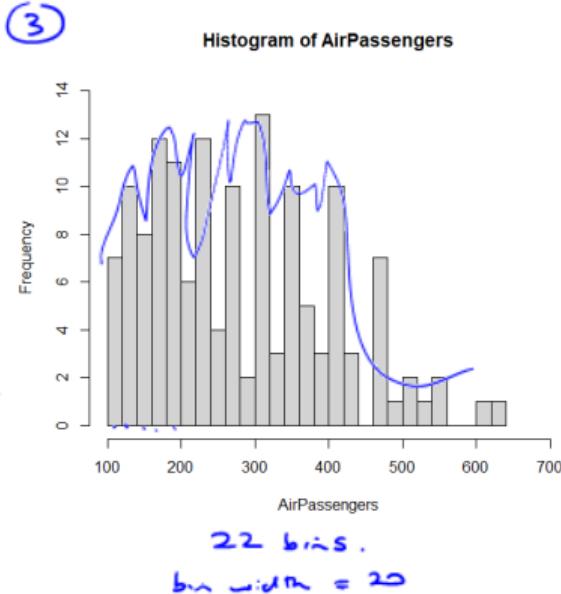
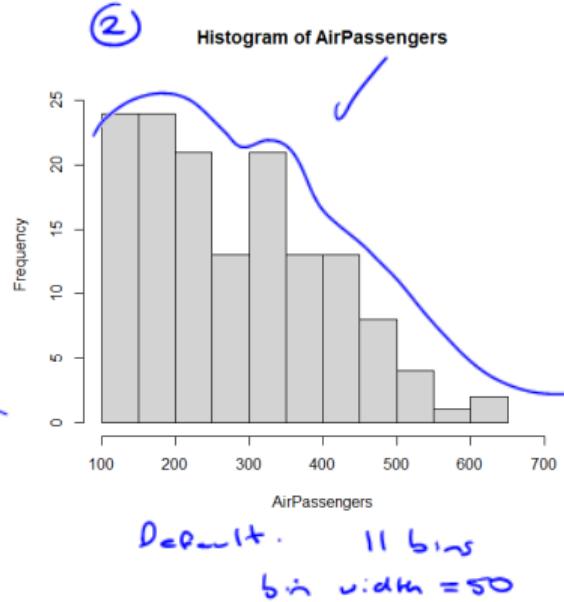
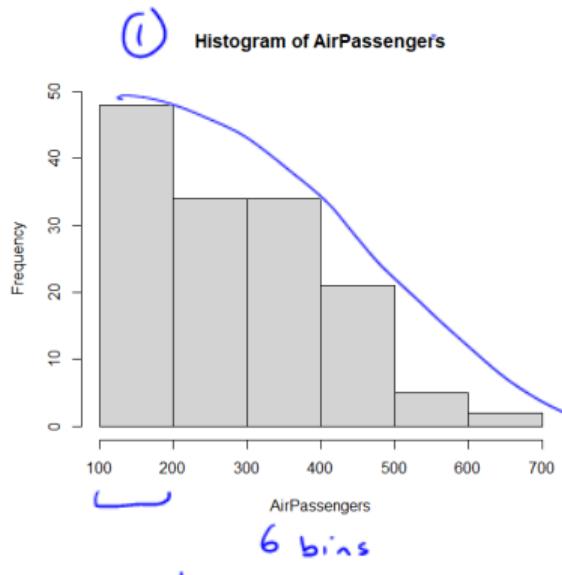
Appearance varies according to number and choice of bins;

- Rule of thumb: use number of bins as $\approx \sqrt{n}$
- avoid too few (uninformative)
- or too many (bumpy plot)

↑

Histogram: Choice of interval length is important

Number of monthly air passengers (1949-1960): same data plotted left to right with 6, 11, and 22 bins respectively.



In R: Histogram

R code:

#1. Default uses 11 bins (see middle plot)

```
hist(AirPassengers, xlim=c(100, 700), ylim=c(0, 26))
```

#2. Use 6 bins (see left plot)

```
hist(AirPassengers, breaks=6, xlim=c(100, 700), ylim=c(0, 50))
```

#3. Use 22 bins (see right plot)

```
hist(AirPassengers, breaks=22, xlim=c(100, 700), ylim=c(0, 14))
```

Topic: Exploratory Data Analysis (EDA)

Shape & Reporting of Univariate Data

School of Mathematics and Applied Statistics



UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Meaningful Reporting - Univariate Data

① Context

- Units, sampling, design of collection, research findings, daily recommended exposure

② Shape

- Bell, normal, uniform, symmetric, unimodal, skewed, bimodal

③ Outliers/Extremes

- different software packages may define differently

④ Centre

- Mean, median, mode, trimmed mean

⑤ Spread

- Range, IQR, Variance, Standard Deviation

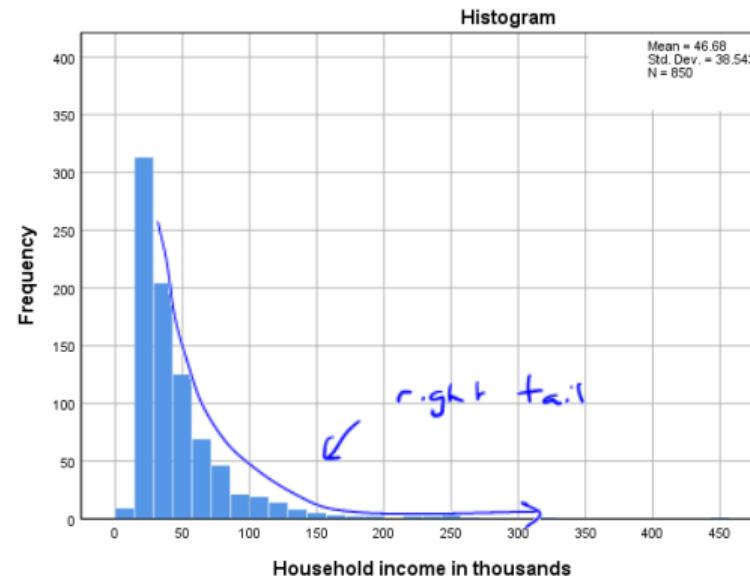
⑥ Patterns

- are there any?

Tails of Distribution

- The **left-hand tail** is the region of **lowest** data values.
- The **right-hand tail** is the region of **highest** data values (don't confuse with highest frequency).

E.g. income distributions typically have a **long** right-hand tail; a minority have much higher incomes than the majority.

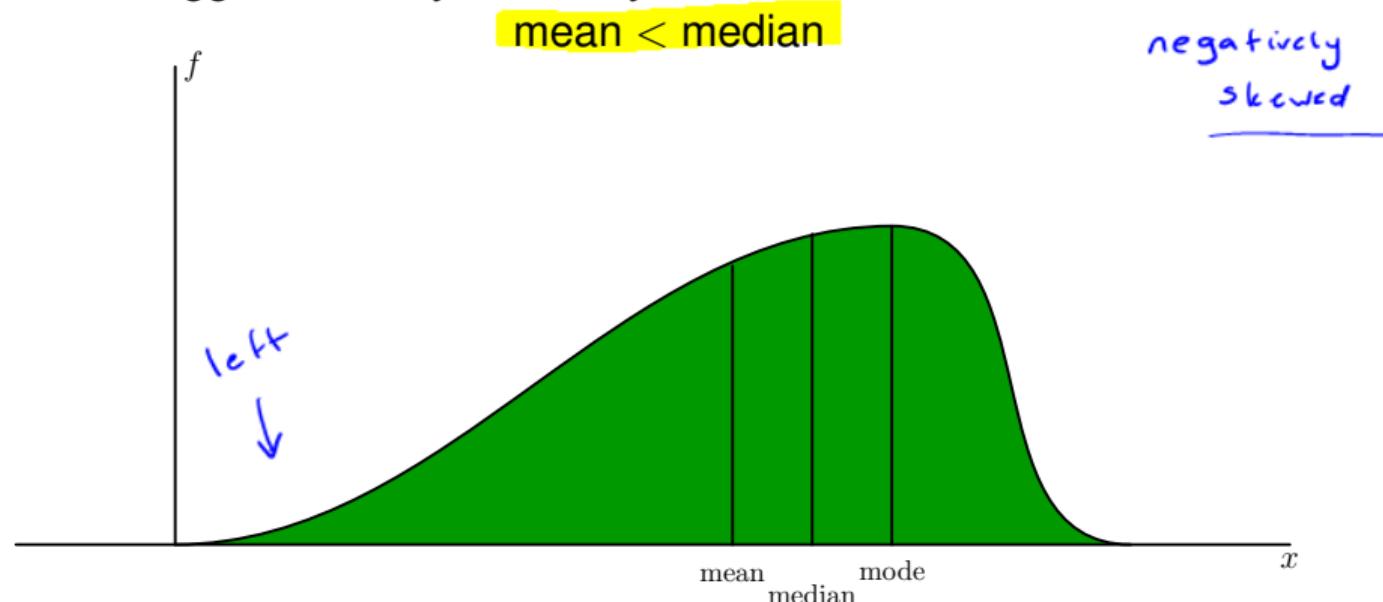


Skewness

The direction of the **skew** is determined by the **location of the tail**

If the tail is on left then the distribution is **skewed to the left**

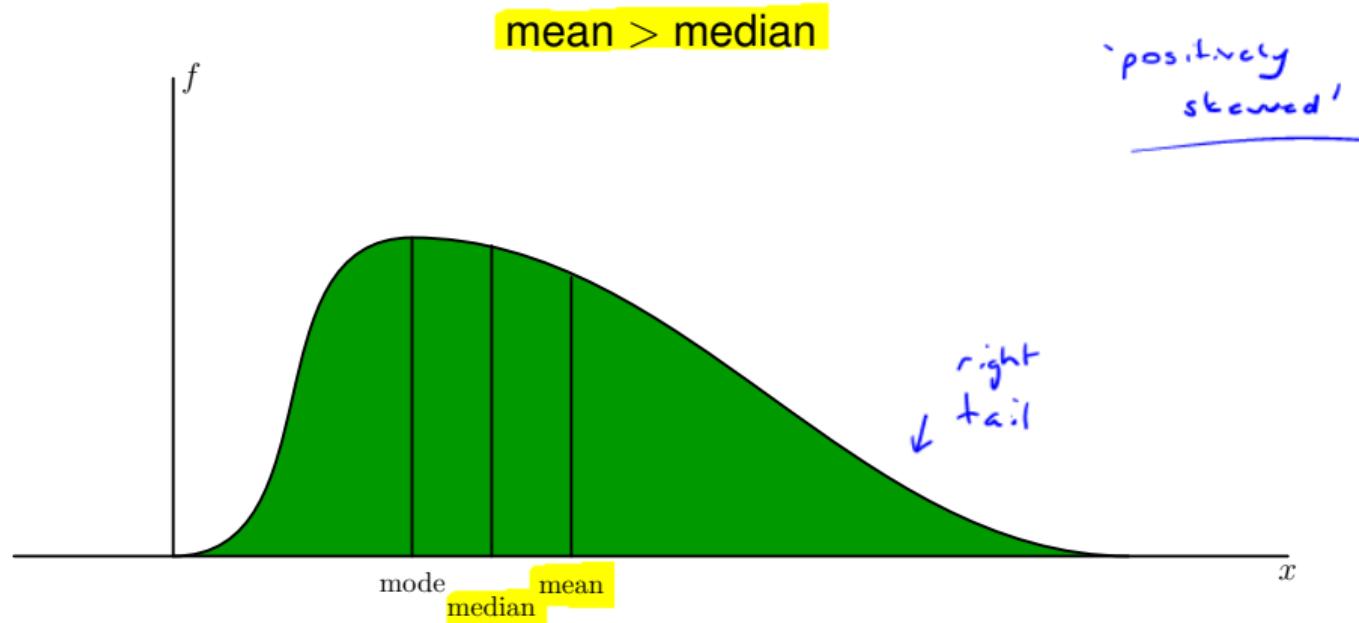
The mean is dragged down by unusually small values in the left tail,



Skewed Distribution

If the tail is on right then the distribution is skewed to the right

The mean is inflated due to unusually large data values in the right tail



Meaningful Reporting - Univariate Data

① Context

- Units, sampling, design of collection, research findings, daily recommended exposure

② Shape

- Bell, normal, uniform, symmetric, unimodal, skewed, bimodal

③ Outliers/Extremes

-defined differently with different plots/data

④ Centre

- Mean, median, mode, trimmed mean

⑤ Spread

- Range, IQR, Variance, Standard Deviation

⑥ Patterns

Meaningful Paragraph: Describing Data

Univariate analysis - one variable at a time.

The aim is to turn data into **meaningful information** covering **all major aspects** of the data, **with precision** AND to **communicate** that information (paragraphs).

- Example:
 - Each row represents one item of food ⇒ can of soup
 - Two columns of data ⇒ 2 variables: fat and sodium
 - both of type quantitative and ratio.

Fat	Sodium
0.5	120
9.0	20
3.0	140
1.0	65
0.5	110
2.0	300
2.0	160
0.0	150
6.0	240
3.0	320
0.5	210
0.5	220
1.5	200
2.0	280
3.5	210
1.0	190
1.0	270
0.5	230
0.0	300
0.0	300
0.0	120
6.0	170
0.0	170
0.0	210
1.0	140
1.0	210
1.0	170
1.0	150
1.5	210

Meaningful Paragraph: Context

Here is **nutrition** information (fat) taken from cans of soup.

To make sense of this data we need to know:

- How is fat **measured?** What **units** are used? (g)
- How were they sampled? eg. Brands? Flavours? Shelves in supermarket/s?
- Sample information? Sample size? $n = 76$
- Previous study findings on similar products:- general context
 - what is the maximum or minimum recommended daily intake for adults? for children?
 - mean?
 - or spread?
 - or...?
 - groupings?

Fat
0.5
9.0
3.0
1.0
0.5
2.0
2.0
0.0
6.0
—

Meaningful Paragraph: Shape

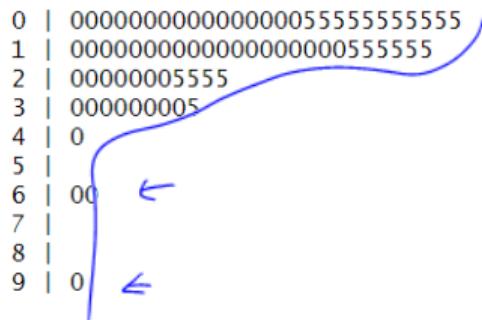
- Quantitative – look at different plots
 - stem & leaf
 - histogram \Rightarrow grouped discrete or continuous
 - bar chart \Rightarrow a small no. of discrete values
 - boxplot \Rightarrow 5-number summary
 - dot plot
- Different versions of the same plot
 - Drawing by hand with different stems
 - Different scales or bin widths
- Why?
 - Helps determine analysis – which
 - measure of centre
 - and spread to use
 - Helps to find unusual values //

Meaningful Paragraph: Shape

Quantitative – look at 3 different plots of same data

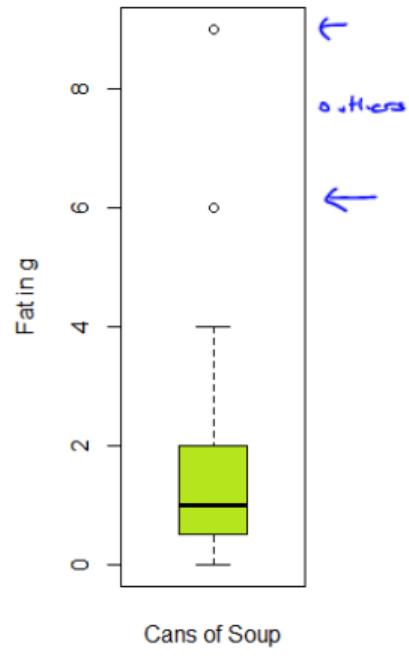
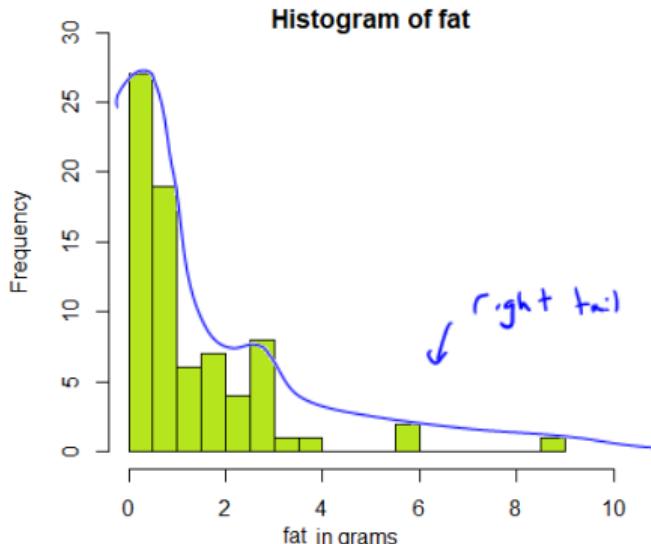
```
> stem(fat)
```

The decimal point is at the |



Description

- longer tail of high values \implies high fat content *for a small no. cans.*
- skewed distribution
 - Positively skewed or skewed to the right



Meaningful Paragraph: Shape

```
> stem(fat)
```

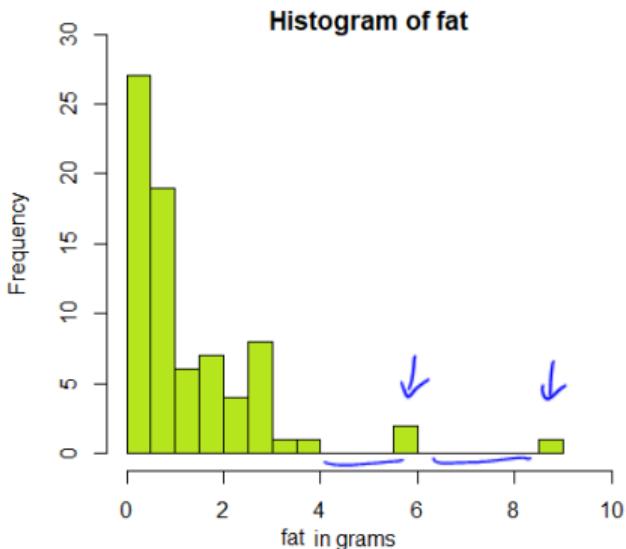
The decimal point is at the |

```
0 | 0000000000000000055555555555  
1 | 0000000000000000000000555555  
2 | 00000005555  
3 | 000000005  
4 | 0  
5 |  
6 | 00  
7 |  
8 |  
9 | 0
```

Stem & leaf

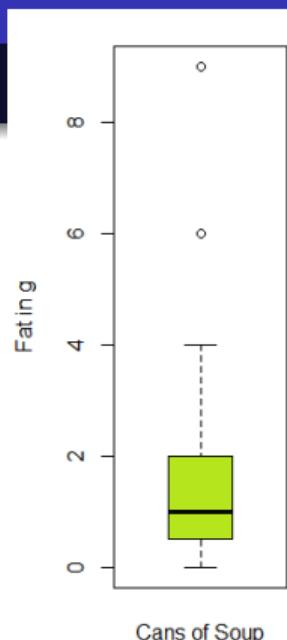
See Detail:

- Patterns
- Centre
- Spread
- Extremes not so easily



Histogram

- Loss of detail
- See outliers
- See gaps

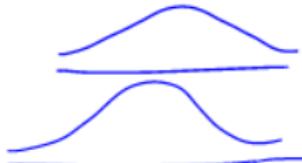


Boxplot: summary plot

- See outliers – not all packages do so
- Symmetry, skewness
- Poorer shape and detail

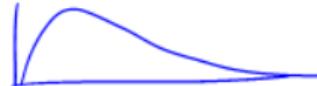
Describing Shape

- Bell-shaped



- Normal

- Skewed to the right (or positively skewed)
⇒ Longer tail of high values



- Skewed to the left (negative) ⇒ Longer tail of low values

- Uniform



- Unimodal versus Bimodal



- Exponential



- Symmetric - two halves same about centre



Activity: Sketch an example for each

Different versions of one type of plot

Take care when describing shape

- Different no. of stems in a stem & leaf plot
- or different no. of bins in histogram
- Whether or not the plot shows outliers

Spot the differences:

fat Stem-and-Leaf Plot

SPSS

Frequency Stem & Leaf

16.00	0 . 0000000000000000	0 - 4
11.00	0 . 5555555555	5 - 9
19.00	1 . 0000000000000000	
6.00	1 . 55555	
7.00	2 . 000000	
4.00	2 . 555	
8.00	3 . 0000000	
1.00	3 . 5	
1.00	4 . 0	
3.00 Extremes	(≥ 6.0)	←

Stem width: 1.00 ✓

Each leaf: 1 case(s)

> stem(fat)

R.

The decimal point is at the |

0 0000000000000000555555555555
1 0000000000000000055555555555
2 00000005555
3 0000000005
4 0
5
6 00 ✓
7
8
9 0 ✓

Meaningful Paragraph: Centre

- The **mean** uses all information in the sample because each value is added to the sum
- **Mean** is subject to error if spurious values are entered
- **Median** is less affected by “wild” values i.e. it is robust.

If the median is similar to the mean:

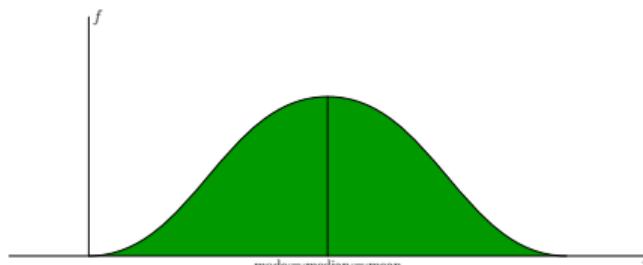
- Use the mean as it uses all data
- It is easier to work with means

If they are different because of non-symmetric distributions

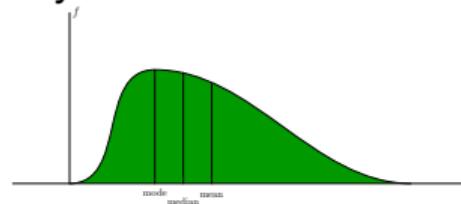
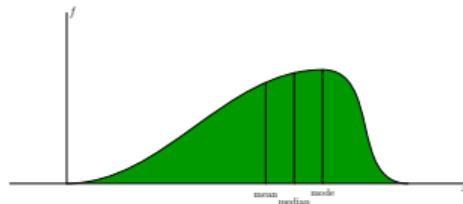
- Can be useful to report both
- The context of what the data are used for may also determine which is the appropriate statistic

Centre: which measure to use?

- Symmetrical mean & median similar



- Skewed - use mean & median as they will differ



- When outliers present - use a robust measure, so outliers do not influence eg. median
- Use mode for nominal data

Meaningful Paragraph: Spread or Variability

Think: Which measure of spread is more appropriate for the variable Fat content?

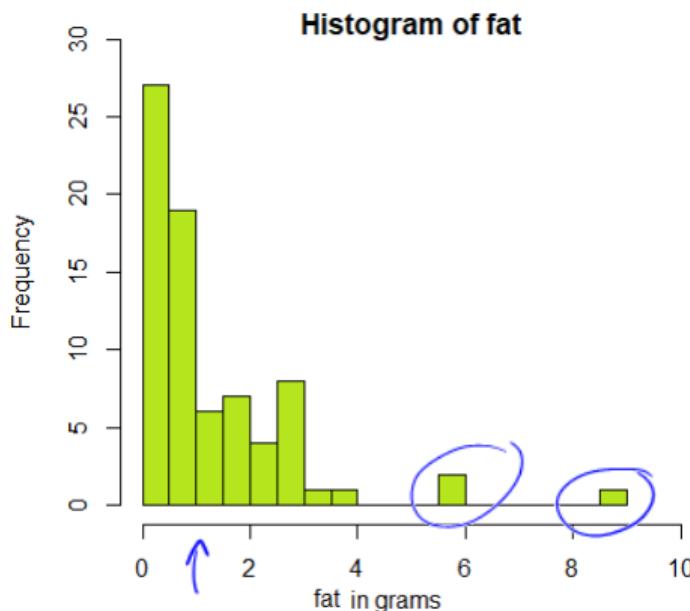
```
> stem(fat)
```

The decimal point is at the |

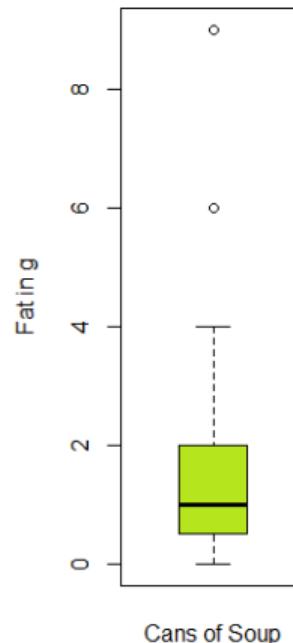
```
0 | 0000000000000000055555555555
1 | 000000000000000005555555
2 | 00000005555
3 | 000000005
4 | 0
5 |
6 | 00
7 |
8 |
9 | 0
```

$$\text{mean} = 1.447 \text{ g}$$

$$\text{median} = 1.0 \text{ g}$$



IQR ✓



Example: which measure of variability to use?

Think: which measure of variability might be appropriate for the variable Monthly Average Temperature?

$$n=114.$$

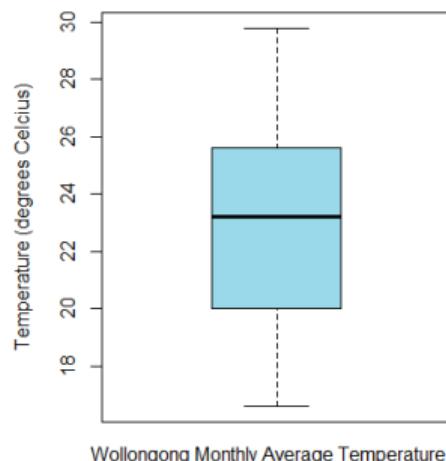
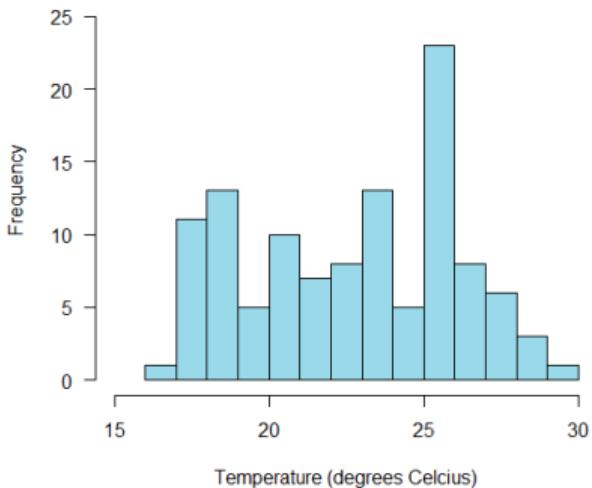
The decimal point is at the |

16	6
17	11333357789
18	122333556789
19	0348
20	00244445789
21	012499
22	0033444689
23	1224567777889
24	1679
25	0123334455666677889999
26	001256789
27	014468
28	0113
29	8

$$\text{mean} = 22.77^\circ \text{C}$$

$$\text{median} = 23.2^\circ \text{C}$$

Wollongong Monthly Average Temperatures



Wollongong Monthly Average Temperature

Meaningful Paragraph: Patterns

Why Look?

- There may be something unusual detected about measurement
- Eg 1. Blood pressure taken in different countries may have used different instruments
- Eg 2. those measuring may be more or less prone to rounding numbers, for example at border between grades in exams
- There may be some importance attached to the pattern
eg. ECG heart attack

Meaningful Paragraph: Patterns

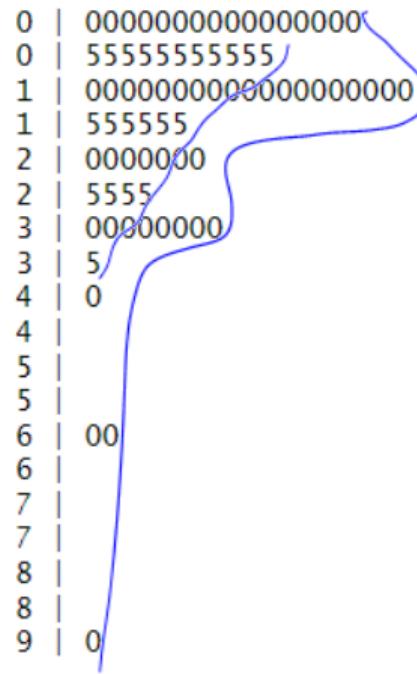
Patterns are not often seen

Fat variable:

- Within this data set the measurement is rounded to the whole or half gram

```
> stem(fat, scale = 2)
```

The decimal point is at the |



Summary

When reporting, need to consider

- type of variable
- different types of plots - reveal different characteristics of data
- shape of distribution: centre & spread
- whether there are outliers: centre & spread
- patterns

Topic: Exploratory Data Analysis (EDA)

Presentation of Bivariate Data

Part A: Two categorical variables

School of Mathematics and Applied Statistics



UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Bivariate data: Two Variables

Different tables / plots for different data types . . .

For **two qualitative** variables:

- two-way ^{tables}
- stacked bar graphs
- clustered bar graphs

For **one quantitative** and **one qualitative** variable:

- side-by-side box plots
- back-to-back stem & leaf plots

For **two quantitative** variable/s:

- scatterplots
- line plots (against time)

Contingency Table

A two-way table or **contingency table** summarises bivariate data of two categorical variables.

Example: Titanic data

Survived	1 st Class	2 nd Class	3 rd Class	Crew	Total
No	122	167	528	673 /	1490
Yes	203	118	178	212	711
Total	325	285	706	885	2201

Is there any association between the two variables?

Is the proportion of *Survived* the same for *Class of passenger*?

Contingency Table - Conditional probability

Example: Titanic: Observed data

Survived	1 st Class	2 nd Class	3 rd Class	Crew	Total
No	122	167	528	673	1490
Yes	203	118	178	212	711
Total	325	285	706	885	2201

$$P(\text{1st} | \text{No}) = \frac{122}{1490} = 0.082 \\ = 8.2\%$$

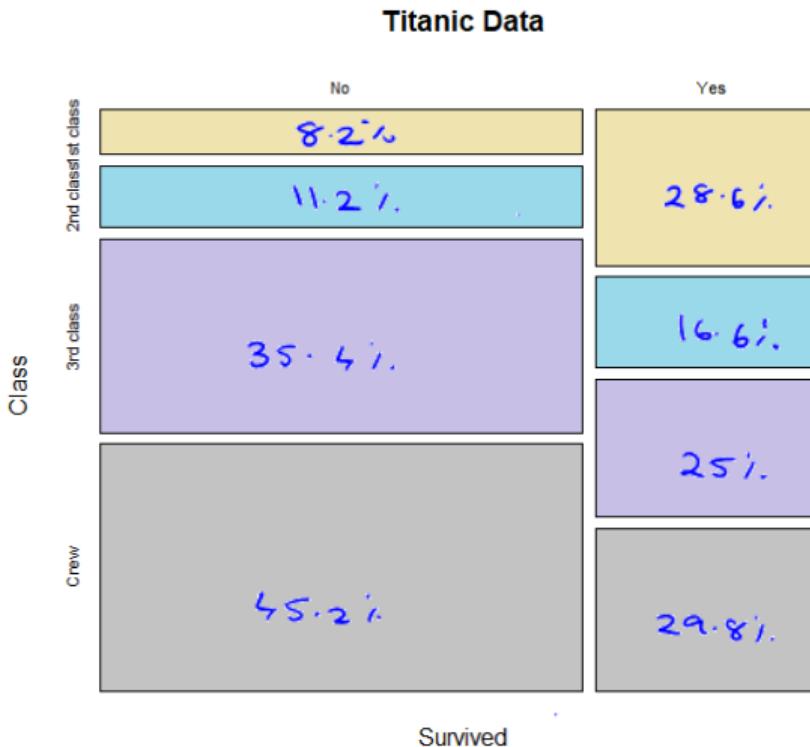
Exercise: Determine the row percentages:

Survived	1 st Class	2 nd Class	3 rd Class	Crew	Total
No	122/1490=	167/1490=	528/1490 =	673/1490=	100%
	8.2%	11.2%	35.4%	45.2%	
Yes	203/711 =	118/711=	178/711=	212/711 =	100%
	28.6%	16.6%	25.0%	29.8%	



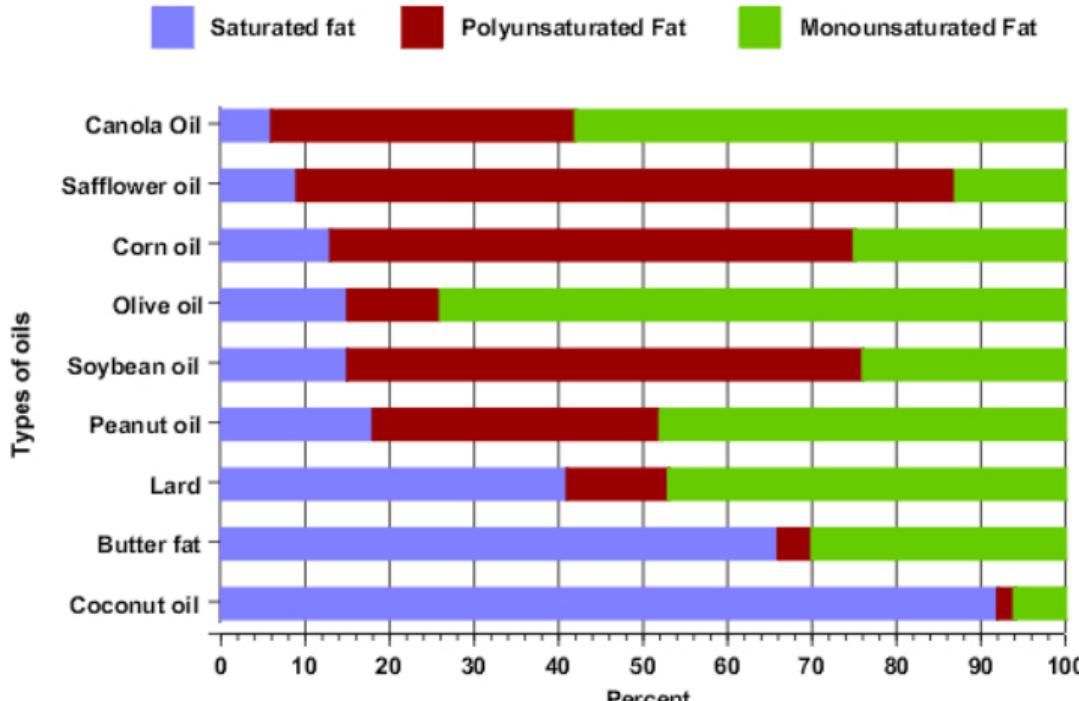
Mosaic Plot

This **mosaic plot** displays counts of two-way table, as areas proportional to frequencies within a row.



Example of a stacked bar chart

Stacked bar graphs are often used to represent parts of a whole.



9 different types
cooking oil.

lowest % of sat. fat

≈92% Coc. oil
Sat. fat.

Conditional probability - another look

Example: Phone carriers and Gender

	Optus	Telstra	Vodafone	Total
Female	19	9	4	32
Male	36	17	20	73
Total	55	26	24	105

Q1: Given a female, what is the probability that they

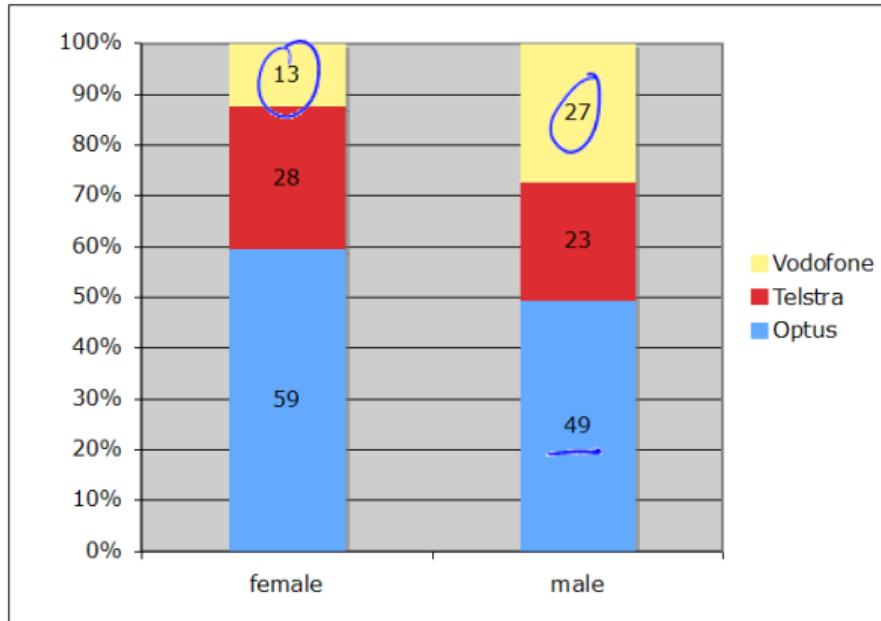
- Use Optus? $P(O|F) = 19/32 = 0.594$
- Use Telstra? $P(T|F) = 9/32 = 0.281$
- Use Vodafone? $P(V|F) = 4/32 = 0.125$

Q2: Write down

- $P(\text{Optus}|\text{Male}) = 36/73 = 0.493$
- $P(\text{Telstra}|\text{Male}) = 17/73 = 0.233$
- $P(\text{Vodafone}|\text{Male}) = 20/73 = 0.274$

Q3: Is the pattern of usage the same for males and females?

Stacked Bar Charts & Independence



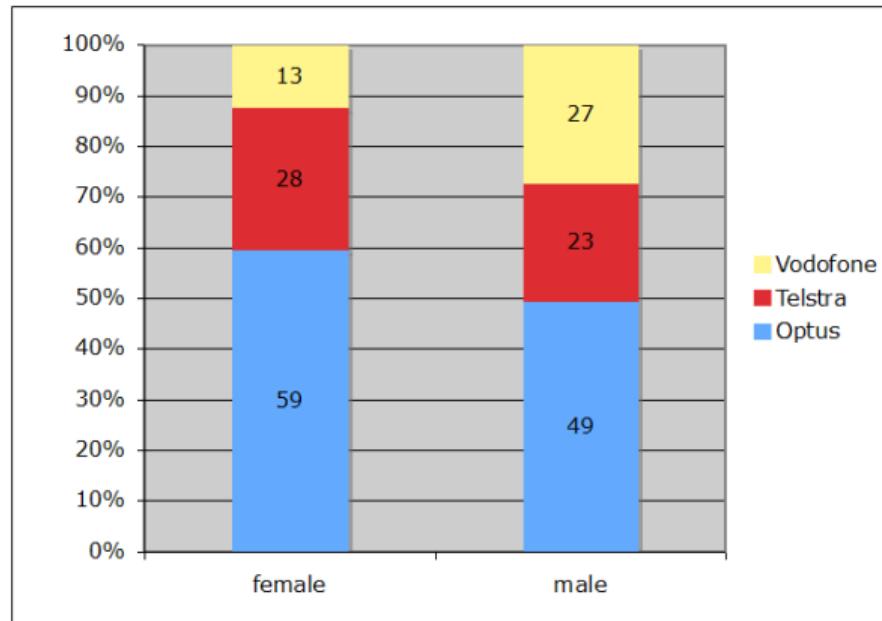
The proportion of Males using Vodafone is greater (27%) than that for females. (13%)

For Optus, it is less than 59% than that for females.. 59%.

Is this just the sample or is it a pattern evident in the population?

Stacked Bar Charts & Independence

When the pattern of use by males and females is the same then we have independence in the population



When we use a sample to infer something about the population then similar patterns imply independence.

BUT

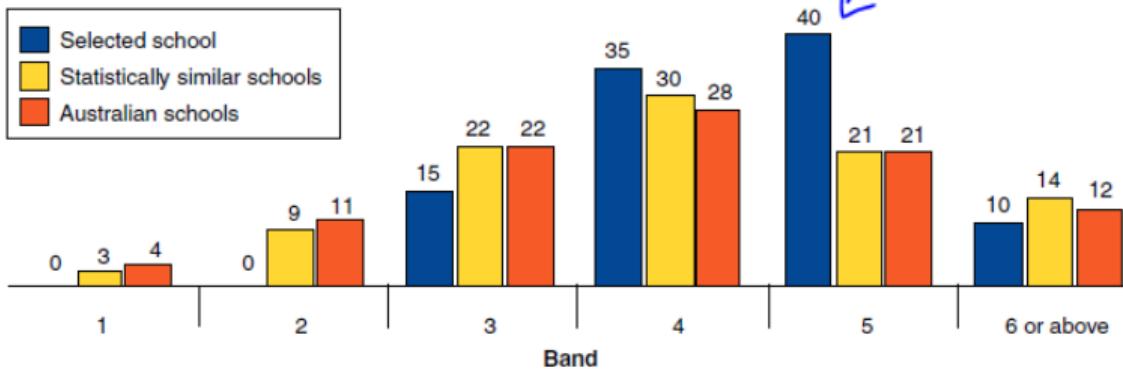
How similar is similar?

Bar Charts can be: . . . clustered

For **two qualitative** variables:

Year 3 Reading

Percentage of students in each band



Source: Lantites - Sample Questions p15, ACER

This graph shows the percentage of Year 3 students in six achievement bands for reading, for a selected school. It also shows comparable percentages for statistically similar schools and for all Australian schools.

Topic: Exploratory Data Analysis (EDA)

Presentation of Bivariate Data

Part B: One quantitative and one qualitative variable

School of Mathematics and Applied Statistics



UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Bivariate Data: Two Variables

Different tables / plots for different data types . . .

For **two qualitative** variables:

- two-way tables
- stacked bar graphs
- clustered bar graphs

Bivariate Data: Two Variables

Different tables / plots for different data types . . .

For **two qualitative** variables:

- two-way tables
- stacked bar graphs
- clustered bar graphs

✓ For **one quantitative** and **one qualitative** variable:

- side-by-side box plots
- back-to-back stem & leaf plots

Bivariate Data: Two Variables

Different tables / plots for different data types . . .

For **two qualitative** variables:

- two-way tables
- stacked bar graphs
- clustered bar graphs

For **one quantitative** and **one qualitative** variable:

- side-by-side box plots
- back-to-back stem & leaf plots

For **two quantitative** variable/s:

- scatterplots
- line plots (against time)

Comparing Batches: One quantitative and one qualitative variable

Question: Is there a **difference** between two or more batches of data?

- One quantitative variable
- Two batches (male & female)
- Or many batches (eg brands)

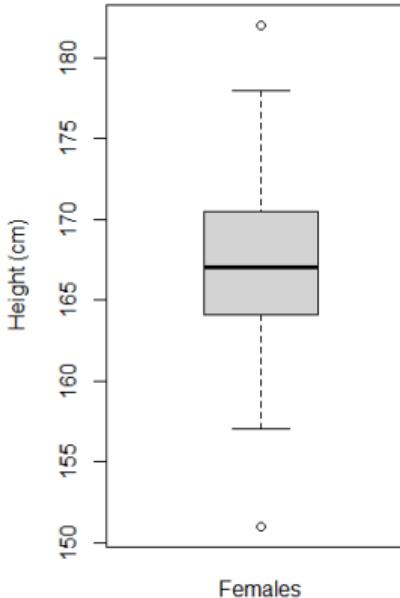
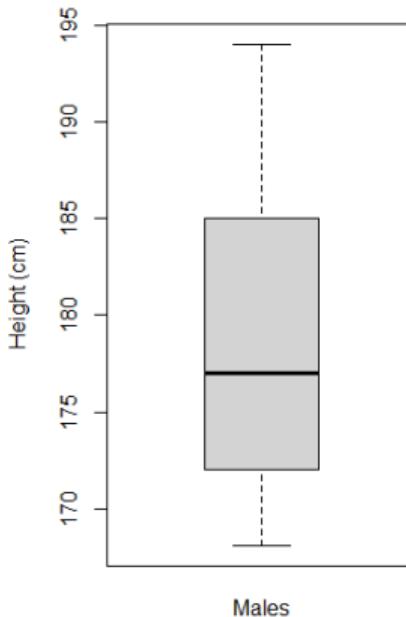
The aim is to turn data into meaningful information AND to **communicate it effectively**

- Plots should be on the same scale
- Do NOT use two separate plots
- Different plots will show different aspects of the data

Later we examine hypothesis tests - eg. are the population means significantly different?

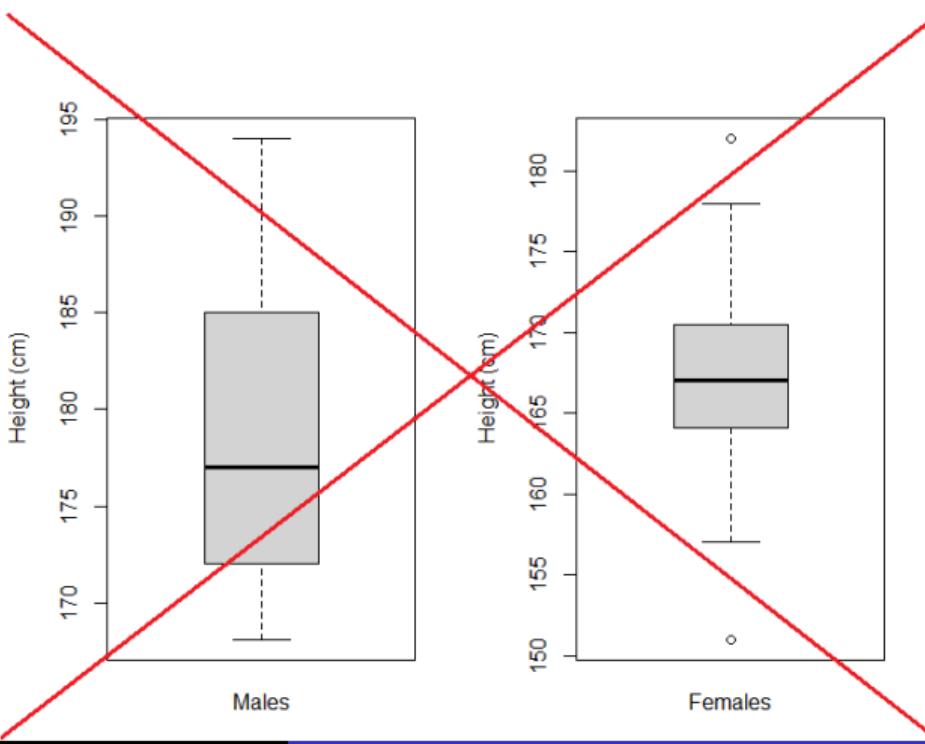
Comparing batches

Example: Measured Heights (cm) for 46 M and 19 F



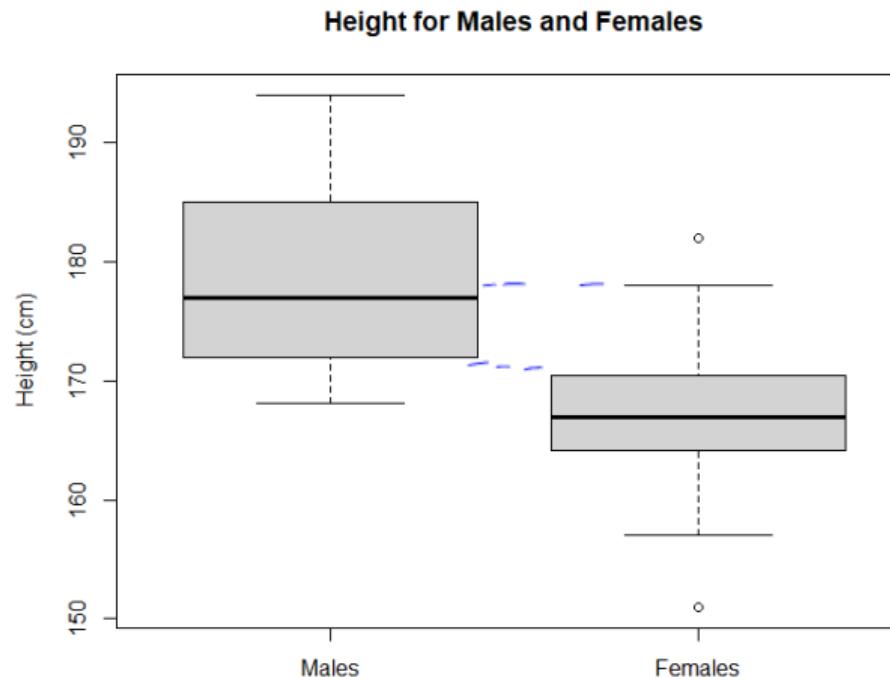
Comparing batches

Example: Measured Heights (cm) for 46 M and 19 F



To compare: Use one plot, one set of axes

Example: Measured Heights (cm) for 46 M and 19 F



Comparing batches: Communication

Key descriptors involve comparison

Based on comparative techniques make **comparative statements**

- Greater than ...
- Similar to ...
- Less than ...

Comparing batches: Communication

Key descriptors involve comparison

Based on comparative techniques make **comparative statements**

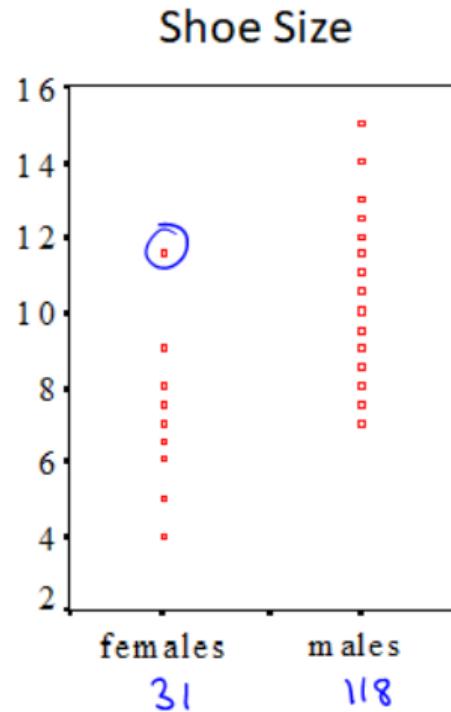
- Greater than ...
- Similar to ...
- Less than ...

For all key features

- Contexts
- Shape of distribution
- Outliers/Extremes
- Centre
- Spread
- Patterns

Comparison - Dot plots

Comparison of male and female shoe size



We can easily see:

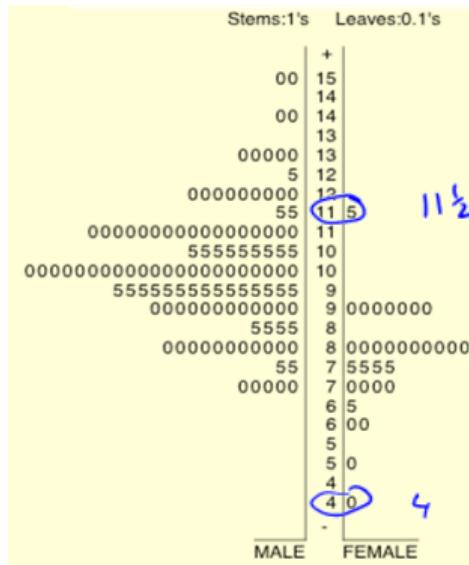
- spread
- possible outliers

What can't we see?

- shape of the distribution
- centre of data
- density of dots (ie how many people with one shoe size) as dots overlaid

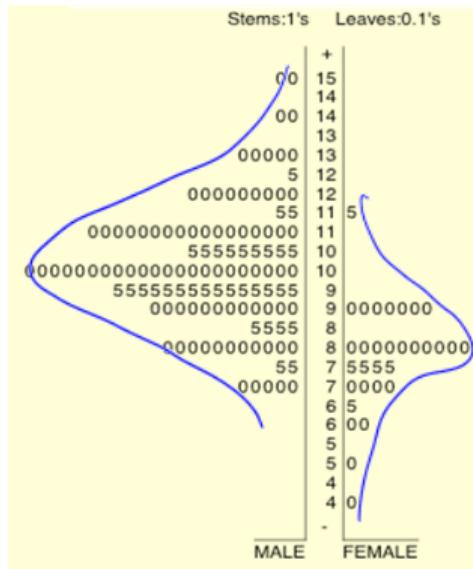
Comparison: Back-to-back stem-and-leaf plots

What does the data reveal?



Comparison: Back-to-back stem-and-leaf plots

What does the data reveal?

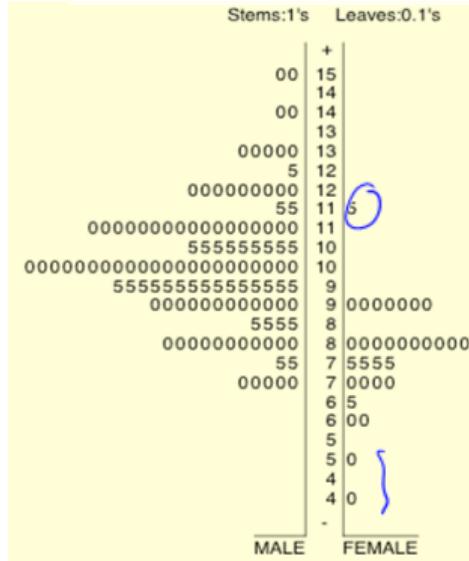


1 Distribution shape:

- Male: bell-shaped
 - Female: skewed with a longer tail of small shoe sizes (negative skew, skewed to left)

Comparison: Back-to-back stem-and-leaf plots

What does the data reveal?



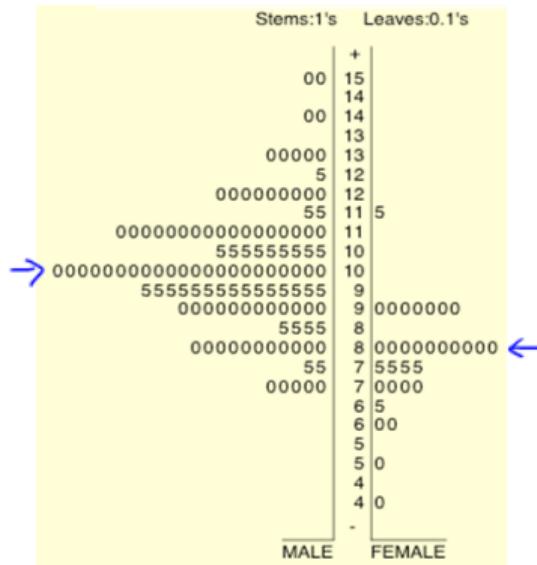
1 Distribution shape:

- Male: bell-shaped
- Female: skewed with a longer tail of small shoe sizes (negative skew, skewed to left)

2 Outliers - Possible - but not shown here

Comparison: Back-to-back stem-and-leaf plots

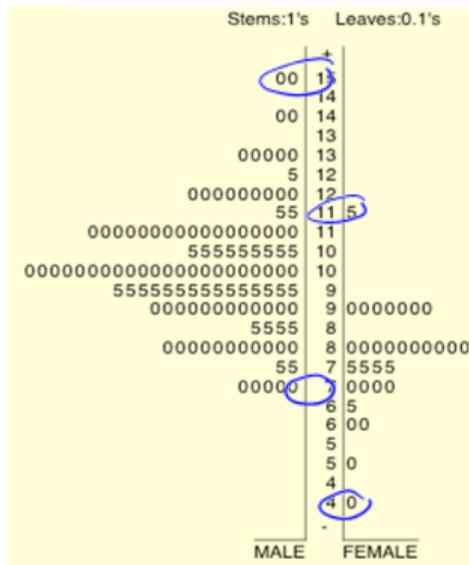
What does the data reveal?



- 1 Distribution shape:
 - Male: bell-shaped
 - Female: skewed with a longer tail of small shoe sizes (negative skew, skewed to left)
 - 2 Outliers - Possible - but not shown here
 - 3 Centre - can only see mode
 - mode for males is 10 &
 - is higher than mode for females (8)

Comparison: Back-to-back stem-and-leaf plots

What does the data reveal?



1 Distribution shape:

- Male: bell-shaped
- Female: skewed with a longer tail of small shoe sizes (negative skew, skewed to left)

2 Outliers - Possible - but not shown here

3 Centre - can only see mode

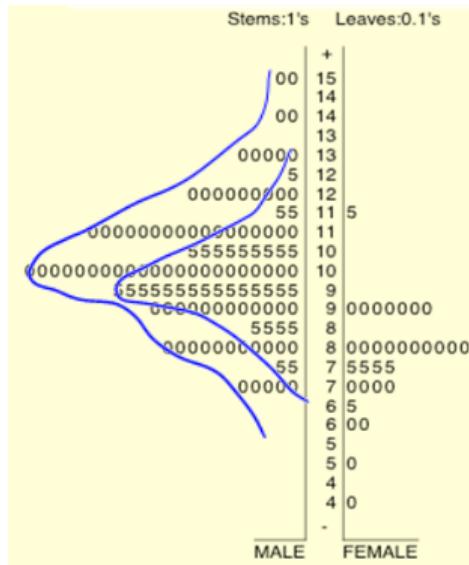
- mode for males is 10 &
- is higher than mode for females (8)

4 Spread can determine range

- for males is 7-15 and females 4-11.5
- so range is a little wider for males 8 than females 7.5

Comparison: Back-to-back stem-and-leaf plots

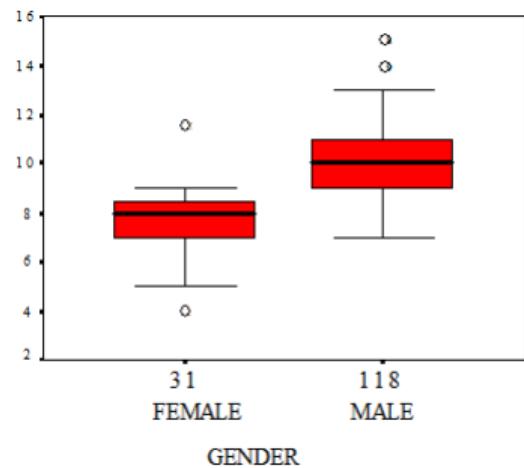
What does the data reveal?



- ➊ Distribution shape:
 - Male: bell-shaped
 - Female: skewed with a longer tail of small shoe sizes (negative skew, skewed to left)
- ➋ Outliers - Possible - but not shown here
- ➌ Centre - can only see mode
 - mode for males is 10 &
 - is higher than mode for females (8)
- ➍ Spread can determine range
 - for males is 7-15 and females 4-11.5
 - so range is a little wider for males 8 than females 7.5
- ➎ Pattern
 - M: bell within a bell; F: not so clear
 - M & F: fewer half sizes

Comparison: Box Plots

What does the data reveal?

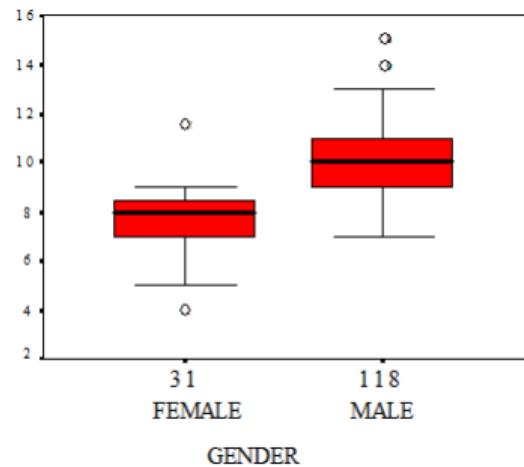


① Context: shoe size

- 118 Males & 31 Females

Comparison: Box Plots

What does the data reveal?



① Context: shoe size

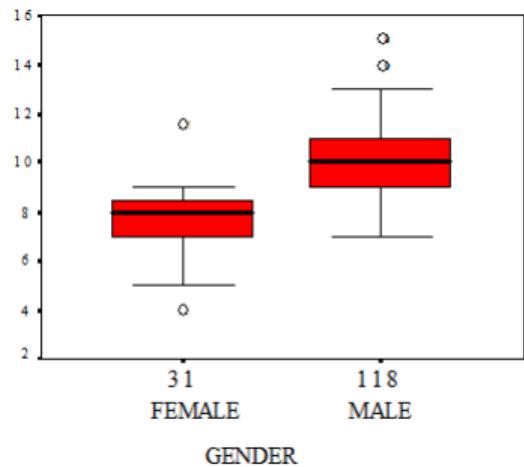
- 118 Males & 31 Females

Comparison: Box Plots

What does the data reveal?

② Distribution shape:

- F is more asymmetric than M with relatively shorter tail of upper values

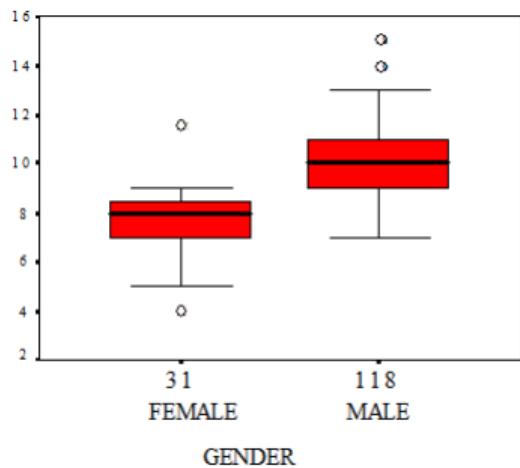


① Context: shoe size

- 118 Males & 31 Females

Comparison: Box Plots

What does the data reveal?



② Distribution shape:

- F is more asymmetric than M with relatively shorter tail of upper values

③ Outliers

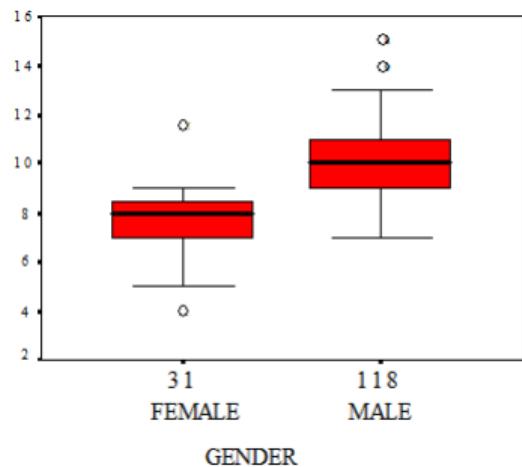
- M: two high (sizes 14 & 15)
- F: a low (size 4) & a high (size 11.5)

① Context: shoe size

- 118 Males & 31 Females

Comparison: Box Plots

What does the data reveal?



② Distribution shape:

- F is more asymmetric than M with relatively shorter tail of upper values

③ Outliers

- M: two high (sizes 14 & 15)
- F: a low (size 4) & a high (size 11.5)

④ Centre - can only see median

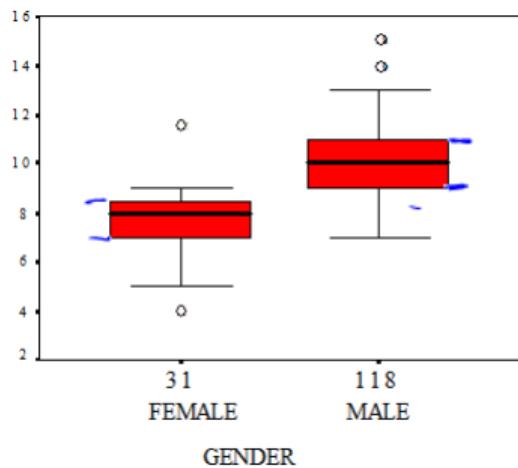
- median for M is size 10 &
- is higher than for F (size 8)

① Context: shoe size

- 118 Males & 31 Females

Comparison: Box Plots

What does the data reveal?

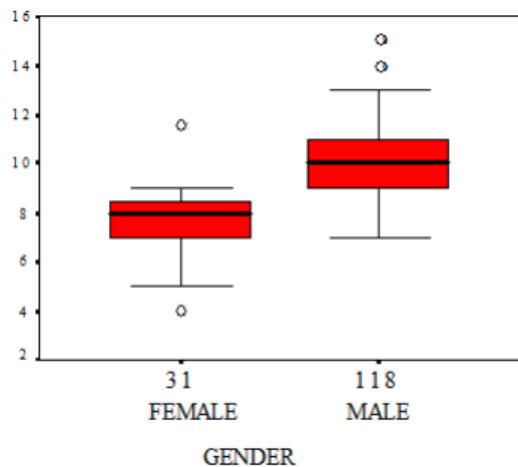


- Context: shoe size
 - 118 Males & 31 Females

- Distribution shape:
 - F is more asymmetric than M with relatively shorter tail of upper values
- Outliers
 - M: two high (sizes 14 & 15)
 - F: a low (size 4) & a high (size 11.5)
- Centre - can only see median
 - median for M is size 10 &
 - is higher than for F (size 8)
- Spread: can determine IQR and range
 - IQR for M is $11-9=2$ and F $8.5-7=1.5$
 - IQR is slightly greater for M than F
 - range is a little wider for M 8 than F 7.5

Comparison: Box Plots

What does the data reveal?



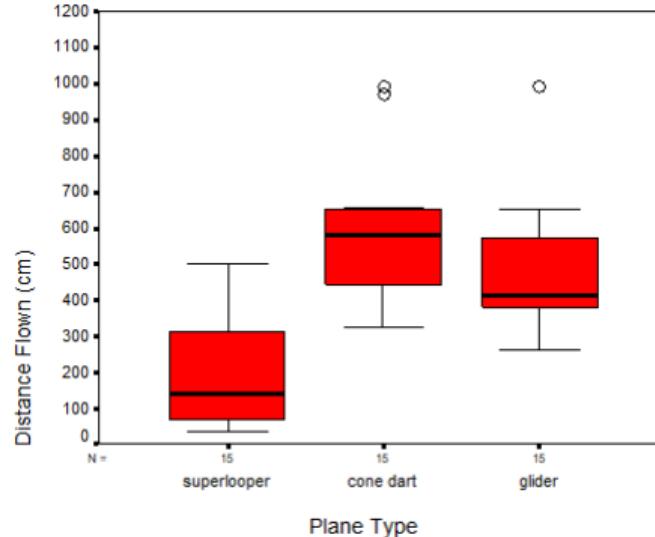
- ① Context: shoe size
- 118 Males & 31 Females

- ② Distribution shape:
 - F is more asymmetric than M with relatively shorter tail of upper values
- ③ Outliers
 - M: two high (sizes 14 & 15)
 - F: a low (size 4) & a high (size 11.5)
- ④ Centre - can only see median
 - median for M is size 10 &
 - is higher than for F (size 8)
- ⑤ Spread: can determine IQR and range
 - IQR for M is $11-9=2$ and F $8.5-7=1.5$
 - IQR is slightly greater for M than F
 - range is a little wider for M 8 than F 7.5
- ⑥ Pattern - cannot be seen in this plot

Utility: Boxplots versus Stem-and-leaf Plots

• Boxplots

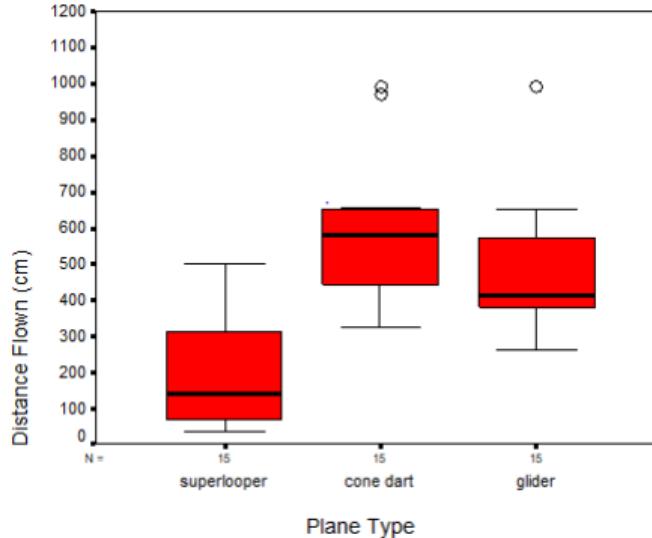
- are especially useful for comparing ≥ 2 samples or batches.
- show the 5-number summary and outliers
- but not the individual values.



Utility: Boxplots versus Stem-and-leaf Plots

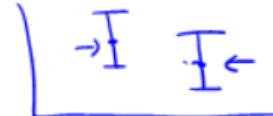
• Boxplots

- are especially useful for comparing ≥ 2 samples or batches.
- show the 5-number summary and outliers
- but not the individual values.



• Stem-and-leaf plots

- show individual values, and
- give a better picture of the shape of the spread,
- but their detail makes them unsuitable for comparing more than two groups (back-to-back)
- not suitable when a large no. of observations



Topic: Exploratory Data Analysis (EDA)

Presentation of Bivariate Data

Part C: Two quantitative variables

School of Mathematics and Applied Statistics



UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Bivariate data: Two Variables

Different tables / plots for different data types . . .

For **two qualitative** variables:

- two-way tables
- stacked bar graphs
- clustered bar graphs

For **one quantitative** and **one qualitative** variable:

- side-by-side box plots
- back-to-back stem & leaf plots

For **two quantitative** variable/s:

- ✓
- scatterplots
 - line plots (against time)

Two Continuous Variables: Where in the statistical process?

- Ethics
- Nature of the question to be answered

Is there a linear relationship between two quantitative variables?

- Context/Expertise
- Design:
 - Experiment vs. observational study
 - Sampling
 - Measurement
- **Description and analysis**
Scatterplots
- Conclusions and decision making



VARIATION

Bivariate Data Analysis

	<i>x</i>	<i>y</i>
1	Mid Session	Exam
2	1	51.7
3	2	96.7
4		38.3
5		78.3
6		91.7
7		68.3
8		83.3
9		63.3
10		58.3
11		60.0
12		53.3
13		98.3
14		73.3
15		81.7
16		70.0
17		80.0
18		93.3
19		91.7
20		83.3
21		90.0
22	21	65.0

Two quantitative variables



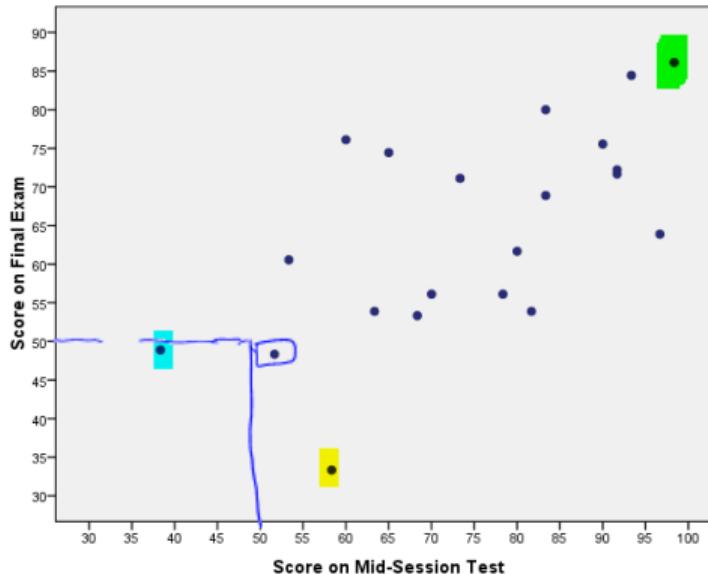
- Plot the (x,y) pairs of points on a **scatterplot**
 - one to be a response variable which is on the y-axis
It is sometimes called the dependent variable
 - and an explanatory variable on the x-axis
It is sometimes called the independent variable

$$n = 21$$

Bivariate: Scatterplot

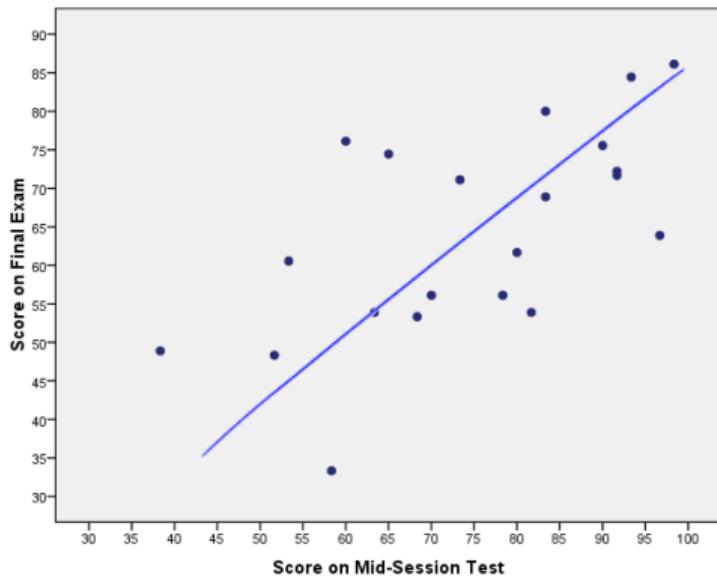
Consider the mark on a mid-session test and the mark on the final exam for 21 students studying a statistics course.

1	Mid Session	Exam
2	51.7	48.3
3	96.7	63.9
4	38.3	48.9
5	78.3	56.1
6	91.7	72.2
7	68.3	53.3
8	83.3	80.0
9	63.3	53.9
10	58.3	33.3
11	60.0	76.1
12	53.3	60.6
13	98.3	86.1
14	73.3	71.1
15	81.7	53.9
16	70.0	56.1
17	80.0	61.7
18	93.3	84.4
19	91.7	71.7
20	83.3	68.9
21	90.0	75.6
22	65.0	74.4



Bivariate: Scatterplot

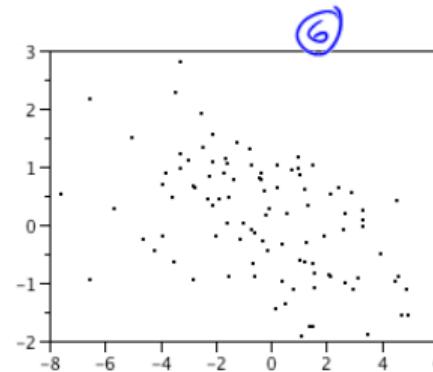
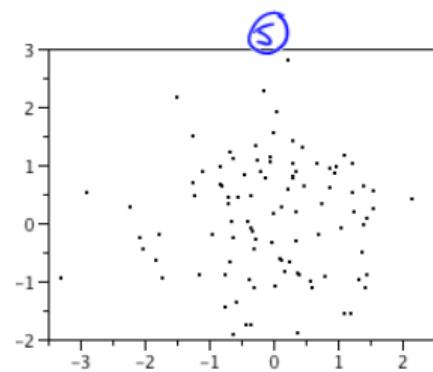
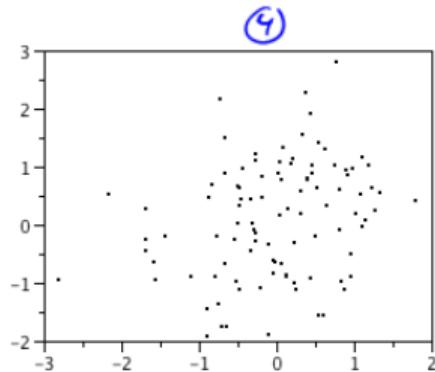
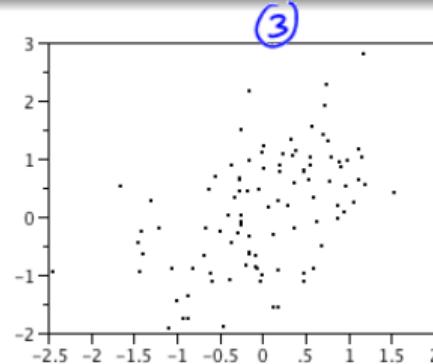
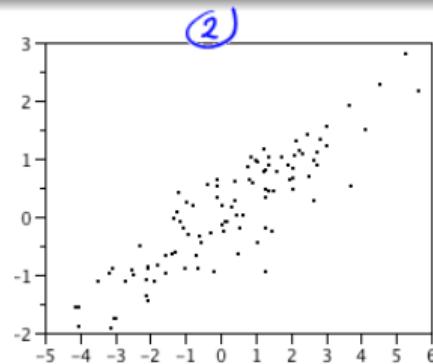
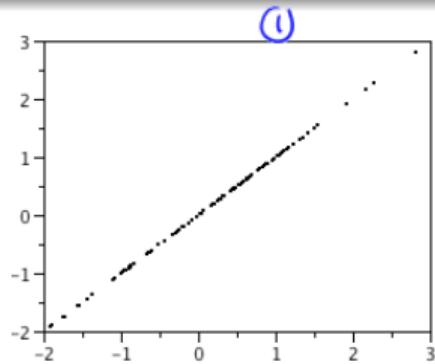
Scatterplot: Examines how two variables vary together.



Consider the mark on a mid-session test and the mark on the final exam for 21 students studying a statistics course.

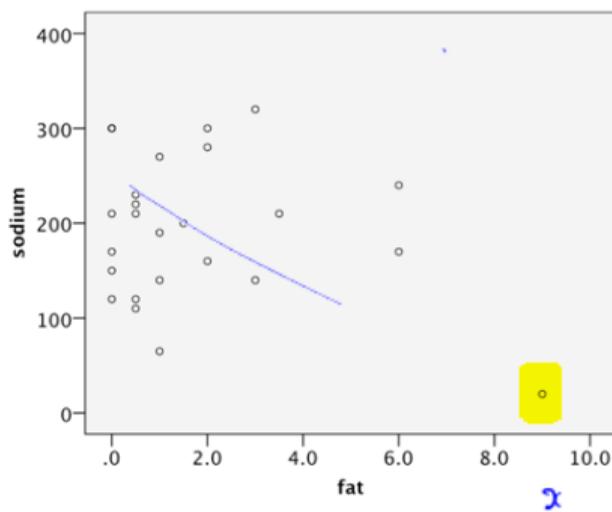
As marks on Mid-session test increase, marks on Final exam tend to increase.

More Examples



Check Plot - Make Sense of Data in Context

Context is used to make sense of real data: Cans of soup.



What do you see?

Sodium (Y) versus Fat (X)

- There is one outlier for Fat
- Is it also an outlier for Sodium?
- It is a bivariate outlier

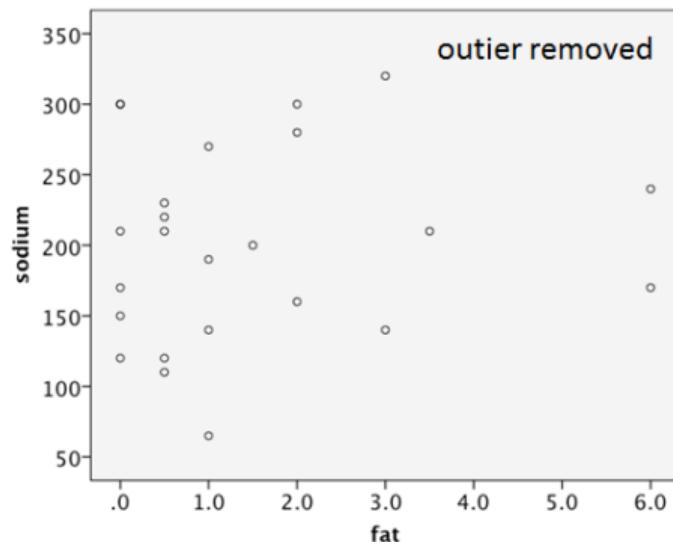
What do we do?

- Check original data for possible recording error since any further analysis will be affected by it.
- Remove the point and redo the plot.

Redo Plot

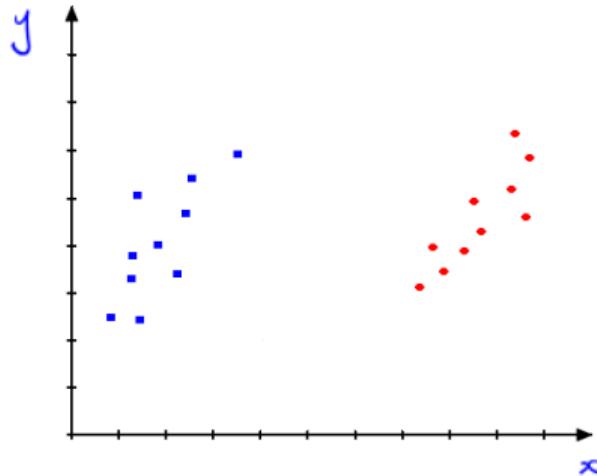
Context is used to make sense of real data: Cans of soup.

Scatterplot with outlier removed:



Plot Reveals Clusters of Points

Sometimes data points separate into two or more different **clusters** which may indicate different groups that should be examined separately:



Ask: Is there another **factor or qualitative variable** which explains the separate clusters?

Topic: Exploratory Data Analysis (EDA)

Presentation of Bivariate Data

Part D: Time Series

School of Mathematics and Applied Statistics



UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Two Continuous Variables with one as Time: Where in the statistical process?

- Ethics
- **Nature of the question to be answered**
 - * *What is the nature of variation over time?*
- Context/Expertise
- Design:
 - Experiment vs. observational study
 - Sampling
 - Measurement
- **Description and analysis:**
 - Line plots and Scatterplots*
- Conclusions and decision making

VARIATION

Time Series: How do data vary over time?

A time series is a collection of univariate data in which the values are recorded at successive time intervals

- measurements may be discrete or continuous
- observations may be observed in discrete or continuous time
- **Examples:**
 - Temperature; Rainfall
 - Sales
 - Road deaths by month
 - Share prices
 - Employment
 - Tourist arrivals ...

What might we be interested in when we look at measurements collected over successive time intervals?

An example: BHP share prices

Daily opening share price from 1st Sept to 31st Dec 2015 (n=85 obs)



Sequence	Date	Open Price
168	1-Sep-2015	34.42
169	2-Sep-2015	35.27
170	3-Sep-2015	34.94
171	4-Sep-2015	33.75
172	8-Sep-2015	34.94
173	9-Sep-2015	35.36
174	10-Sep-2015	34.04
175	11-Sep-2015	34.1
176	14-Sep-2015	33.88
177	15-Sep-2015	33.66
178	16-Sep-2015	34.93
179	17-Sep-2015	35.35
180	18-Sep-2015	34.78
181	21-Sep-2015	34.3
...

Sequence	Date	Open Price
168	1-Sep-2015	34.42
169	2-Sep-2015	35.27
170	3-Sep-2015	34.94
171	4-Sep-2015	33.75
172	8-Sep-2015	34.94
173	9-Sep-2015	35.36
174	10-Sep-2015	34.04
175	11-Sep-2015	34.1
176	14-Sep-2015	33.88
177	15-Sep-2015	33.66
178	16-Sep-2015	34.93
179	17-Sep-2015	35.35
180	18-Sep-2015	34.78
181	21-Sep-2015	34.3
...

- What do these data reveal?

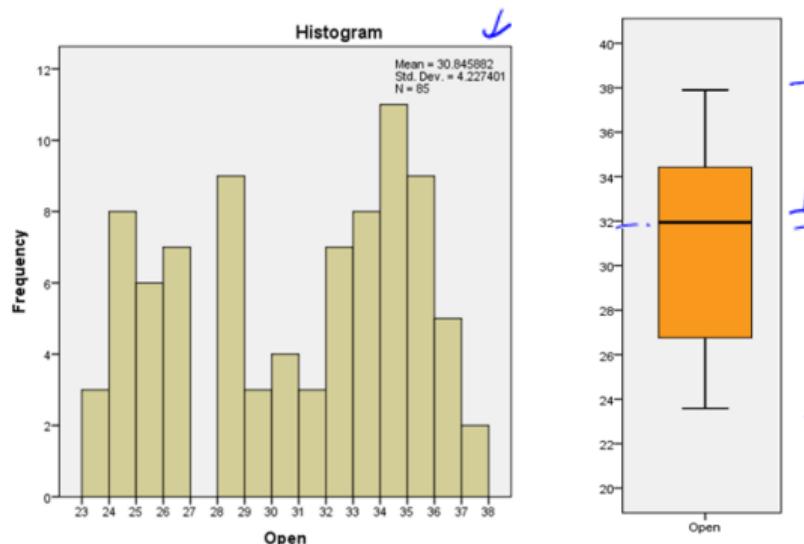
BHP Share Prices: Univariate distribution

What do the univariate plots reveal? no order considered

BHP Open Price Stem-and-Leaf Plot

Frequency	Stem & Leaf
3.00	2 . 333
14.00	2 . 44444444555555
7.00	2 . 6666666
12.00	2 . 88888888999
7.00	3 . 0000111
15.00	3 . 22222233333333
20.00	3 . 44444444444555555555
7.00	3 . 6666677

Stem width: 10.00000
Each leaf: 1 case(s)



- Shape: Slightly skewed to left
- Centre: median about \$32 and mean is \$30.85
- Spread: range is $\approx 37 - 23 = 15$ and sd is \$4.23
- Outliers: None
- Patterns: None

Time Series

In simple univariate analysis, use

- Stem-and-leaf plots,
- Boxplots
- histograms

to find centre, shape of the distribution, spread, outliers, patterns (S&L)

But these plots ignore the **time dimension**

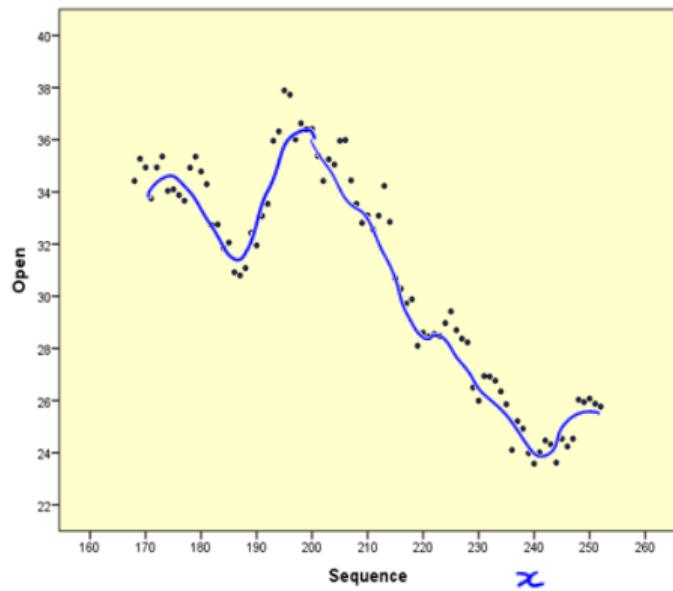
So if there is a series of data points over time, we need a
different technique to reveal the components of the time series

To represent a **time series** a line plot is more appropriate.

BHP Share Prices: Time Series

Scatterplot: Opening share price on vertical axis against time on the horizontal axis.

What does it reveal?



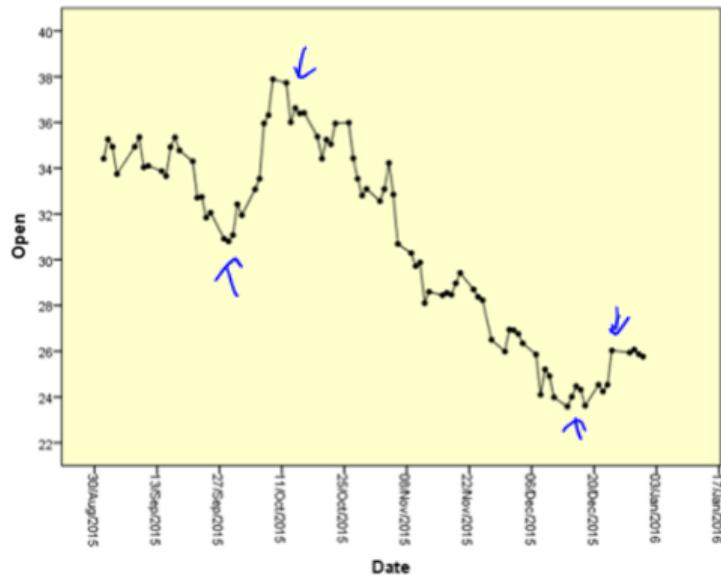
- There is quite a bit of variation or scatter
- But there is an overall pattern of price over time
- This is called the trend.

BHP Share Prices: Time Series

Line Plot: Price by time: Join the data points in the scatterplot in sequence order

What does it reveal?

BHP Daily Opening Share Price: 1st Sept - 31st Dec 2015



- There are **fluctuations** within the overall trend
- We can see highs (peaks) and lows (troughs)

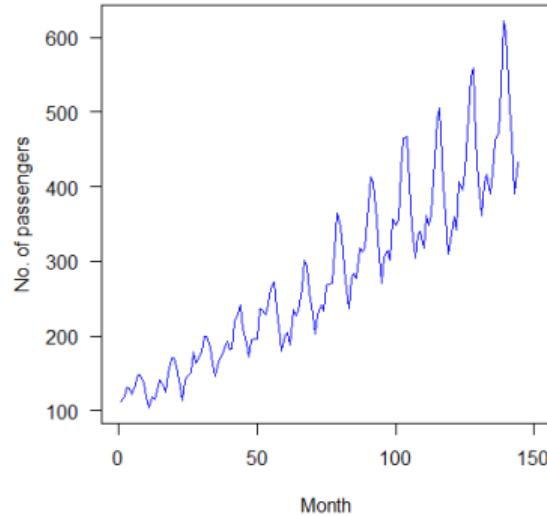
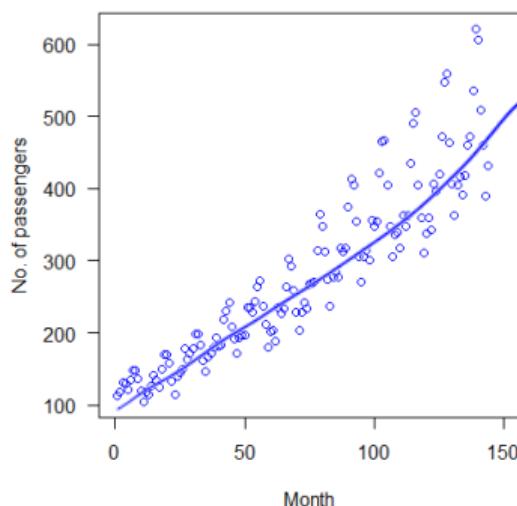
Components of a Time Series

A time series may consist of components such as

- a trend
- cyclical and /or seasonal variation
- random variation

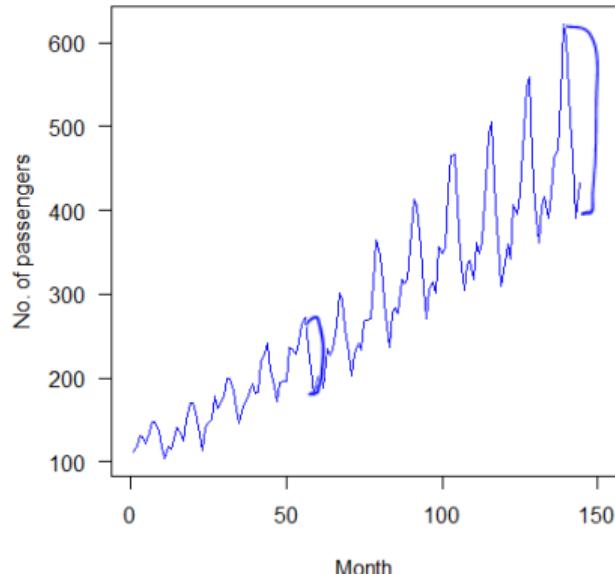
Example: Number of airline passengers over time

1949 → 1960
 $T_{an} - V_{an}$
⇒ 144 data points



Seasonal Patterns

Example: Number of airline passengers over time

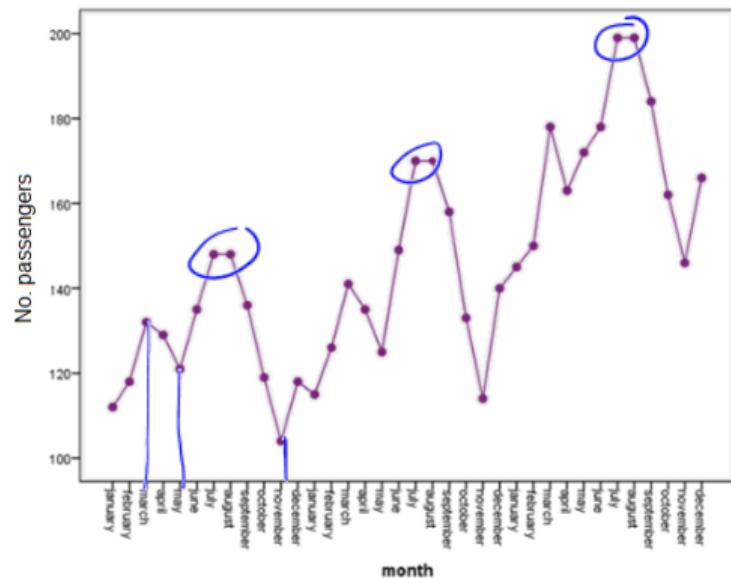


What do you observe?

- increase in number of passengers over time.
- Appears to be a seasonal pattern
- Peaks within the repeating pattern are getting higher over time
- More variation in later years
- This is called a multiplicative time series

Seasonal Patterns - Zooming in

Number of airline passengers over time : 3-year slice



What else can you observe?

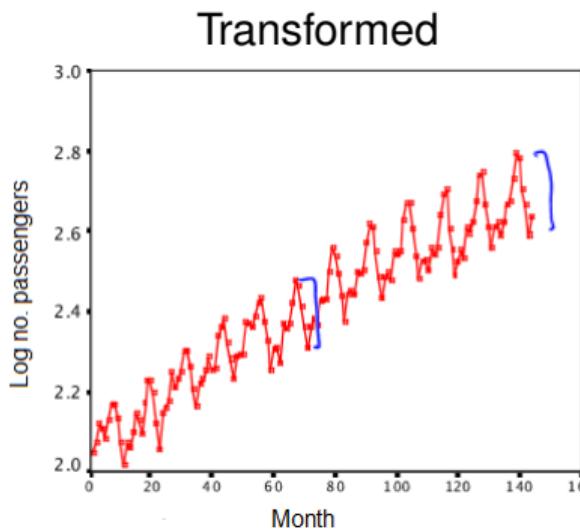
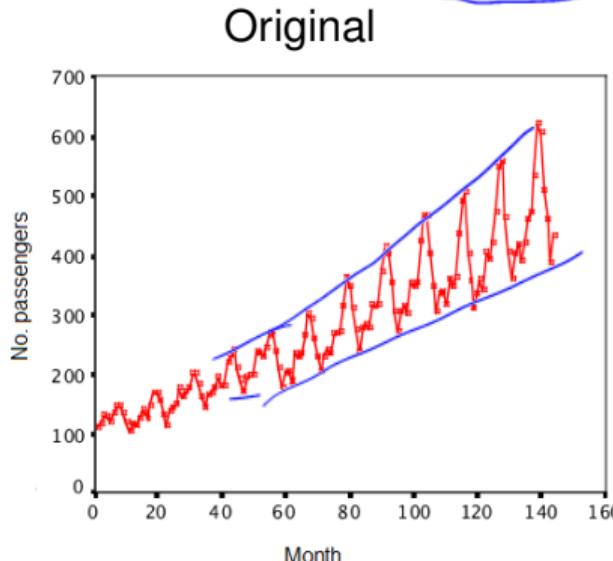
- Appears to be a seasonal pattern
 - First peak each year at about March
 - Dip at about May
 - High peak at about Jul - Aug
 - Low occurs about Nov., 3 months after high peak
- Not entirely same pattern each year

Transformed data: $\log(\text{passengers})$

Example: Transform by taking the log of Number of airline passengers over time

What is the impact of the transformation?

- Removes the multiplicative behaviour of the seasonal pattern
- This results in an additive time series



Dynamic Plots Example: CBA ASX Chart

You may see dynamic interactive plots such as this one (see link below): customize the plot by choosing the time period, labelling a particular point, adding a moving average, comparing to another series.

Commonwealth Bank of Australia Shares



Ref: <https://www.marketindex.com.au/asx/cba>

Dynamic plots - try these

Go to:

<https://ourworldindata.org/covid-vaccinations>

Share of people who received at least one dose of COVID-19 vaccine -
- go to website - play videos

- Select countries
- CHART
- MAP

Also go to:

<https://ourworldindata.org/>

to find plots on other interesting topics such as Artificial Intelligence.

Topic: Measuring Uncertainty with Probability

Introduction, Language and Notation

School of Mathematics and Applied Statistics



UNIVERSITY
OF WOLLONGONG
AUSTRALIA

How do we measure uncertainty?

Probability is a measure of **uncertainty or likelihood**.

The probability of an event **certain** to happen is set to 1.

In **everyday life**, it is often expressed as a percentage:

- There is a 50-50 chance of getting a tail when a fair coin is tossed.
- The weather forecast states: *Today there is an 80% chance of rain.*
- There is a 20% chance that the train to Sydney from Wollongong is likely to arrive late.

What is probability?

Probability is used to quantify **unpredictability** and describe it precisely.

The probability of an event is a number between 0 and 1 indicating how **likely** it is that the event will occur when an ‘experiment’ is carried out.

What is probability?

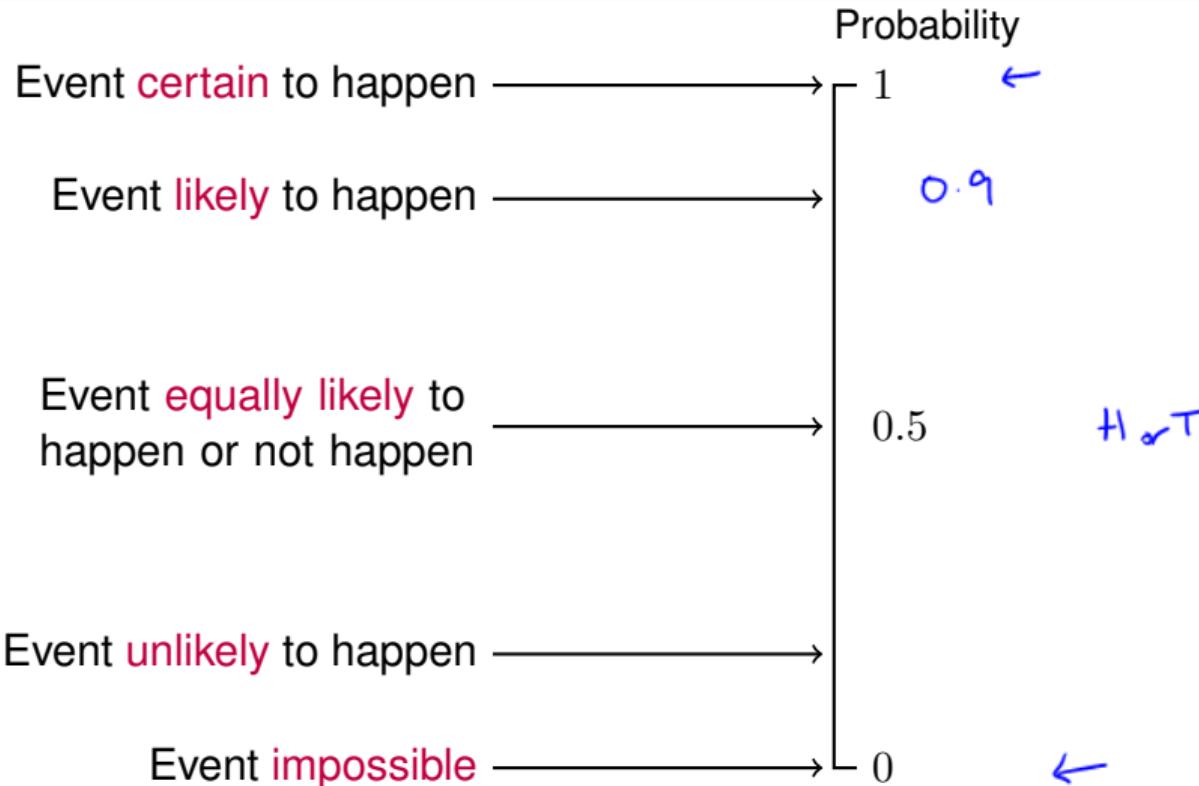
Probability is used to quantify **unpredictability** and describe it precisely.

The probability of an event is a number between 0 and 1 indicating how **likely** it is that the event will occur when an ‘experiment’ is carried out.

A probability model describes the uncertainty in an experiment by assigning probabilities to the possible outcomes.

- Weather forecast models
- Artificial intelligence uses statistics to **make independent learning decisions** where a decision is an outcome that has highest probability.

Probability Scale



Language and Notation

- **Random** phenomenon: cannot be predicted with certainty in advance
- **Experiment**: the observation of any phenomenon that is uncertain
- **Outcome** is a single observed result of random phenomenon
 - is a result of an experiment which cannot be reduced to simpler results
Example: getting a Head or Tail on the toss of a coin
- The **sample space** S is the set of all possible outcomes or sample points
 - S may be finite, countably infinite, or uncountably infinite.
 - A **discrete sample space** contains a finite or countable number of distinct sample points
 - Example: $S = \{1, 2, 3, 4, 5, 6\}$ is set of all outcomes for a throw of a die
 - $P(S) = 1$, as the sample space includes all possibilities.

Language and Notation: Events

- The subsets of S are called **events**.
 - E is a subset of the sample space $S : E \subseteq S$
 - Events are collection of outcomes, including both S and \emptyset (the null or empty set).
 - An event in a discrete sample space S is a collection of sample points, any subset of S
 - e.g. in the die experiment the event 'getting an even number' is the collection of outcomes $\{2, 4, 6\}$
- **Null event** $\{\}$ or \emptyset
 - The empty set (no outcomes) is an event which can **never** occur. e.g. even and odd: \emptyset
 - $P(\emptyset) = 0$, as \emptyset contains no possibilities. $P(\emptyset) = 0$

Language and Notation

- Intersection of events: $P(A \cap B)$

- The event that A and B both occur.

e.g. 1 die rolled: $P(\text{even } \underline{\text{and}} \text{ greater than } 4)$

$$A := \{2, 4, 6\}$$

$$B := \{5, 6\}$$

$$\epsilon_1 = \{6\}$$

- Union of events: $P(A \cup B)$

- e.g. 1 die rolled: $P(\text{even } \underline{\text{or}} \text{ greater than } 4)$

$$\epsilon_2 = \{2, 4, 5, 6\}$$

- Disjoint events

- have no outcomes in common
- If $A \cap B = \emptyset$, the events A and B are said to be disjoint or mutually exclusive; i.e. they cannot occur simultaneously.
- 1 die rolled: $P(\text{even } \underline{\text{and}} \text{ odd})=0$

$$\{2, 4, 6\}$$

$$\{1, 3, 5\}$$

Probability Axioms

- ① The **probability** of each individual outcome is a number between 0 (“can’t happen”) and 1 (“certain to happen”).
i.e. For any event E : $0 \leq P(E) \leq 1$
≤ ‘less than or equal to’
- ② **Total probability** of all outcomes = 1
i.e. $P(S) = 1$ where S represents the sample space
- ③ The probability $P(E)$ of an event E is obtained by adding probabilities of **disjoint** outcomes in E .
i.e. $P(E_1 \text{ or } E_2 \text{ happens}) = P(E_1 \cup E_2) = P(E_1) + P(E_2)$
3 4 \cup

From these basic rules of probability (the axioms) other properties of probabilities can be derived.

Summary

In this lecture segment we have considered:

- Probability - as a measure of uncertainty which is used to quantify unpredictability.
- Language and notation
- Basic probability laws or axioms

Reference: Wackerley D.D., Mendenhall W. & Scheaffer R.L. [WMS] (2008) "Mathematical Statistics with Applications", 7th ed. Duxbury, Belmont . (Library: 519.5/40).

Topic: Measuring Uncertainty with Probability

Concepts of Probability

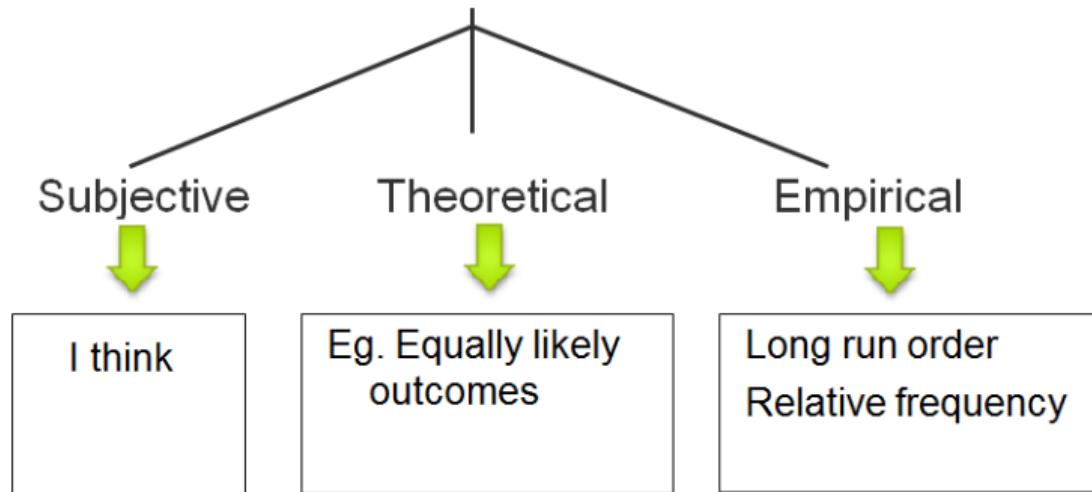
School of Mathematics and Applied Statistics



UNIVERSITY
OF WOLLONGONG
AUSTRALIA

How do we measure uncertainty?

Probability is a measure of unpredictability



Measuring Uncertainty with Probability

- **Subjective:** 'I believe the probability is'

Measuring Uncertainty with Probability

- **Subjective**: 'I believe the probability is'
- **Theoretical** eg equally likely

$$P(\text{single outcome}) = \frac{1}{\text{Total of outcomes in sample space}}$$

$$P(\text{event}) = \frac{\text{Number of outcomes}}{\text{Total no. of outcomes in sample space}}$$

Measuring Uncertainty with Probability

- **Subjective**: 'I believe the probability is'
- **Theoretical** eg equally likely

$$P(\text{single outcome}) = \frac{1}{\text{Total of outcomes in sample space}}$$

$$P(\text{event}) = \frac{\text{Number of outcomes}}{\text{Total no. of outcomes in sample space}}$$

- **Empirical**

- use observed outcomes to determine or estimate probability
- repeat experiment indefinitely long run frequency

$$P(\text{event}) = \frac{\text{Number of times event occurs}}{\text{Total no. of times experiment in repeated}}$$

Theoretical Probability - Activity

When I toss this die, what is the probability of ...

$$S = \{1, 2, 3, 4, 5, 6\}$$



① Getting a 1? $P(1) = \frac{1}{6}$

② Getting a 2 or a 3?

$$P(2 \text{ or } 3) = \frac{2}{6} = \frac{1}{3}$$

③ Getting an even number
or a no. greater than 4?

$$\{2, 4, 6\} \quad \{5, 6\}$$

$$E_1 = \{2, 4, 5, 6\}$$

$$P(E_1) = \frac{4}{6}$$

④ Getting an even number
and a no. greater than
4?

$$\{2, 4, 6\} \quad \{5, 6\}$$

$$E_2 = \{6\}$$

$$P(E_2) = \frac{1}{6}$$

Theoretical Probability - Activity

What thinking gave you those answers? or
What have you assumed in your calculations?

- fair die
- equally likely outcomes



Theoretical Probability - Activity

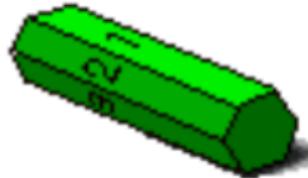
What thinking gave you those answers? or
What have you assumed in your calculations?



Could you use the same thinking for these dice? If not, what could you do?

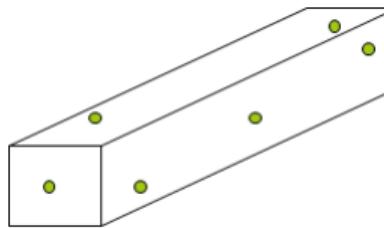
What is the probability of getting a 1 when this die is tossed?

$$\frac{1}{6}$$



Or this one?

?



Equally Likely Outcomes

- The most common experiments for which **equally likely** outcomes can be assumed are related to gambling.
- This definition of probability is **rarely** applicable in experiments involving natural phenomena.
- It would seem to be applicable, for example, to the day of the week (Monday, Tuesday, ..., Sunday) on which a wild animal is born.
- However, for humans, and perhaps even for domesticated animals, various effects such as **medical intervention** (Caesareans, induced births,...) makes births more likely on any weekday (Monday to Friday) than on Saturday or Sunday.

Equally likely outcomes cont.

When all possible outcomes are equally likely,

$$P(E) = \frac{n(E)}{n(S)}$$

where $n(E)$ = no. of outcomes in E ; and $n(S)$ = no. of outcomes in S

Example 1: A coin is tossed twice, the sequence of heads (H) and tails (T) is recorded.

- Sample Space $S = \{\textcircled{HH}, \textcircled{HT}, \textcircled{TH}, \textcircled{TT}\}$
- Let $E =$ denote the event “same result for both tosses”.

Then $E = \{\textcircled{HH}, \textcircled{TT}\}^2$ and $P(E) = \frac{n(E)}{n(S)} = \frac{2}{4} = \frac{1}{2}$

Empirical - Relative Frequency

Outcomes are **not always** equally likely: consider

- an unbalanced coin; or
- a loaded die; or
- the proportion of emails received each day of the week.

Empirical - Relative Frequency

Outcomes are **not always** equally likely: consider

- an unbalanced coin; or
- a loaded die; or
- the proportion of emails received each day of the week.

Relative frequency is used to estimate the **theoretical** probability.

- the estimate is calculated from available data
- or by experiment

A principle concern of Statistics is

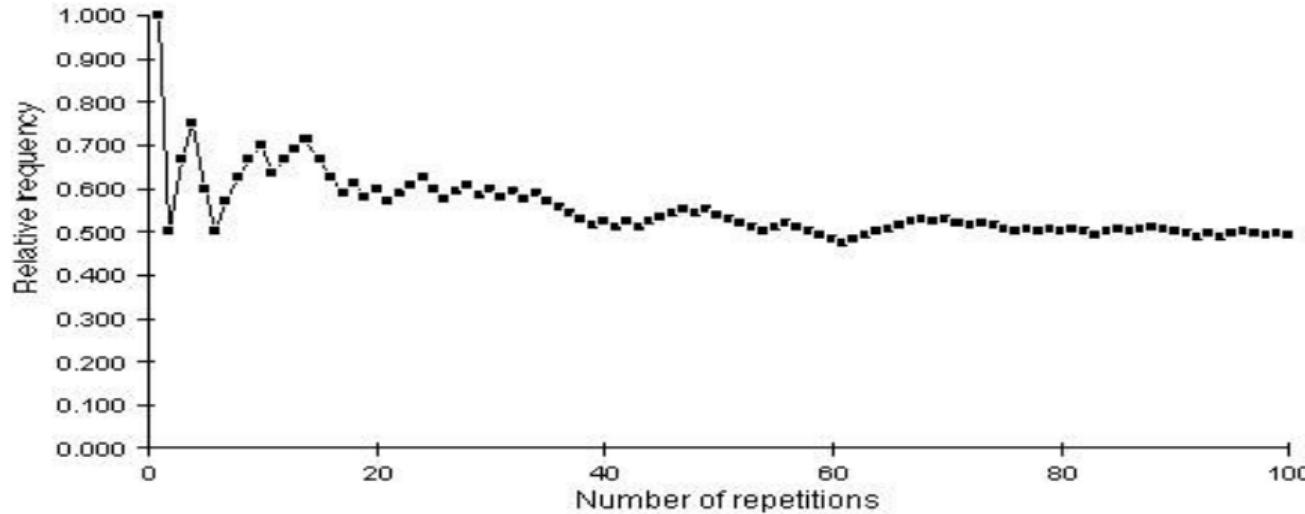
- estimating probabilities from samples...
- and using those estimates to make inferences about populations

Relative Frequency

- If we repeat the experiment indefinitely, the relative frequency of the random event from all past repetitions will **initially fluctuate markedly** but will tend to settle down or **stabilise** at a narrow band of values.
- The more times this experiment is repeated, the closer will the relative frequencies be to a particular value which we define to be the **probability** of the event.

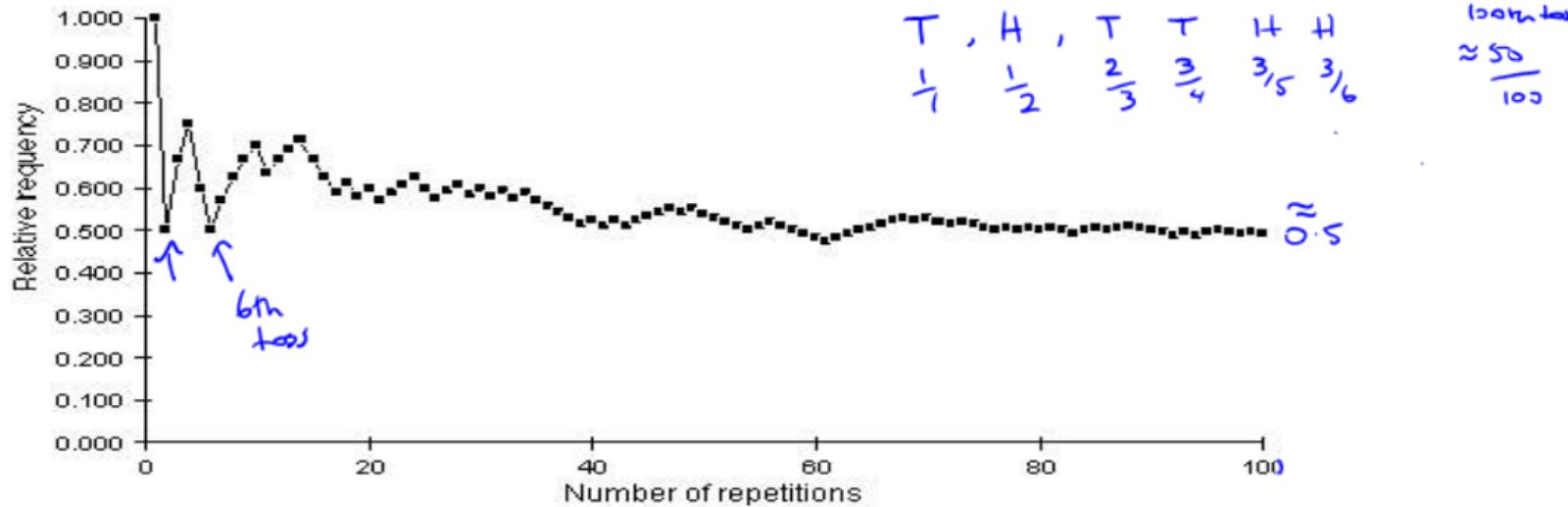
Relative Frequency

- This is an example of the phenomenon known as the **Law of Large Numbers**.



Relative Frequency

- This is an example of the phenomenon known as the **Law of Large Numbers**.



- Outcomes are **not always** equally likely; sometimes probabilities are estimated as **long-run relative frequencies**.

Example 2

The grades obtained by 500 students completing a certain subject over a number of years are as follows:

						^{TF}
F	PC	P	C	D	HD	
75	25	175	100	75	50	Total 500

- ① Estimate $P(C)$, $P(D)$, and $P(HD)$

$$P(C) = \frac{100}{500} \\ = 0.2$$

$$P(D) = \frac{75}{500} \\ = 0.15$$

$$P(HD) = \frac{50}{500} \\ = 0.1$$

- ② Let E denote the event that a randomly selected student obtains at least a credit. Estimate $P(E)$.

$$P(E) = P(C \text{ or } D \text{ or } HD)$$

$$= P(C) + P(D) + P(HD) = 0.2 + 0.15 + 0.1 \\ = 0.45$$

- ③ Are the events C , D and HD disjoint?

Yes.

Summary - Concepts of Probability

In this lecture segment we have looked at three ways that probability is measured:

- Subjective - depends on the person and information they draw on
- Theoretical - may have equally likely outcomes or not
- Empirical - relative frequencies are estimates of probabilities

Probability theory is the foundation used in **inferential statistics**.

Reference: Wackerley D.D., Mendenhall W. & Scheaffer R.L. [WMS] (2008) "Mathematical Statistics with Applications", 7th ed. Duxbury, Belmont . (Library: 519.5/40).

Topic: Measuring Uncertainty with Probability

Venn Diagrams

School of Mathematics and Applied Statistics



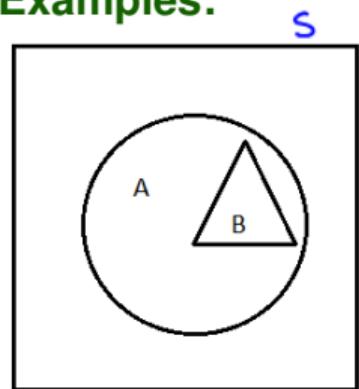
UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Introduction to Venn Diagrams

A **Venn diagram** represents events as subregions of a larger region representing the entire sample space.

It is a convenient way to represent the relationship between sets.

Examples:



B is a subset of A

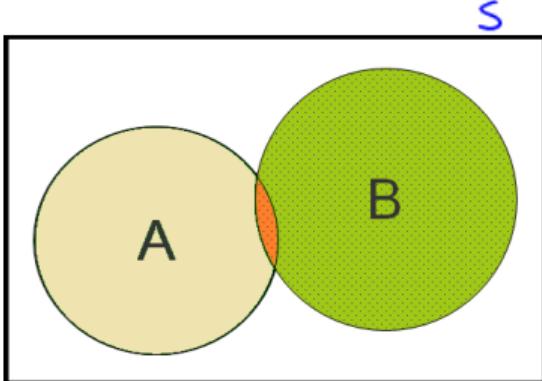
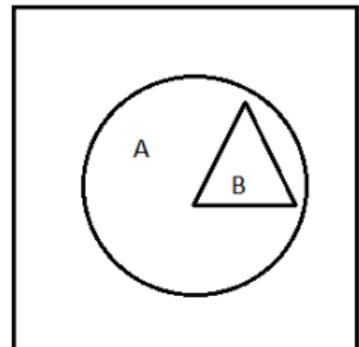
$$B \subset A$$

Introduction to Venn Diagrams

A **Venn diagram** represents events as subregions of a larger region representing the entire sample space.

It is a convenient way to represent the relationship between sets.

Examples:



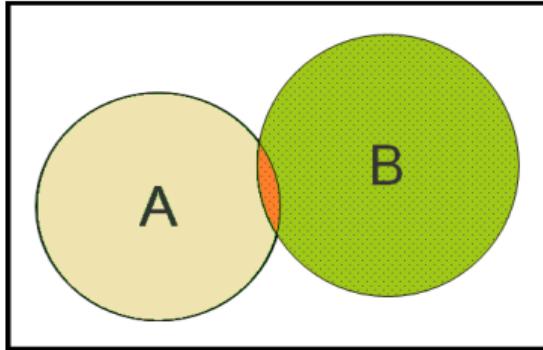
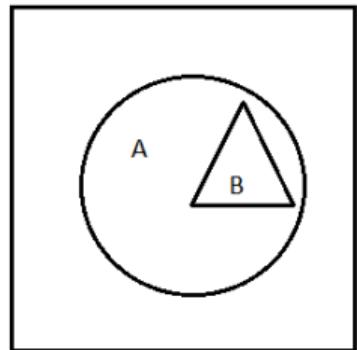
intersection

Introduction to Venn Diagrams

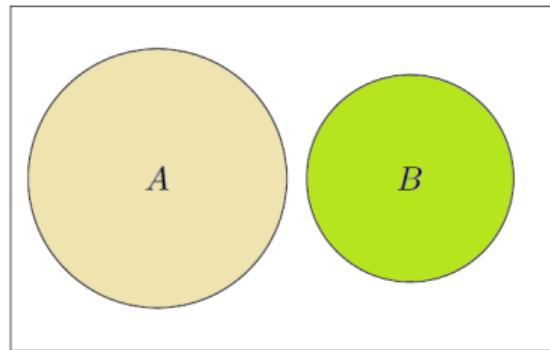
A **Venn diagram** represents events as subregions of a larger region representing the entire sample space.

It is a convenient way to represent the relationship between sets.

Examples:



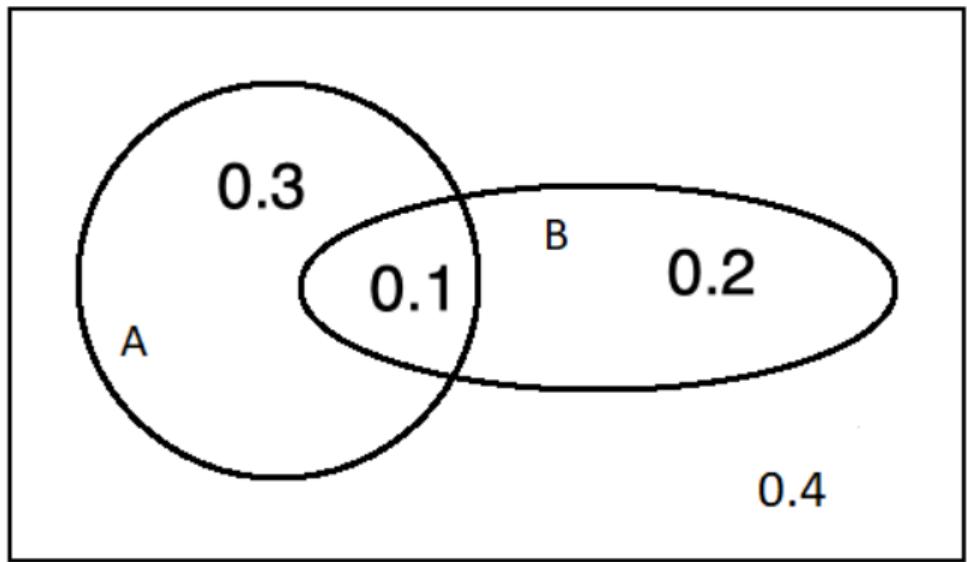
intersection



no intersection

Introduction to Venn Diagrams

Probabilities are represented as **areas** (not necessarily drawn to scale).
Numerical values (counts) can also be shown.



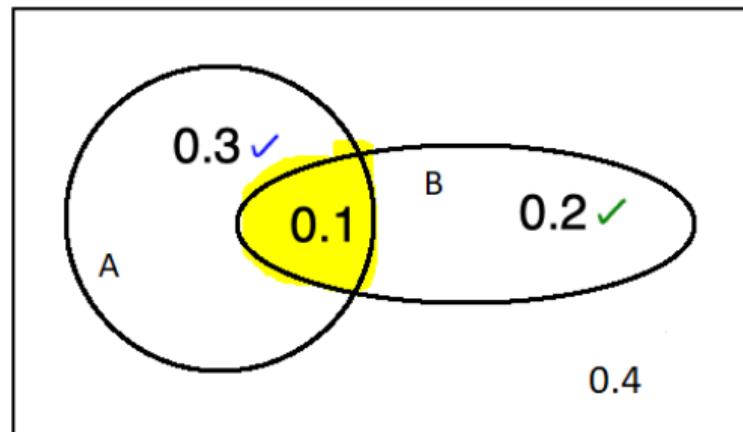
S

$$\begin{aligned} & 0.3 + 0.1 + 0.2 + 0.4 \\ & = 1.0 \end{aligned}$$

Total area = 1

Venn Diagrams & Two-way Tables

A **Venn diagram** and a **two-way table** are two different ways to represent the same information:



Total area = 1

S

	B	not B	Total
A	0.1	0.3 ✓	0.4
not A	0.2 ✓	0.4	0.6
Total	0.3	0.7	1.0 *

marginal totals

$$P(A) = 0.3 + 0.1 = 0.4$$

$$P(B) = 0.2 + 0.1 = 0.3$$

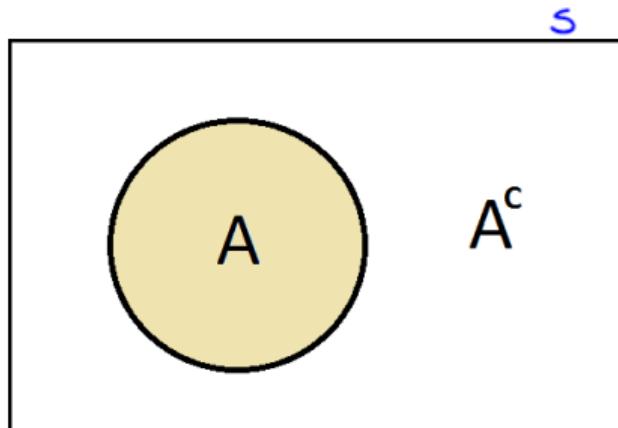
The Complement (Not)

The complement of an event is the event not occurring.

Different notations for the complement of A include: $A^c = A' = \bar{A}$

$$P(A \text{ occurs}) + P(A \text{ does not occur}) = P(A) + P(A^c) = 1$$

$$P(\text{not } A) = P(A^c) = 1 - P(A)$$



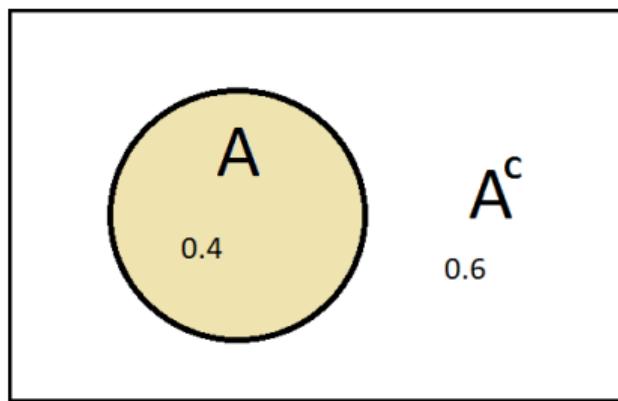
The Complement (Not)

The complement of an event is the event not occurring.

Different notations for the complement of A include: $A^c = A' = \bar{A}$

$$P(A \text{ occurs}) + P(A \text{ does not occur}) = P(A) + P(A^c) = 1$$

$$P(\text{not } A) = P(A^c) = 1 - P(A) = 1 - 0.4 = 0.6.$$

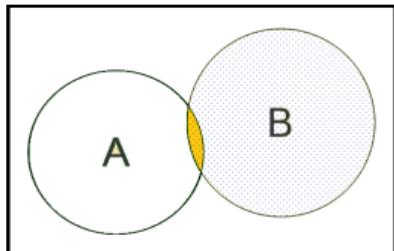


Venn diagram with two events

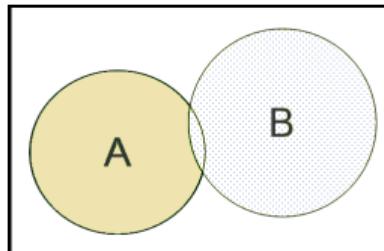


A Venn diagram with 2 events is subdivided into 4 regions:

① $A \cap B$ intersection



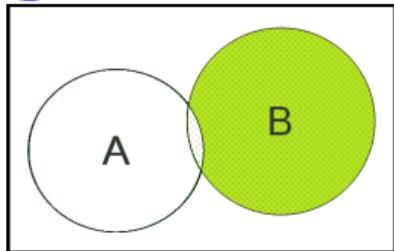
② $A \cap B^c$



① → ②

$$= P(A)$$

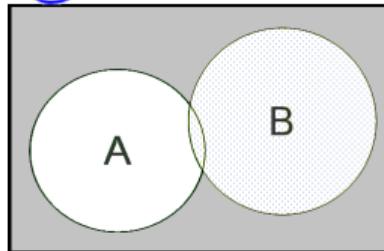
③ $A^c \cap B$



①+③

$$= P(B) \quad \text{The probabilities of these 4 events sum to 1.}$$

④ $A^c \cap B^c$



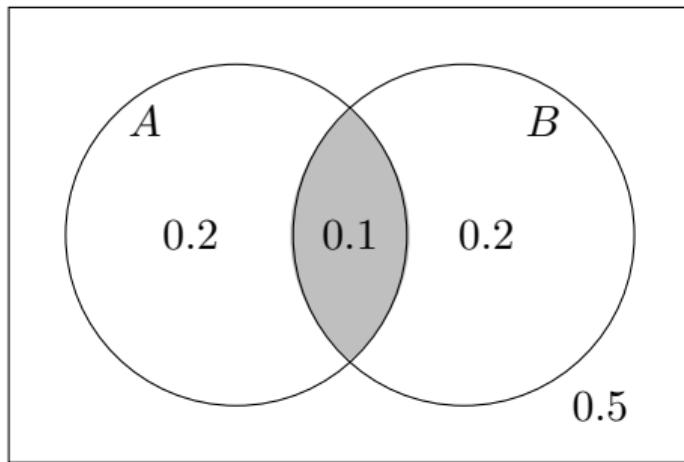
$$\textcircled{3} + \textcircled{4} = P(A^c)$$

$$\textcircled{2} + \textcircled{4} = P(B^c)$$

Intersection (And)

The **intersection** $A \cap B$ is the event that both A **and** B occur.

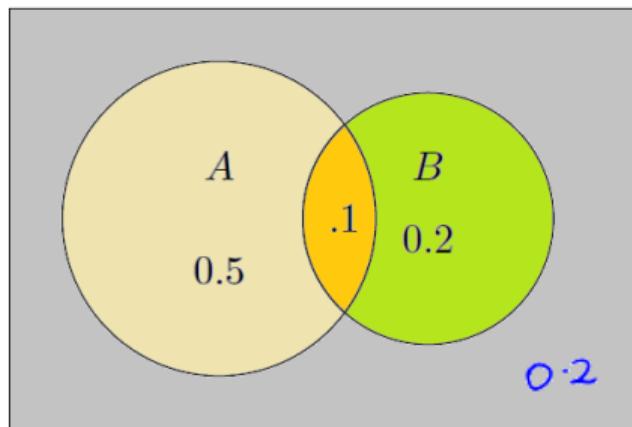
Another example:



$$P(A \cap B) = 0.1$$

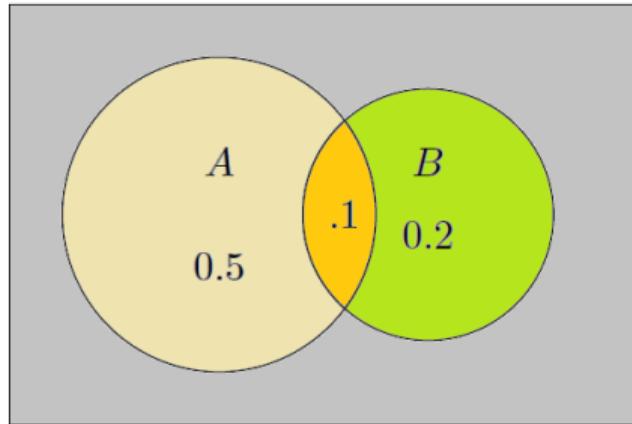
Union (Or)

The **union** $A \cup B$ consists of outcomes that are in A **or** B (or both).



Union (Or)

The **union** $A \cup B$ consists of outcomes that are in A **or** B (or both).



$$P(A) = 0.6$$

$$P(B) = 0.3$$

$$P(A \cap B) = 0.1$$

Additive Law of Probability:

The probability of the union of two events A and B is

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

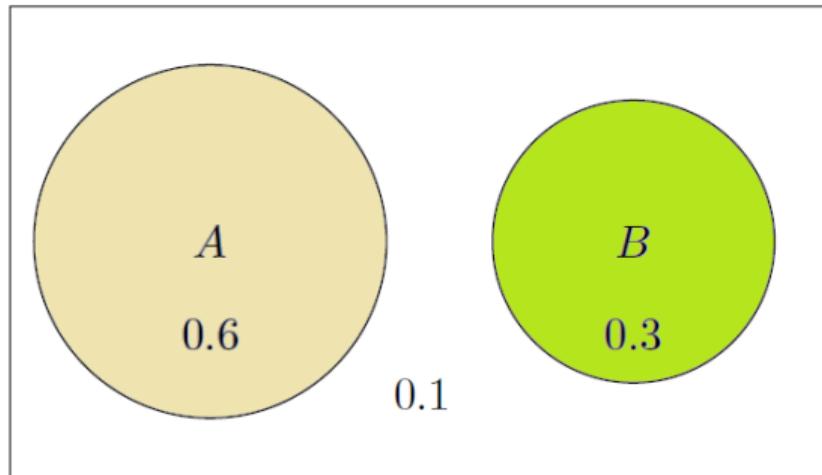
$$= 0.6 + 0.3 - 0.1$$

$$= \underline{\underline{0.8}}$$

Disjoint Events

If $A \cap B = \emptyset$ (i.e. no intersection), then the two events are said to be **mutually exclusive** or **disjoint**.

For disjoint (mutually exclusive) events, the previous result can be simplified since $P(A \cap B) = 0$:



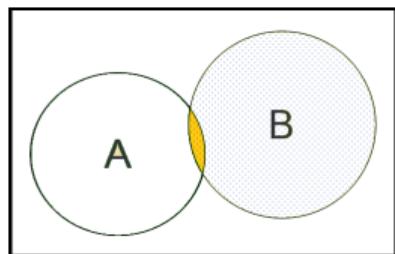
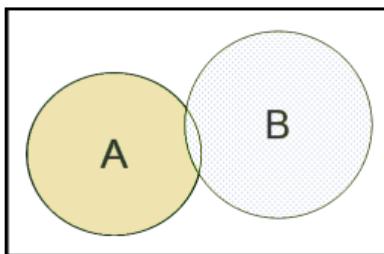
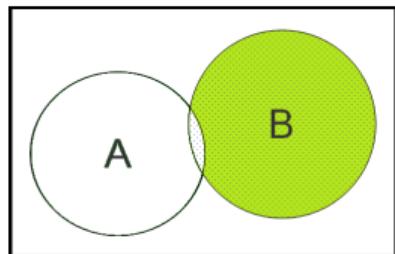
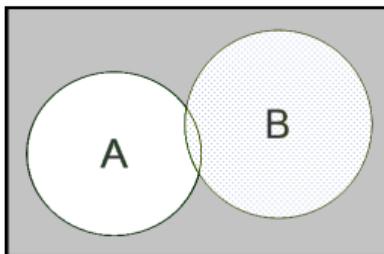
$$P(A \cup B) = P(A) + P(B)$$

$$= 0.6 + 0.3$$

$$\underline{= 0.9}$$

Summary - Venn Diagrams

If we have a 2×2 contingency table we can present the probabilities for outcomes as such:

 $A \cap B$  $A \cap B^c$  $A^c \cap B$  $A^c \cap B^c$ 

	B	$\text{not } B$	Total
A	$A \cap B$	$A \cap B^c$	$P(A)$
$\text{not } A$	$A^c \cap B$	$A^c \cap B^c$	$P(A^c)$
Total	$P(B)$	$P(B^c)$	1.0

Summary

In this lecture segment we have looked at Venn Diagrams as a convenient way to represent relationships between sets and applied set notation:

- Complement: $P(A^c) = 1 - P(A)$
- Intersection: $P(A \cap B)$
- Union: $P(A \cup B)$
- Disjoint events: $P(A \cap B) = 0$

Venn Diagrams can be extended to represent relationships between 3 or more events.

Reference: Wackerley D.D., Mendenhall W. & Scheaffer R.L. [WMS] (2008) "Mathematical Statistics with Applications", 7th ed. Duxbury, Belmont . (Library: 519.5/40).

Topic: Measuring Uncertainty with Probability

Conditional Probability

School of Mathematics and Applied Statistics



UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Conditional Probability

Definition: The probability of an event (A) occurring when it is known that some event (B) has already occurred is called a **conditional probability**.

- The **conditional** probability of event A given that event B has occurred is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$


The term $P(A \cap B)$ is annotated with a blue handwritten-style 'and' above the intersection symbol.

Conditional Probability

Definition: The probability of an event (A) occurring when it is known that some event (B) has already occurred is called a **conditional probability**.

- The **conditional** probability of event A given that event B has occurred is

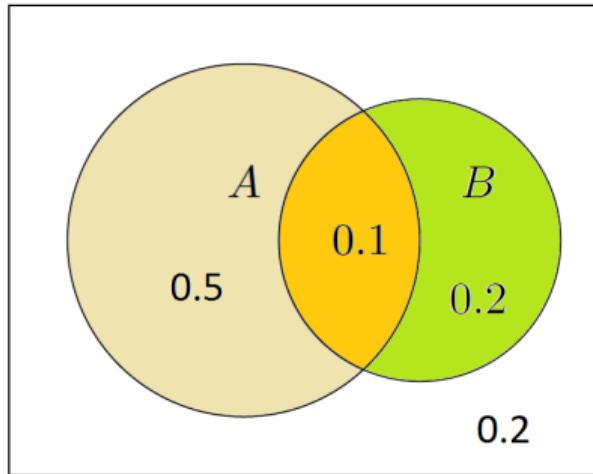
$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$P(A|B)$, can only be applied if $P(B) \neq 0$

Note: All probabilities can be considered as conditional probabilities, since $P(A)$ is really shorthand of $P(A|S)$

Conditional Probability $P(A|B)$

In terms of Venn diagrams, all of the sample space lying outside B is discarded, and B becomes the new sample space.



$$\begin{aligned}
 P(A|B) &= \frac{P(A \cap B)}{P(B)} \\
 &= \frac{0.1}{0.2 + 0.1} = \frac{0.1}{0.3} \\
 &= \frac{1}{3}.
 \end{aligned}$$

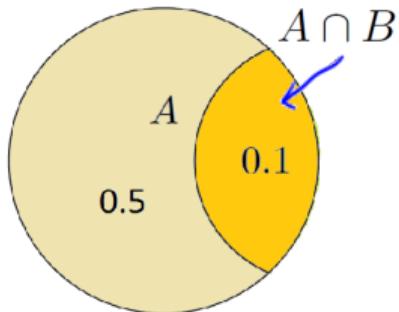
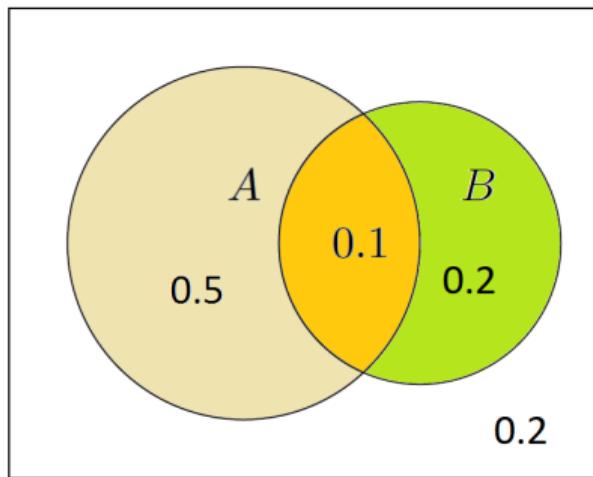
What about $P(B|A)$?

We just reverse A with B in the formula so that:

conditional probability of event B given that event A has occurred is

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{0.1}{0.5+0.1} = \frac{1}{6}$$

$(\neq P(A|B))$.

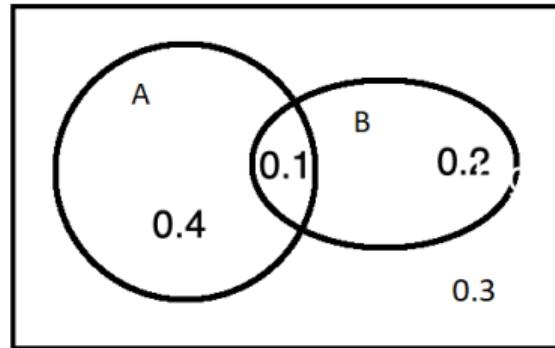


$$P(B \cap A) = P(A \cap B)$$

Conditional probability using a two-way table

To find a conditional probability using a two-way table, divide the intersection value by the appropriate marginal total.

Example:

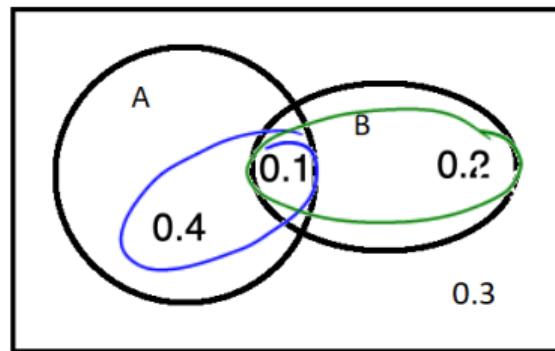


	B	B'	Total
A	0.1	0.4	0.5
A'	0.2	0.3	0.5
Total	0.3	0.7	1

Conditional probability using a two-way table

To find a conditional probability using a two-way table, divide the intersection value by the appropriate marginal total.

Example:



	B	B'	Total
A	0.1	0.4	0.5
A'	0.2	0.3	0.5
Total	0.3	0.7	1

$$P(A \cap B)$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$= \frac{0.1}{0.3} = \frac{1}{3}$$

$$P(B|A) = \frac{P(B \cap A)}{P(A)}$$

$$= \frac{0.1}{0.5} = \frac{1}{5}$$

Multiplicative Law of Probability

Sometimes, it is more convenient to start with information on the conditional probability and use it to find the joint probability (intersection).

We can just **rearrange** the rule for conditional probability:

$$P(A \text{ given } B) = P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Multiplicative Law of Probability

The probability of the intersection of two events A and B is

$$\underline{P(A \cap B) = P(B) \times P(A|B)} \quad \leftarrow$$

or $P(B \cap A) = P(A) \times P(B|A)$ ✓

Example continued

Example: If $P(A) = 0.5$, $P(B) = 0.3$, and $P(A|B) = 1/3$, determine $P(A \cap B)$

$$P(A \cap B) = P(B) \times P(A|B)$$

$$= 0.3 \times \frac{1}{3}$$

$$= \frac{3}{10} \times \frac{1}{3}$$

$$= \frac{1}{10}$$

$$= \underline{\underline{0.1}}$$

$$P(B \cap A) =$$

$$\underline{\underline{0.1}}$$

Summary

In this lecture segment we have looked at **Conditional probabilities**:

$$P(A \text{ given } B) = P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- Multiplicative Law:

$$P(A \cap B) = P(B) \times P(A|B)$$

Conditional probabilities can also be determined using **two-way tables**.

Reference: Wackerley D.D., Mendenhall W. & Scheaffer R.L. [WMS] (2008) "Mathematical Statistics with Applications", 7th ed. Duxbury, Belmont . (Library: 519.5/40).

Topic: Measuring Uncertainty with Probability

Conditional Probability - Exercise

School of Mathematics and Applied Statistics



UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Conditional Probability

Recall:

Definition: The probability of an event (A) occurring when it is known that some event (B) has already occurred is called a **conditional probability**.

- The **conditional** probability of event A given that event B has occurred is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Exercise: Conditional Probability

Exercise: Let L be the event that bus leaves on time and A the event that bus arrives on time. Assume that the probability that a bus leaves on time is 0.80, the probability that it arrives on time is 0.75, and the probability that it leaves on time and arrives on time is 0.72.

- 1 Construct a Venn diagram to represent this information
- 2 Fill the following table:

.	A	A^c	Total
L			
L^c			
Total			

- 3 Then find the probability that the bus:
 - (a) arrives on time given it left on time?
 - (b) arrives on time given that it did not leave on time?
 - (c) left on time given it arrives on time?
 - (d) arrives on times or leaves on time?

Exercise: Conditional Probability Solution

Exercise: Let L be the event that bus leaves on time and A the event that bus arrives on time. Assume that the probability that a bus leaves on time is 0.80, the probability that it arrives on time is 0.75, and the probability that it leaves on time and arrives on time is 0.72.

① Venn diagram ✓

L A

$$P(L) = 0.80$$

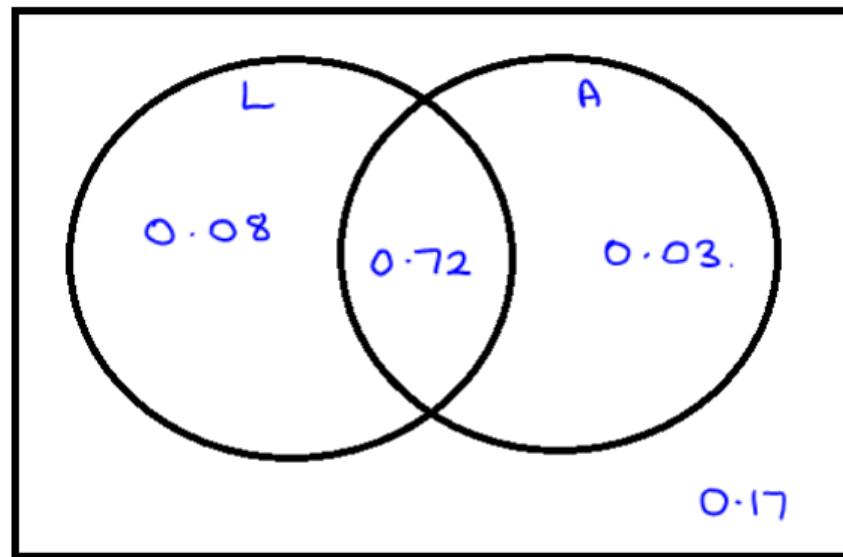
$$P(A) = 0.75$$

$$* P(L \cap A) = 0.72$$

$$0.8 - 0.72 = 0.08$$

$$0.75 - 0.72 = 0.03$$

$$\begin{aligned} P(L^c \cap A^c) &= 1 - (0.08 + 0.72 + \\ &\quad 0.03) \\ &= 0.17 \end{aligned}$$



$$\checkmark \text{Total} = 1$$

Exercise: Conditional Probability Solution cont.

- ② The two-way table: ✓

$$P(L) = 0.8$$

$$P(A) = 0.75$$

$$P(L \cap A) = 0.72$$

$$P(L^c \cap A) = 0.75 - 0.72 = 0.03$$

$$P(L \cap A^c) = 0.08$$

$$P(L^c) = 0.20$$

$$P(A^c) = 0.25$$

$$P(L^c \cap A^c) = 0.17$$

	A	A^c	Total
L	0.72	0.08	0.80
L^c	* 0.03	0.17	0.20
Total	0.75	0.25	1.0

Exercise: Conditional Probability Solution cont.

③ The probability that the bus:

(a) arrives on time given it left on time?

$$\begin{aligned} P(A | L) &= \frac{P(A \cap L)}{P(L)} \\ &= \frac{0.72}{0.80} = \underline{\underline{0.9}} \end{aligned}$$

(b) arrives on time given that it did not leave on time?

$$P(A | L^c) = \frac{P(A \cap L^c)}{P(L^c)} = \frac{0.03}{0.20} = \frac{3}{20} = \underline{\underline{0.15}}$$

(c) left on time given it arrives on time?

$$P(L | A) = \frac{P(L \cap A)}{P(A)} = \frac{0.72}{0.75} = \underline{\underline{0.96}}$$

Exercise: Conditional Probability Solution cont.

- ③ The probability that the bus:

(d) arrives on time or leaves on time?

$$\begin{aligned} P(A \text{ or } L) &= P(A \cup L) \\ &= P(A) + P(L) - P(A \cap L) \\ &= 0.75 + 0.80 - 0.72 \\ &= \underline{0.83} \end{aligned}$$

Topic: Measuring Uncertainty with Probability

Probability - Exercise

School of Mathematics and Applied Statistics



UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Recall: Equally likely outcomes

When all possible outcomes are equally likely,

$$P(E) = \frac{n(E)}{n(S)}$$

where $n(E)$ = no. of outcomes in E ; and

$n(S)$ = no. of outcomes in S

Recall: Conditional Probability

Recall:

Definition: The probability of an event (A) occurring when it is known that some event (B) has already occurred is called a **conditional probability**.

- The **conditional** probability of event A given that event B has occurred is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$



Exercise: Probability Solution

Exercise: Two fair dice are rolled.

- ① Draw up a table to show all the possible outcomes of the sums.
- ② What is the probability of getting a 1 on Die 1 and a 3 on Die 2?
- ③ What is the probability that the sum of faces is 4?
- ④ Given that the sum is 4, what is the probability of “doubles”?

Exercise: Probability Solution

- ① Table of possible outcomes of the sums. *Dice 2.*

Dice 1

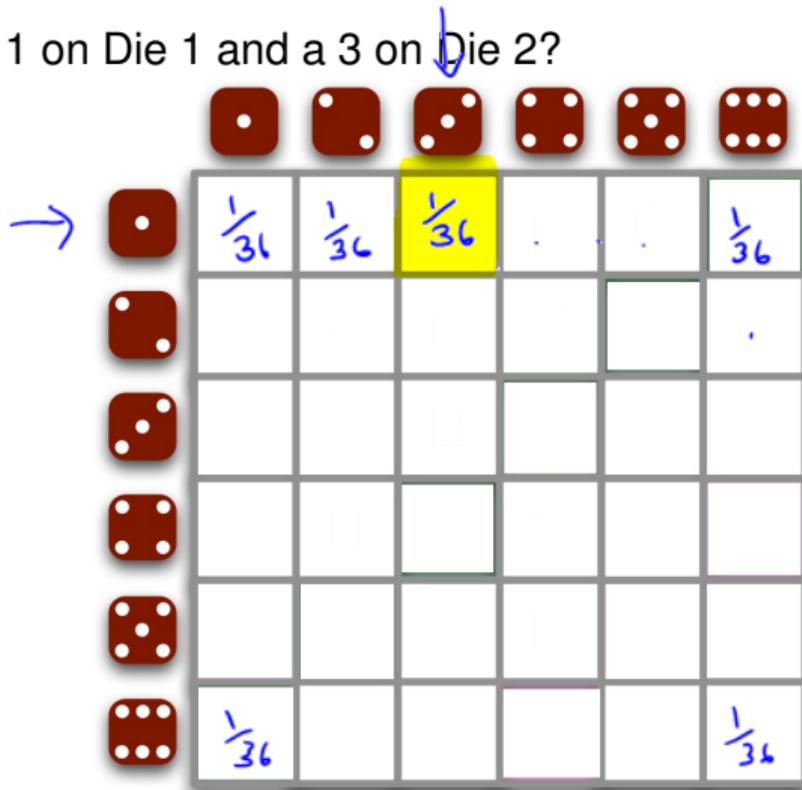
	2	3	4	5	6	7
	3	4	5	6	7	8
	4	5	6	7	8	9
	5	6	7	8	9	10
	6	7	8	9	10	11
	7	8	9	10	11	12

36 possible outcomes

Exercise: Probability Solution cont.

- ② What is the probability of getting a 1 on Die 1 and a 3 on Die 2?

$$P(1, 3) = \frac{1}{36}$$



Exercise: Probability Solution cont.

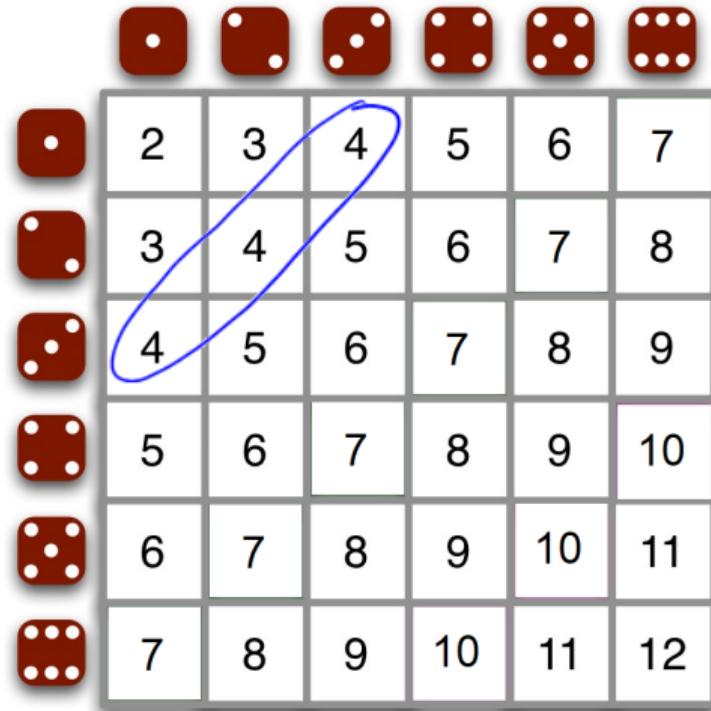
- ③ What is the probability that the sum of faces is 4?

$$\begin{array}{rcl} 1 + 3 & = 4 \\ & \text{sum.} \end{array}$$

$$2 + 2 = 4$$

$$3 + 1 = 4$$

$$\begin{aligned} P(\text{sum is } 4) &= \frac{n(E)}{n(S)} \\ &= \frac{3}{36} \\ &= \frac{1}{12} \end{aligned}$$



Exercise: Probability Solution cont.

- ④ Given that the sum is 4, what is the probability of “doubles”? B

$$P(\text{doubles} \mid \text{sum is } 4)$$

$$= P(A \mid B)$$

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

$$= \frac{1/36}{1/12}$$

$$= \frac{1}{36} \times \frac{12}{1}$$

$$= \frac{1}{3}. \quad \checkmark$$

•	2	3	4	5	6	7
•	3	4	5	6	7	8
•	4	5	6	7	8	9
•	5	6	7	8	9	10
•	6	7	8	9	10	11
•	7	8	9	10	11	12

Topic: Measuring Uncertainty with Probability

Probability - Introduction to Tree Diagrams

School of Mathematics and Applied Statistics



UNIVERSITY
OF WOLLONGONG
AUSTRALIA

What is a Probability Tree Diagram?

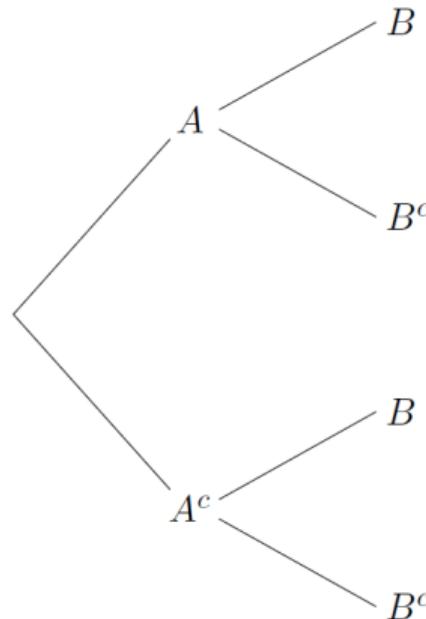
A **tree diagram** is a helpful tool when working with **multi-stage experiments** or **composite events**.

They can help with

- determining the sample space
- calculating probabilities.

Write

- each outcome at the end of the branch
- the probability on the branch



Tree Diagrams

- Conditional probabilities correspond to second (or higher) level branches in a **tree diagram**.
 $P(B|A)$ $P(B|A^c)$

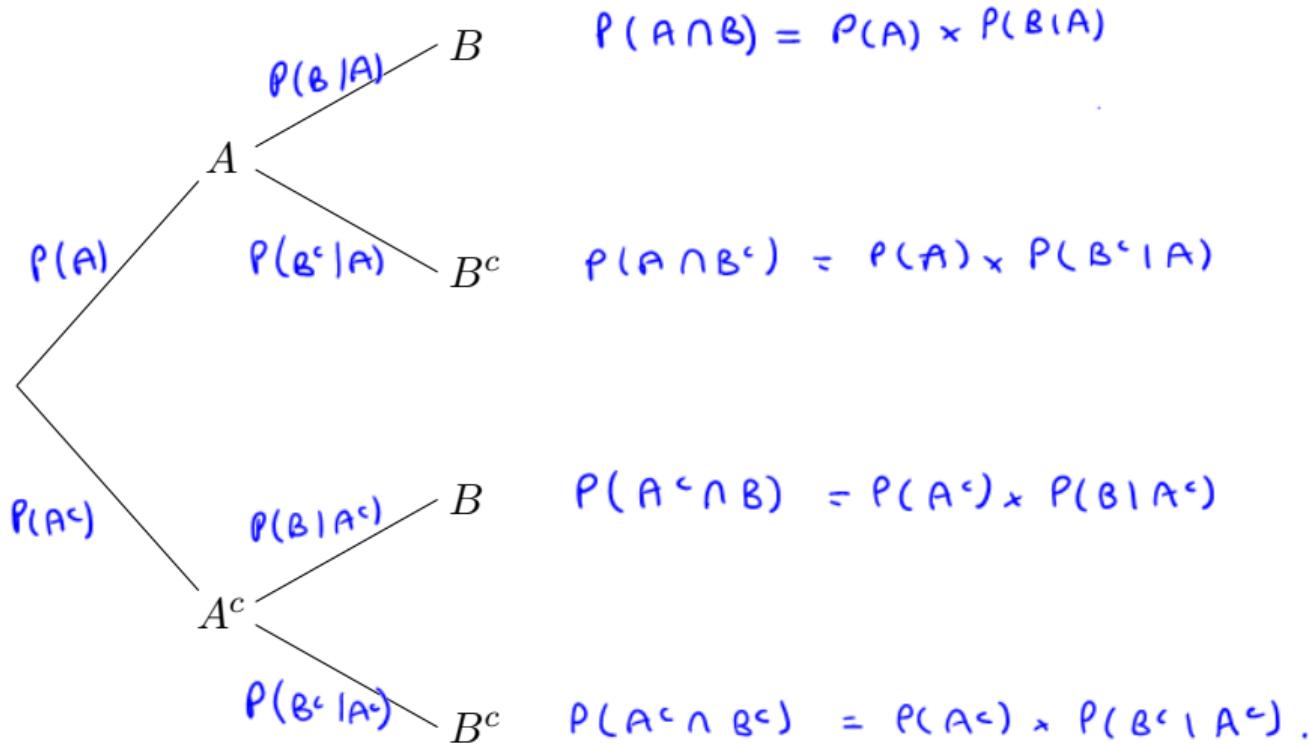
- Multiply** probabilities of all branches along a **path** to find the probability of a single outcome (using the multiplicative law of probability):

$$\underbrace{P(A \cap B)}_{\checkmark} = P(A) \times P(B|A) \quad \checkmark$$

- Sum** probabilities of all paths leading to an **event** to find its probability. The paths represent mutually exclusive outcomes.

$$P(B) = P(A \cap B) + P(A^c \cap B).$$

Structure of a Tree Diagram



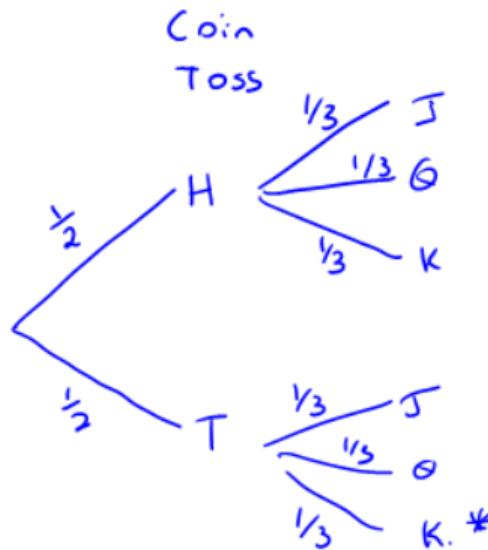
Example

H/T

A fair coin is tossed and one card is drawn from a set of 3 cards labelled: Jack, Queen, King.

- a. Create a tree diagram to determine the outcomes in the sample space and $n(S)$.
 $n(S) = |S| = \text{no. outcomes in } S$.
- b. Determine the probability of getting a Tail and a King.

Example cont.



(a)

HJ

HQ

HK

TJ

TQ

TK

$n(S) = 6$ possible
outcomes

(b)

$$P(T \cap K)$$

$$= P(T) \times P(K|T)$$

$$= \frac{1}{2} \times \frac{1}{3}$$

$$= \frac{1}{6}.$$

Topic: Measuring Uncertainty with Probability

Probability - Tree Diagram Exercise

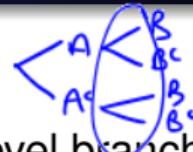
School of Mathematics and Applied Statistics



UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Tree Diagrams

Recall:



- Conditional probabilities correspond to second (or higher) level branches in a **tree diagram**.
- Multiply** probabilities of all branches along a **path** to find the probability of a single outcome (using the multiplicative law of probability):

$$P(A \cap B) = P(A) \times \underbrace{P(B|A)}$$

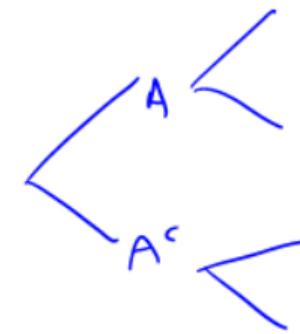
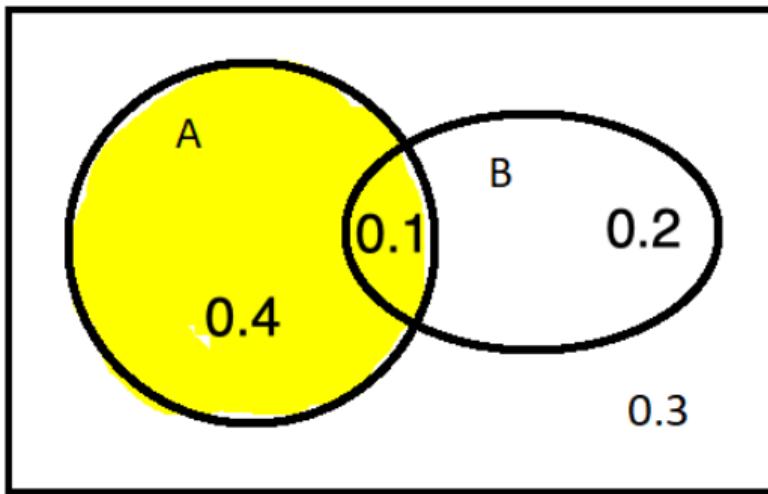
$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

- Sum** probabilities of all paths leading to an **event** to find its probability. The paths represent mutually exclusive outcomes.

$$P(B) = P(A \cap B) + P(A^c \cap B).$$

Exercise: Tree diagrams from Venn Diagrams

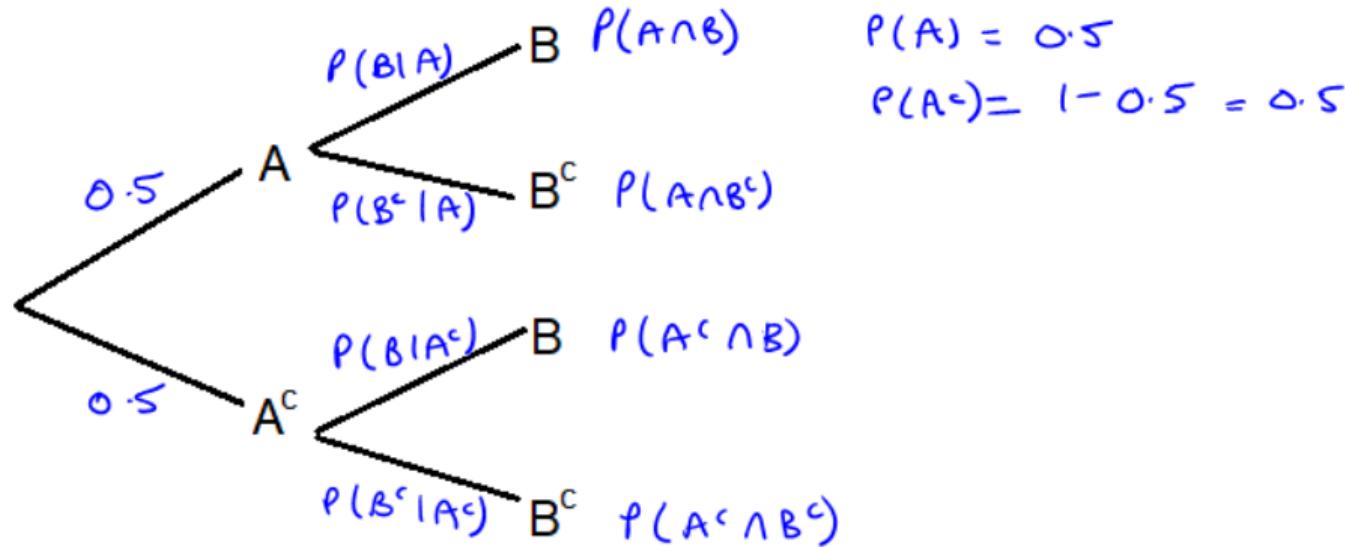
Example: Use the given Venn diagram to build a tree diagram and calculate all conditional probabilities branching on A first.



$$\begin{aligned}P(A) &= 0.4 + 0.1 \\&= 0.5\end{aligned}$$

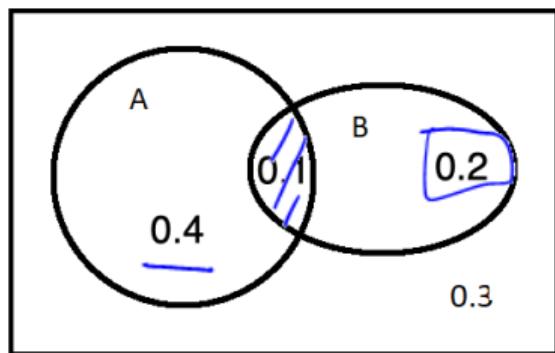
Exercise: Step 1

Step 1: Draw the tree diagram and fill in $P(A)$, $P(A^c)$ on first set of branches.



Exercise: Step 2

Step 2: The probabilities on the second set of branches are the conditional probabilities. But first we need the intersection probabilities. Use the Venn diagram:



$$P(A \cap B) = 0.1 \quad = P(B \wedge A)$$

$$P(A \cap B^c) = 0.4$$

$$P(A^c \cap B) = 0.2$$

$$P(A^c \cap B^c) = 0.3$$

Exercise: Step 3

Step 3: Calculate the conditional probabilities:

$$\begin{aligned} P(B|A) &= \frac{P(B \cap A)}{P(A)} \\ &= \frac{0.1}{0.5} = 0.2 \end{aligned}$$

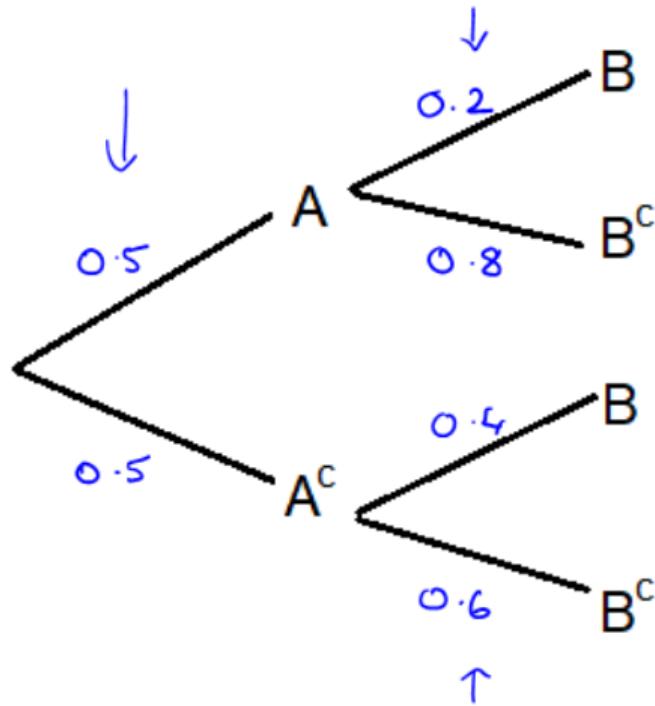
$$\begin{aligned} P(B^c|A) &= \frac{P(B^c \cap A)}{P(A)} \\ &= \frac{0.4}{0.5} = 0.8 \quad \checkmark \end{aligned}$$

$$\begin{aligned} P(B|A^c) &= \frac{P(B \cap A^c)}{P(A^c)} \\ &= \frac{0.2}{0.5} = 0.4 \quad \checkmark \end{aligned}$$

$$\begin{aligned} P(B^c|A^c) &= \frac{P(B^c \cap A^c)}{P(A^c)} \\ &= \frac{0.3}{0.5} = 0.6 \quad \checkmark \end{aligned}$$

Exercise: Step 4

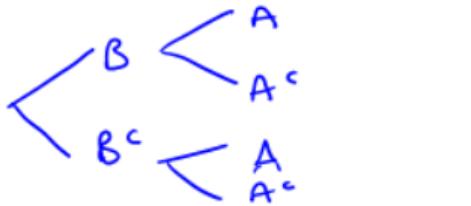
Step 4: Now complete the tree diagram.



Challenge

Try the following:

- 1 Repeat this process, branching on B first.
- 2 Consider how to get from the tree diagram to the corresponding two-way table.



	B	B^c	
A			
A^c			

Topic: Measuring Uncertainty with Probability

Independence

School of Mathematics and Applied Statistics



UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Introduction: Two scenarios

For certain pairs of events, the occurrence of one of them may or may not change the probability of the occurrence of the other.

- ➊ **Coin Tosses:** Flip a coin twice and think about the event of getting tails for Toss 1 and the event of Tails for Toss 2.

A

B

Introduction: Two scenarios

For certain pairs of events, the occurrence of one of them may or may not change the probability of the occurrence of the other.

- ➊ **Coin Tosses:** Flip a coin twice and think about the event of getting tails for Toss 1 and the event of Tails for Toss 2.
Successive coin tosses are not affected by previous results.

Introduction: Two scenarios

For certain pairs of events, the occurrence of one of them may or may not change the probability of the occurrence of the other.

- 1 **Coin Tosses:** Flip a coin twice and think about the event of getting tails for Toss 1 and the event of Tails for Toss 2.
Successive coin tosses are not affected by previous results.
- 2 **Drug Testing:** A drug test is much more likely to give a positive result if the drug is present, so the event ‘positive test result’ will be affected by whether or not the ‘drug is present’

Introduction: Two scenarios

For certain pairs of events, the occurrence of one of them may or may not change the probability of the occurrence of the other.

- 1 **Coin Tosses:** Flip a coin twice and think about the event of getting tails for Toss 1 and the event of Tails for Toss 2.
Successive coin tosses are not affected by previous results.
- 2 **Drug Testing:** A drug test is much more likely to give a positive result if the drug is present, so the event ‘positive test result’ will be affected by whether or not the ‘drug is present’

If the probability that A occurs is not affected by whether or not B occurs, we say that A and B are **independent** events.

Independence

Definition

Two events A and B are said to be **independent** if and only if

$$P(A \cap B) = P(A)P(B)$$

Independence

Definition

Two events A and B are said to be **independent** if and only if

$$P(A \cap B) = P(A)P(B)$$

or

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(A) \cdot P(B)}{P(B)} \end{aligned}$$

This means that

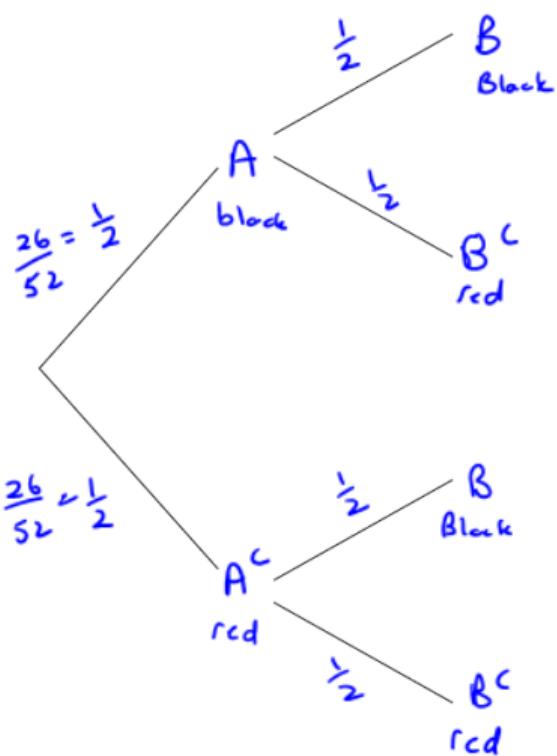
- $P(A|B) = P(A)$ and $P(B|A) = P(B)$
- Conditioning upon B does not alter the probability of A , and vice-versa.
- The knowledge that B has occurred gives no information as to whether A is more or less likely to occur, and vice-versa.

Independence - Exercise

Exercise

- ① A card is drawn from a pack of 52 cards. The card is returned, the pack is reshuffled, and a second card is drawn. Let $A = \{\text{first card is black}\}$, $B = \{\text{second card is black}\}.$
 - a. Draw a tree diagram.
 - b. Are A and B independent?
- ② Repeat Q1 if the first card is not returned to the pack before the 2nd card is drawn.

Exercise 1



$$P(A \cap B) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}. \quad \#$$

$$P(A^c \cap B) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}. \quad \#$$

$$P(B) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

Exercise 1 cont.

Are A & $A \cup B$ independent?

Check: $P(A \cap B) = P(A) \cdot P(B)$.

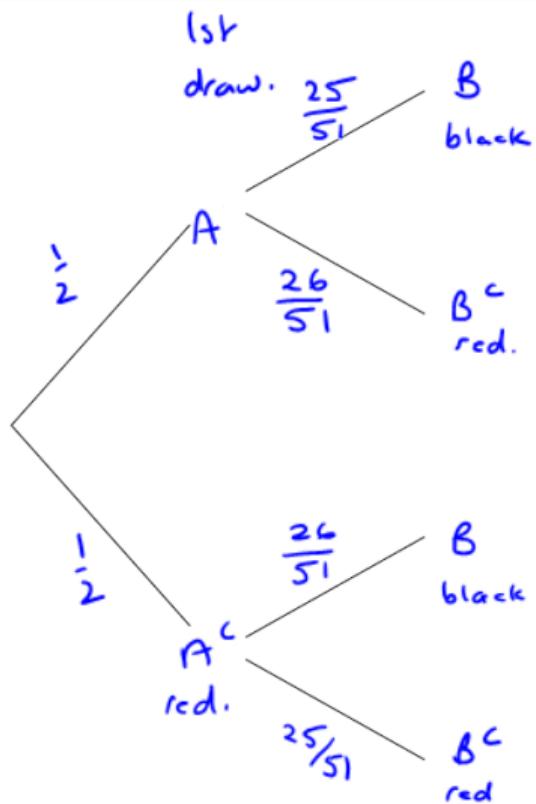
$$\begin{aligned} \text{LHS} &= P(A \cap B) \\ &= \frac{1}{4} \end{aligned}$$

$$P(B) = \frac{1}{2}.$$

$$\begin{aligned} \text{RHS} &= P(A) \cdot P(B) \\ &= \frac{1}{2} \times \frac{1}{2} \\ &= \frac{1}{4}. \end{aligned}$$

$\text{LHS} = \text{RHS} \Rightarrow A$ & B are
independent events.

Exercise 2



$$P(A \cap B) = \frac{1}{2} \times \frac{25}{51} = \frac{25}{102}.$$
*

+

$$P(A^c \cap B) = \frac{1}{2} \times \frac{26}{51} = \frac{26}{102}$$
*

$$P(B) = \frac{51}{102}$$

Exercise 2 cont.

$$\text{LHS} = P(A \cap B)$$

$$= \frac{25}{102} .$$

$$\text{RHS} = P(A) \cdot P(B)$$

$$= \frac{1}{2} \times \frac{51}{102}$$

$$= \frac{51}{204} .$$

$$= \frac{1}{4} .$$

$\text{LHS} \neq \text{RHS} \Rightarrow A \cap B \text{ are NOT indept}$
 $\Rightarrow \text{dependent} .$

Dependence

If two events are not independent, they are dependent.

Sometimes it is useful to think of this as

- mathematical
- or theoretical,
- or population
- or model-based

independence.

This contrasts with how statisticians use data from samples to investigate, and make inferences about,

- dependence
- independence

in populations.

Topic: Measuring Uncertainty with Probability

Law of Total Probability

School of Mathematics and Applied Statistics



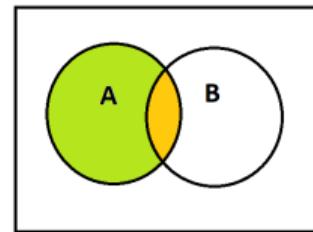
UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Rewriting $P(A)$

For any events A and B , we can rewrite $P(A)$ as:

$$P(A) = \underbrace{P(A \cap B)}_{\text{orange}} + \underbrace{P(A \cap B^c)}_{\text{green}} \quad (1)$$

We know from the Multiplicative rule that



$$P(A \cap B) = P(A|B)P(B)$$

$$P(A \cap B^c) = P(A|B^c)P(B^c)$$

So we can rewrite $P(A)$ from (1) as

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c) \quad (2) \quad \checkmark$$

We have **rewritten or separated** the event A into two parts:

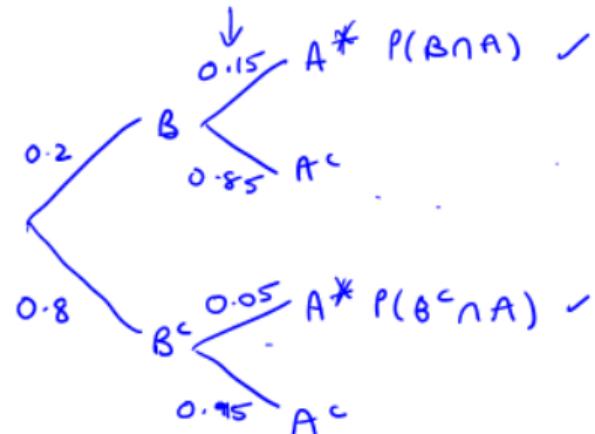
- the **part** which intersects with B , and
- the **part** which intersects with B^c

Example: finding $P(A)$

Example 1: Let $A =$ event a person has lung cancer, and $B =$ the event a person is a smoker, and given $P(A|B) = 0.15$, $P(A|B^c) = 0.05$, $P(B) = 0.2$, find the proportion of people with lung cancer.

$$P(B^c) = 0.8 \checkmark$$

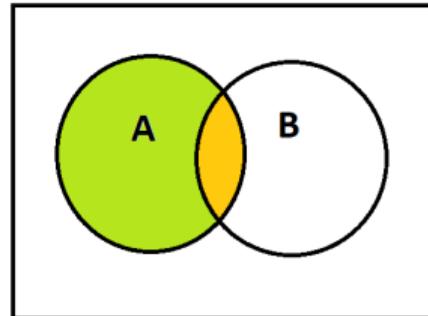
$$\begin{aligned} P(A) &= P(A \cap B) + P(A \cap B^c) \quad \checkmark \\ &= P(A|B)P(B) + P(A|B^c)P(B^c) \\ &= 0.15 \times 0.2 + 0.05 \times 0.8 \\ &= 0.03 + 0.04 \\ &= 0.07 \end{aligned}$$



Partitioning the Sample Space S

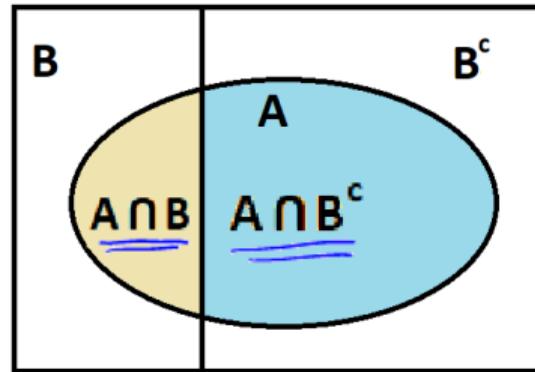
From (1) and (2) we have:

$$\begin{aligned} P(A) &= P(A \cap B) + P(A \cap B^c) \\ &= P(A|B)P(B) + P(A|B^c)P(B^c) \end{aligned}$$



Now consider that S is partitioned such that $S = B \cup B^c$ then we could redraw our diagram as:

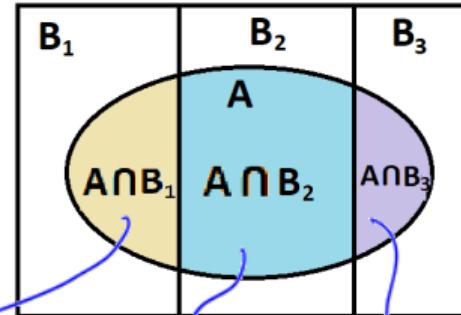
The result above for $P(A)$ still holds.



Partitioning the Sample Space S

Now consider that S is partitioned such that

- $S = B_1 \cup B_2 \cup B_3$ and
- $B_1 \cap B_2 = \emptyset$; and
 $B_1 \cap B_3 = \emptyset$; and
 $B_2 \cap B_3 = \emptyset$



Then A is decomposed as:

$$A = (A \cap B_1) \cup (A \cap B_2) \cup (A \cap B_3)$$

and $P(A)$ is given by

$$\begin{aligned} P(A) &= P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3) \\ &= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3) \end{aligned}$$

Law of Total Probability

More generally, we say that the collection of sets $\{B_1, \dots, B_k\}$ is said to be **a partition of S** when

- for some positive integer k ,

$$B_1 \cup B_2 \cup \dots \cup B_k = \bigcup_{i=1}^k B_i = S, \quad \stackrel{\text{overall union}}{=} \quad i = 1, \dots, k$$

- $\{B_1, \dots, B_k\}$ are non-overlapping such that $B_i \cap B_j = \emptyset$ for all $i \neq j$

In this case,

$$P(A) = \sum_{i=1}^k P(A \cap B_i) = \sum_{i=1}^k P(A|B_i)P(B_i) \quad (3)$$

This is called the **Law of Total Probability**

Topic: Measuring Uncertainty with Probability

Law of Total Probability - Exercise

School of Mathematics and Applied Statistics



UNIVERSITY
OF WOLLONGONG
AUSTRALIA



Probability Rules - Summary

If A and B represent any two events then,

Complement

$$P(\text{not } A) = P(A^c) = 1 - P(A)$$

Additive

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Conditional

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Multiplicative

$$P(A \cap B) = P(A|B)P(B)$$



Law of Total Probability $P(A) = \sum_{i=1}^k P(A \cap B_i) = \sum_{i=1}^k P(A|B_i)P(B_i)$

Exercise: Law of Total Probability

Exercise: In a certain factory, Machines 1, 2, and 3 are all producing springs of the same length. Machines 1, 2, and 3 produce 1%, 4% and 2% defective springs, respectively. Of the total production of springs in the factory, Machine 1 produces 30%, Machine 2 produces ~~25~~²⁵%, and Machine 3 produces 45%.

If one spring is selected at random from the total springs produced in a given day, determine the probability that it is defective.

Let D be the event that the spring is defective.

Let M_i be the event that a spring is produced Machine i ($i = 1, 2, 3$).

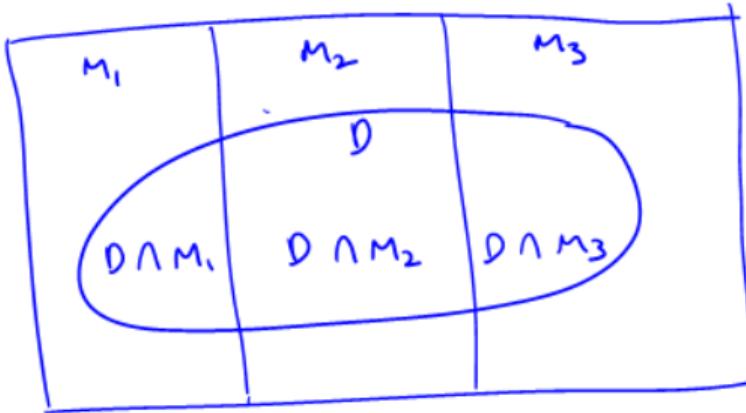
M1 M2 M3.

Ref: From Hogg, McKeon, Craig, (2013) Introduction to Mathematical Statistics. p28 Ex 1.4.8

Exercise cont.: Diagram

- Draw a Venn diagram to show the partitioning of S : How is the sample space partitioned in this context?

3 machines



Exercise cont.: Given information

- Write down known information using correct notation.

Machines 1, 2, and 3 produce 1%, 4% and 2% defective springs, respectively.

Machine 1 produces 30%, Machine 2 produces ~~30~~²⁵%, and Machine 3 produces 45%.

$$P(M_1) = 0.30$$

$$P(D|M_1) = 0.01$$

$$P(M_2) = 0.25$$

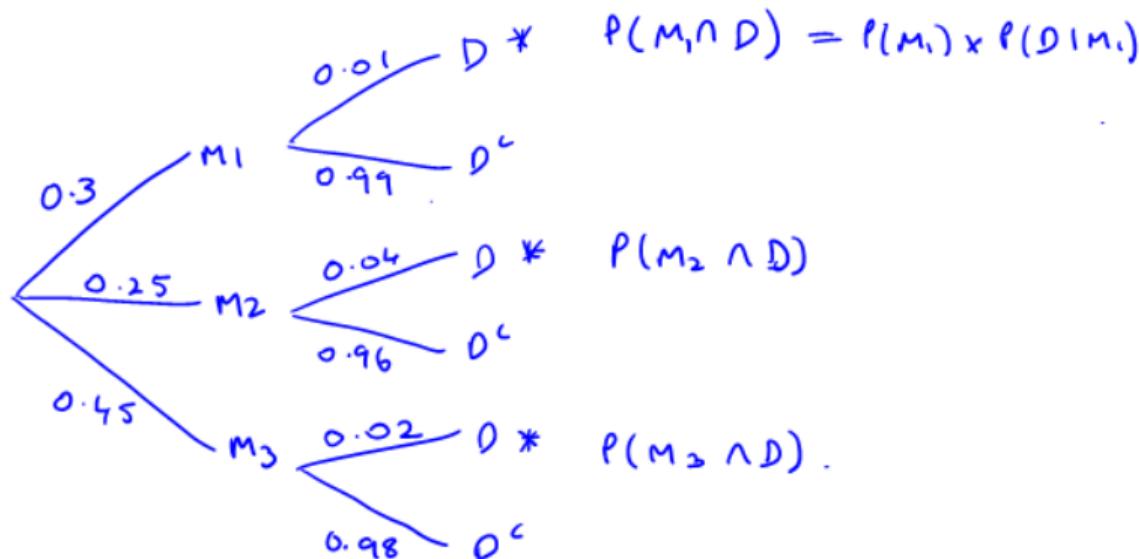
$$P(D|M_2) = 0.04$$

$$\begin{array}{r} P(M_3) = \underline{\underline{0.45}} \\ \hline 1.0 \end{array}$$

$$P(D|M_3) = 0.02$$

Exercise cont.: Tree diagram

- Draw a tree diagram showing all probabilities.



Exercise cont.: Apply the Law of Total Probability

- If one spring is selected at random from the total springs produced in a given day, determine the probability that it is defective.

$$\begin{aligned} P(D) &= P(D \cap M_1) + P(D \cap M_2) + P(D \cap M_3) \quad \# \\ &= P(D|M_1) \cdot P(M_1) + P(D|M_2) \cdot P(M_2) + P(D|M_3) \cdot P(M_3). \\ &= (0.01 \times 0.3) + (0.04 \times 0.25) + (0.02 \times 0.45) \\ &= 0.003 + 0.01 + 0.009 \quad \# \\ &= 0.022 \end{aligned}$$

2.2%



Exercise cont.: Apply the Law of Total Probability

$$P(M_1 \cap D) = 0.003$$

$$+ P(M_1 \cap D^c) = 0.3 \times 0.99 \\ = 0.297$$

$$P(M_2 \cap D) = 0.01$$

$$+ P(M_2 \cap D^c) = 0.25 \times 0.96 \\ = 0.24.$$

$$P(M_3 \cap D) = 0.009$$

$$+ P(M_3 \cap D^c) = 0.45 \times 0.98 \\ = \underline{0.441}.$$

$P(M_i)$

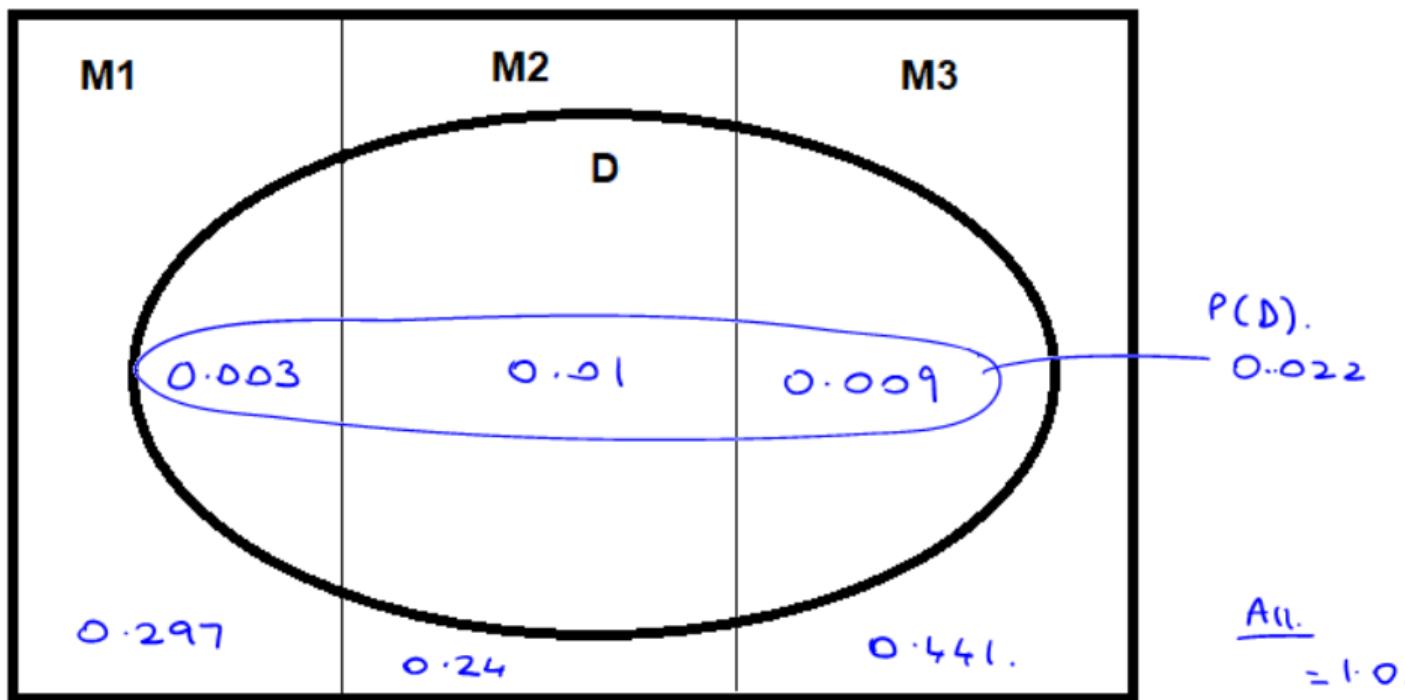
0.3

0.25

0.45

Exercise cont.: Apply the Law of Total Probability

- Complete the Venn diagram showing all probabilities.



Topic: Measuring Uncertainty with Probability

Bayes' Theorem

School of Mathematics and Applied Statistics



UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Recall: Conditional probability

We know

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

and similarly

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A \cap B)}{P(A)}$$

Rearranging

$$P(A \cap B) = P(A|B) \times P(B)$$

and

$$P(A \cap B) = P(B|A) \times P(A)$$

So equating these two expressions for $P(A \cap B)$ we can see

$$P(A|B)P(B) = P(B|A)P(A)$$

Then dividing both sides by $P(A)$ we get

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

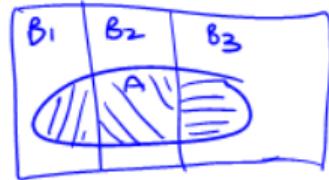
Eqn. (1)

Recall: Law of Total Probability

Recall: that the collection of sets $\{B_1, \dots, B_k\}$ is said to be **a partition of S** when

- for some positive integer k ,

$$B_1 \cup B_2 \cup \dots \cup B_k = \bigcup_{i=1}^k B_i = S, \quad i = 1, \dots, k$$



- where $\{B_1, \dots, B_k\}$ are non-overlapping such that $B_i \cap B_j = \emptyset$ for all $i \neq j$

The **Law of Total Probability**:

$$P(A) = \sum_{i=1}^k P(A \cap B_i) = \sum_{i=1}^k P(A|B_i)P(B_i) \quad (2)$$

✓ -

Recall: Conditional probability

So let's apply the Law of Total Probability if we have two partitions: B and B^c :

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$$

$P(A \cap B) + P(A \cap B^c)$



Then substitute for $P(A)$ into Eqn(1): we obtain **Bayes' Theorem**

$$\underline{P(B|A)} = \frac{\cancel{P(A|B)}P(B)}{\cancel{P(A|B)}P(B) + \cancel{P(A|B^c)}P(B^c)} \leftarrow P(A)$$

Thus, if we know $P(A|B)$, we can determine the $P(B|A)$

Bayes' Theorem cont.

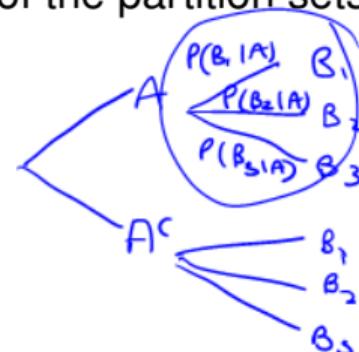
This result can be generalised: If the sets B_1, B_2, \dots constitute a partition of S , then **Bayes' Theorem** may be written as

$$\underline{P(B_i|A)} = \frac{P(A|B_i)P(B_i)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \dots + \cancel{P(A|B_k)P(B_k)}} \leftarrow \text{P(A)}$$

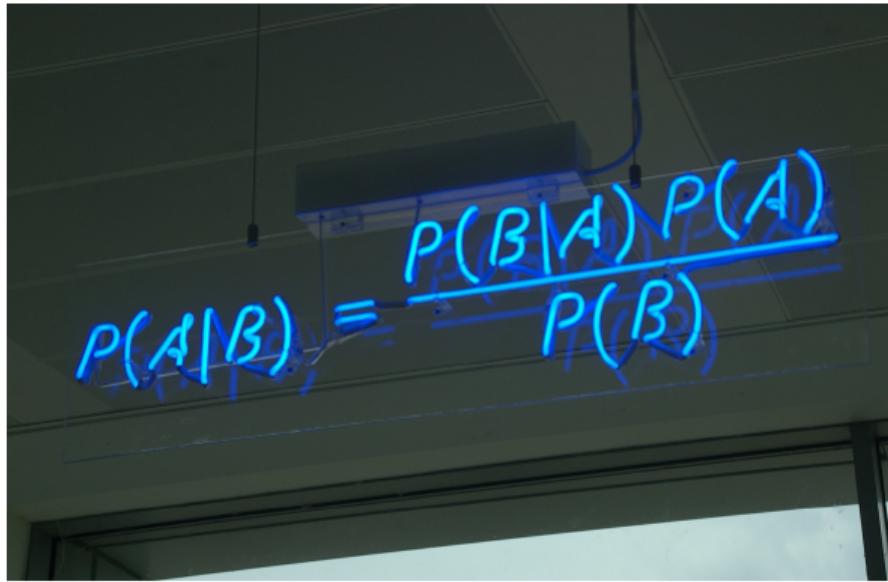
The denominator uses the **Law of Total Probability** from Eqn2

Notice that the sum of these conditional probabilities of the partition sets, given A , is 1.

$$\sum_{i=1}^k P(B_i|A) = 1$$



Bayes' Theorem in Neon Lights!



Bayes' Theorem has many applications in computer science.

Video: Prof Sahami from Stanford University <http://www.youtube.com/watch?v=MSIoBqvTK0Y>

Bayes' Rule

Bayes' Theorem may be conveniently presented in table

$P(B_i)$	$P(A B_i)$	$P(B_i)P(A B_i) = P(B_i \cap A)$	$\frac{P(B_i)P(A B_i)}{P(A)} = P(B_i A)$
x	y	$x \times y$	$\frac{x \times y}{P(A)}$
1		$\sum_i P(B_i)P(A B_i) = P(A)$	1

- The first two columns usually given information
- The 3rd column is product of first two, and the sum is $P(A)$
- The 4th column is the 3rd divided by $P(A)$
these are by Bayes' theorem $P(B_i|A)$

Topic: Measuring Uncertainty with Probability

Bayes' Theorem - Exercise

School of Mathematics and Applied Statistics



UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Probability Rules - Summary

If A and B represent any two events then,

Conditional

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Multiplicative

$$P(A \cap B) = P(A|B)P(B)$$

Law of Total Probability $P(A) = \sum_{i=1}^k P(A \cap B_i) = \sum_{i=1}^k P(A|B_i)P(B_i)$

Bayes' Theorem

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \dots + P(A|B_k)P(B_k)}$$

Exercise.: Bayes' Theorem

In a certain factory, Machines 1, 2, and 3 are all producing springs of the same length. Machines 1, 2, and 3 produce 1%, 4% and 2% defective springs, respectively. Of the total production of springs in the factory, Machine 1 produces 30%, Machine 2 produces 25%, and Machine 3 produces 45%.

Let D be the event that the spring is defective.

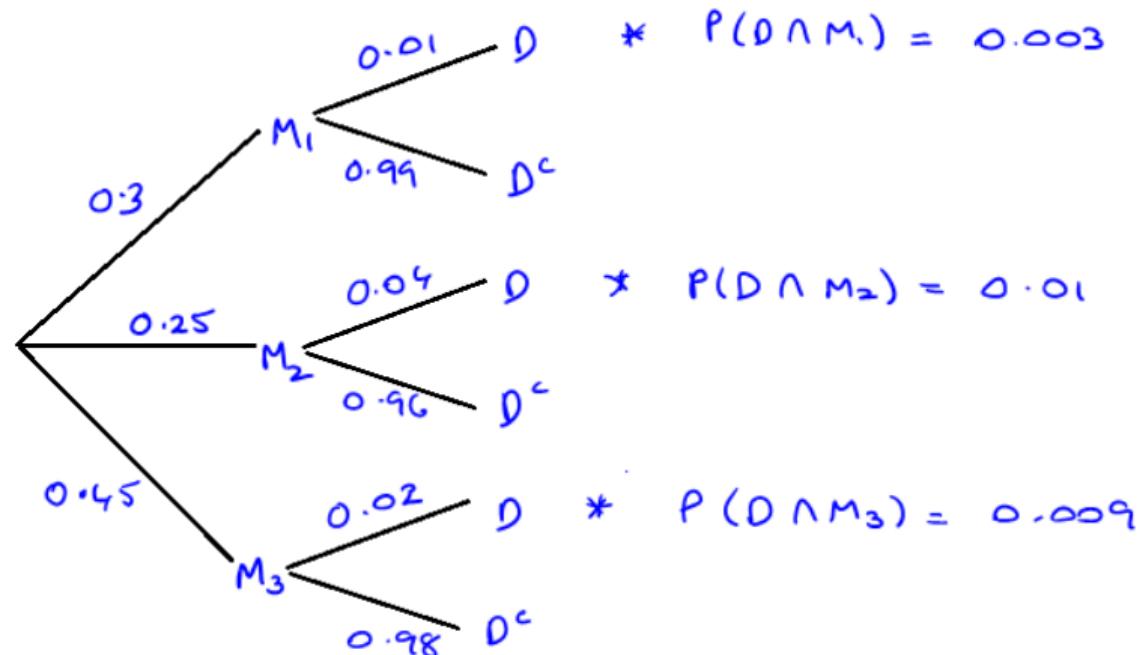
Let M_i be the event that a spring is produced Machine ($i = 1, 2, 3$). M_1, M_2, M_3 .

- a. If one spring is selected at random from the total springs produced in a given day, determine the probability that it is defective.
- b. Given that the ~~defective~~ spring is defective, find the conditional probability that it was produced by Machine 2.
- c. Further, determine $P(M_1|D)$ and $P(M_3|D)$ and demonstrate that $\sum_{i=1}^k P(M_i|D) = 1$.

Ref: From Hogg, McKeon, Craig, (2013) Introduction to Mathematical Statistics. p28 Ex 1.4.8

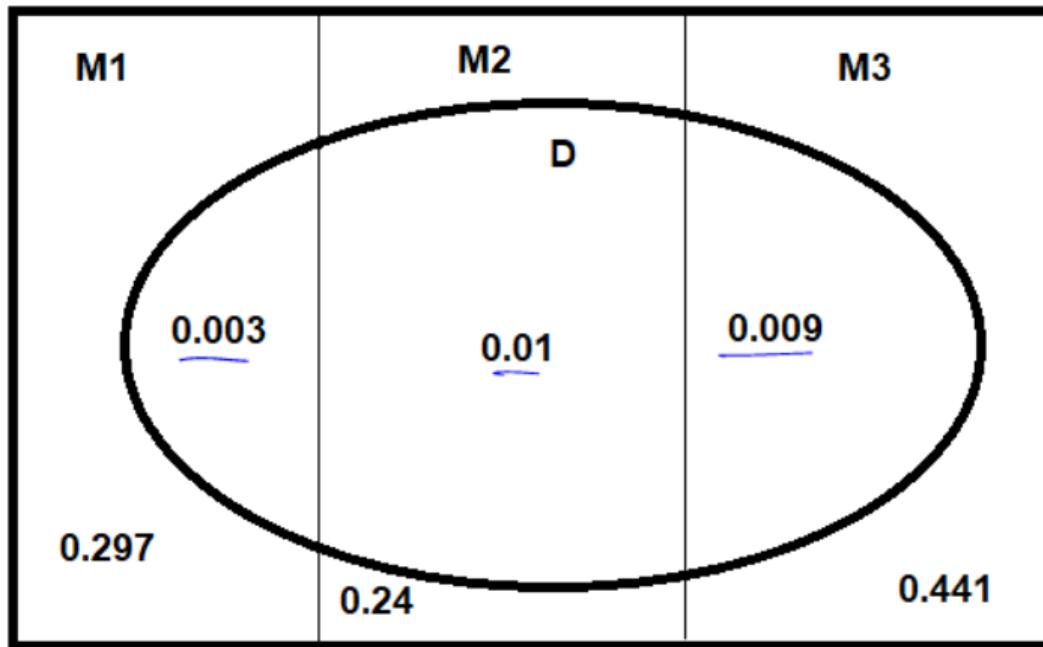
Exercise cont.: Apply the Law of Total Probability

a. Draw the tree diagram



Exercise cont.: Apply the Law of Total Probability

The Venn diagram showing all probabilities (previously determined).



$$\begin{aligned}P(D) &= 0.003 + \\&0.010 \\&0.009 \\&\hline \\&\underline{0.022} \\&\text{or } 2.2\%.\end{aligned}$$

Exercise cont.: Apply Bayes' Theorem

- b. Given that the ~~defective~~ spring is defective, find the conditional probability that it was produced by Machine 2.

$$\begin{aligned}
 P(M_2 | D) &= \frac{P(D|M_2) P(M_2)}{P(D|M_1) P(M_1) + P(D|M_2) P(M_2) + P(D|M_3) P(M_3)} \leftarrow P(D) \\
 &= \frac{0.04 \times 0.25}{0.022} \leftarrow (a). \\
 &= \frac{0.010}{0.022} \\
 &= \frac{10}{22} = \frac{5}{11}.
 \end{aligned}$$

Exercise cont.: Apply Bayes' Theorem

- c. Determine $P(M_1|D)$ and $P(M_3|D)$ and demonstrate that $\sum_{i=1}^3 P(M_i|D) = 1$.

$$\begin{aligned} P(M_1|D) &= \frac{P(M_1 \wedge D)}{P(D)} \\ &= \frac{0.003}{0.022} \\ &= \frac{3}{22}. \end{aligned}$$

$$\begin{aligned} P(M_3|D) &= \frac{P(M_3 \wedge D)}{P(D)} \\ &= \frac{0.009}{0.022} \\ &= \frac{9}{22}. \end{aligned}$$

Exercise cont.: Apply Bayes' Theorem

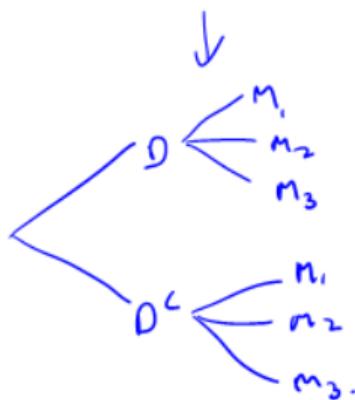
$$\sum_{i=1}^3 P(M_i | D) = P(M_1 | D) + P(M_2 | D) + P(M_3 | D)$$

$\xrightarrow{\quad}$

$$= \frac{3}{22} + \frac{10}{22} + \frac{9}{22}$$

$$= \frac{22}{22}$$

$$= 1 \quad \checkmark \text{ required.}$$



Correlation and Regression



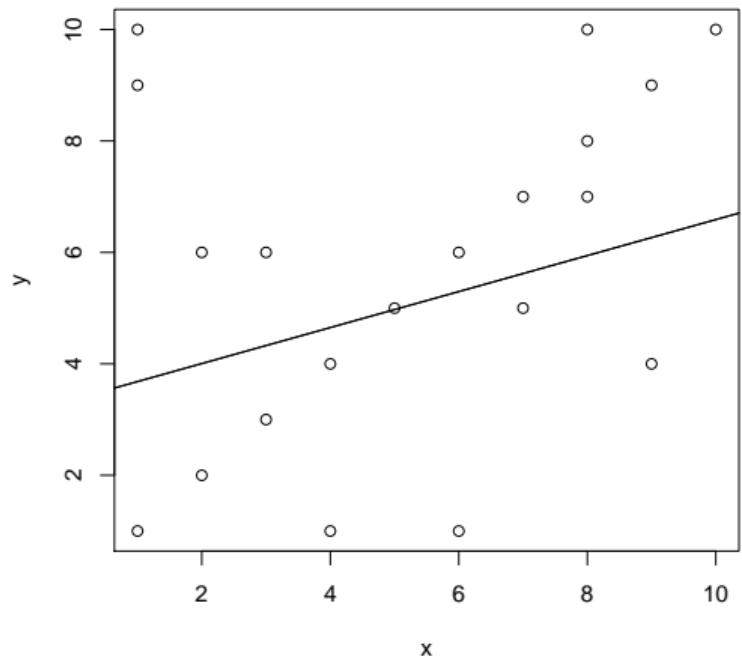
School of Mathematics and Applied Statistics
Faculty of Engineering and Information Sciences

Scatterplot

A scatterplot is a point graph of two quantitative variables, x and y .

i.e. x is used to predict y .

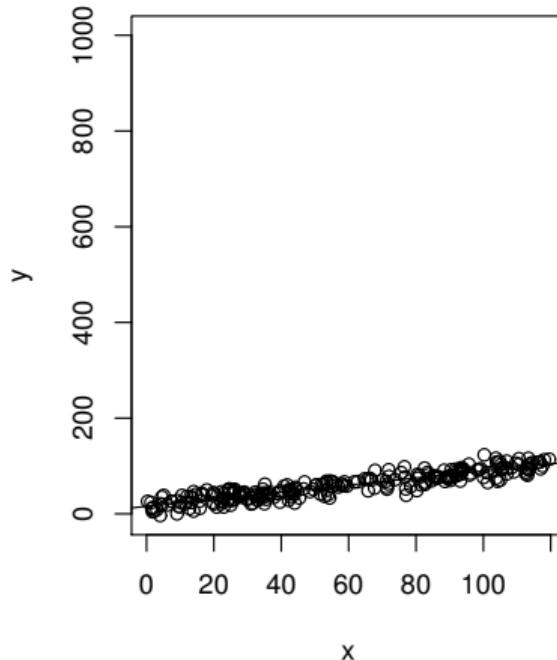
Scatterplot



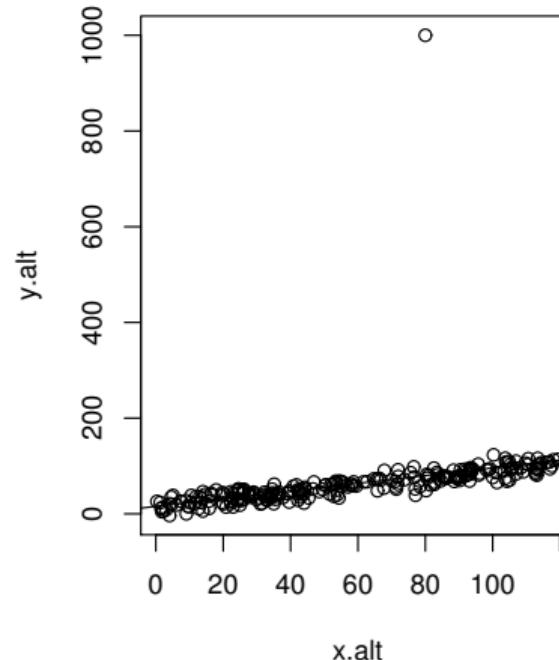
What is an outlier?

An outlier is an observation which differs substantially from the main trend of the data.

no outlier

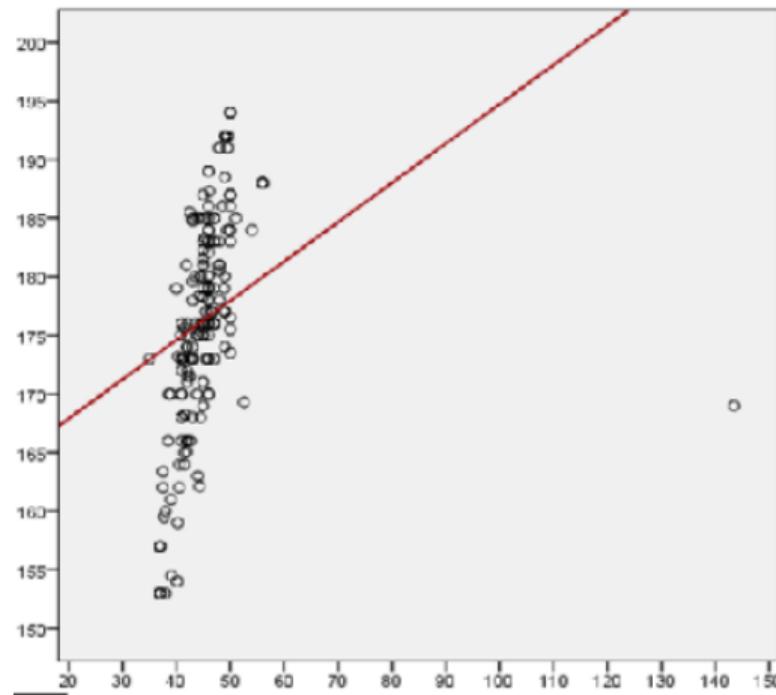


with outlier



What is an influential point?

Every data point affects the line of best fit; an **influential observation** does so more than other points.



Scatterplot

A scatterplot reveals:

- positive/negative and weak/strong association,
- linear or curved relationship,
- outliers,
- influential points that impact the best-fit line,
- clusters and gaps.

Scatterplots

- A relationship between two quantitative variables can be displayed as a *scatterplot*.
- R code: `plot(x,y)`
- Plotting symbols (R `plot()` optional argument: `pch`) and colours (argument: `col`).

Scatterplots: What to look for?

- The direction of the relationship between two variables x and y . Positive? Negative? or No relationship?
- The variability of the y for different value of x . For example, variations of house price in big house is much bigger than in small house.
- Unusual data points/outliers

Classic Example: Anscombe's Quartet

Four bivariate (x, y) datasets with 11 observations each.

obs	X1	Y1	X2	Y2	X3	Y3	X4	Y4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.10	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.10	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89

Four bivariate (x, y) datasets with 11 observations each, *all* having the following properties:

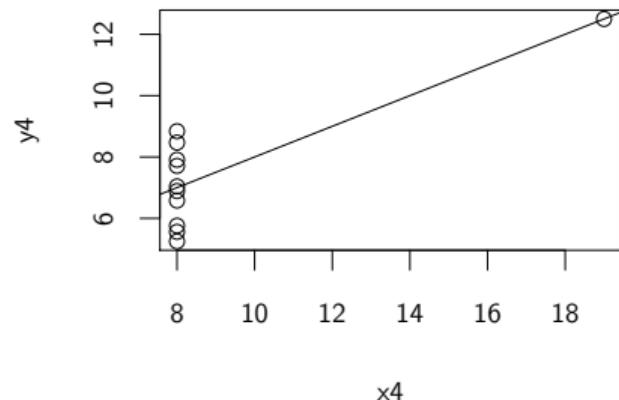
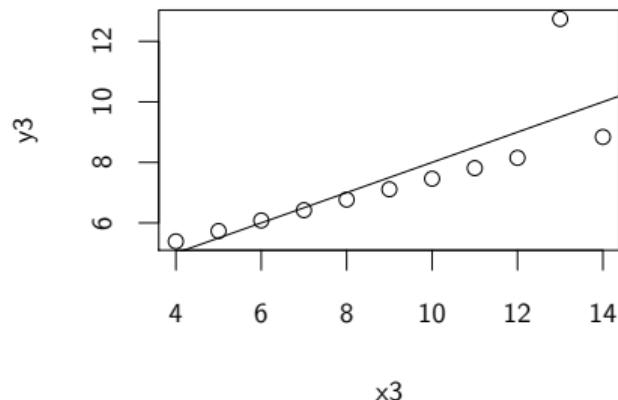
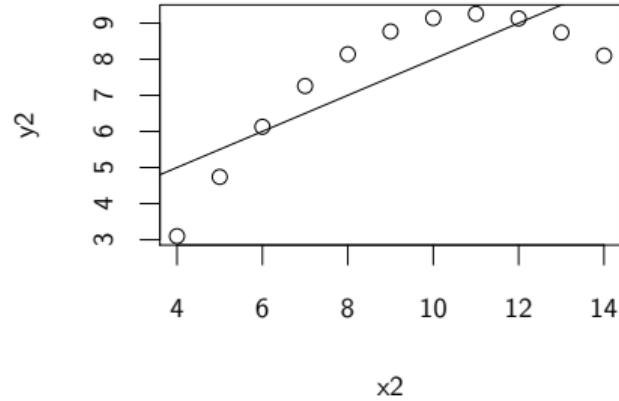
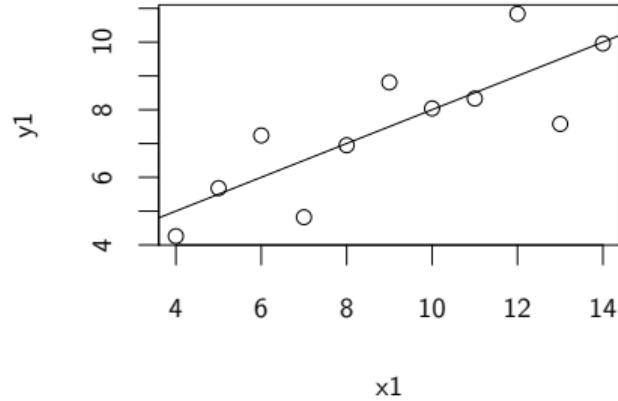
- $\bar{x} = 9$
- $s_x = 3.3166$
- $\bar{y} = 7.5009$
- $s_y = 2.0316$
- $r = 0.8164$

Four bivariate (x, y) datasets with 11 observations each, *all* having the following properties:

- $\bar{x} = 9$
- $s_x = 3.3166$
- $\bar{y} = 7.5009$
- $s_y = 2.0316$
- $r = 0.8164$

How different can they be?

Very different!



Scatterplot Matrix

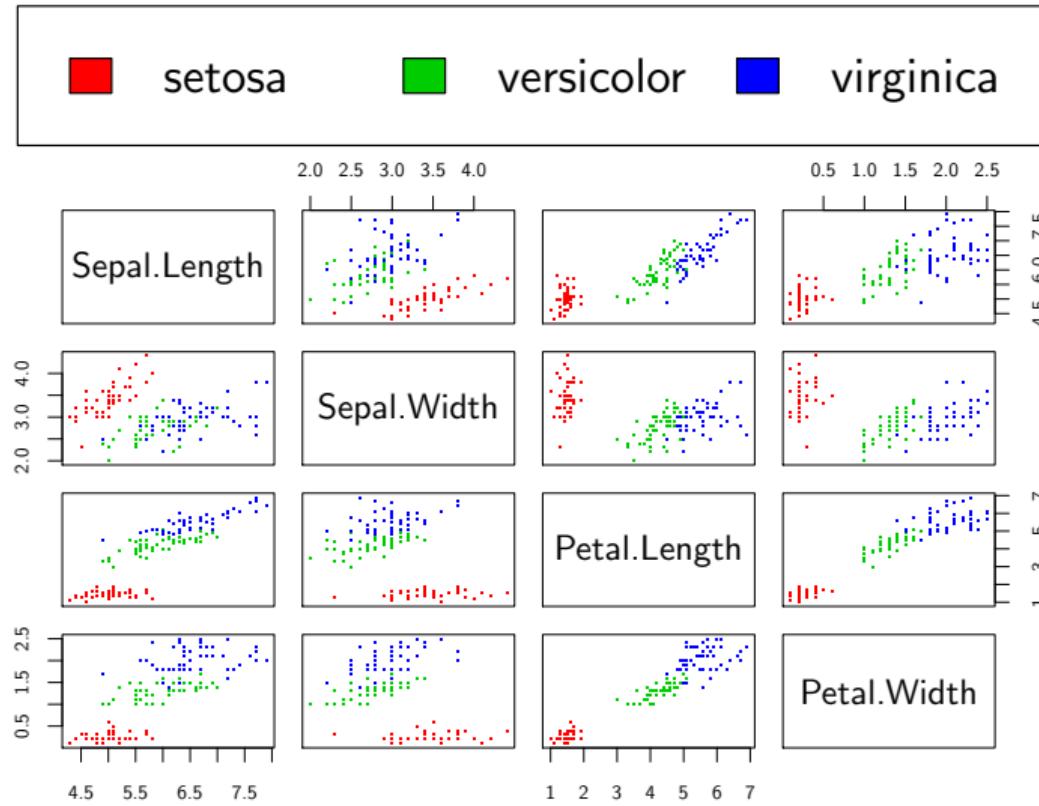
- A *scatter plot matrix* is an array of scatterplots with aligned axes, with each attribute plotted against each other attribute (except itself).
- R code: `pairs(x)` (x has multiple columns)

Example Dataset: Anderson's Iris Data

Measurements on 150 flowers from three species of iris (50 each):

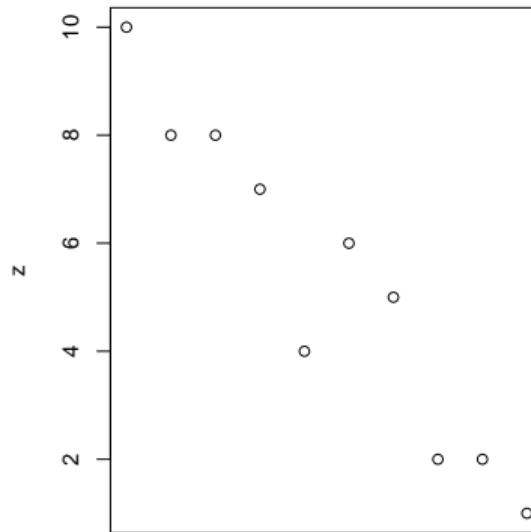
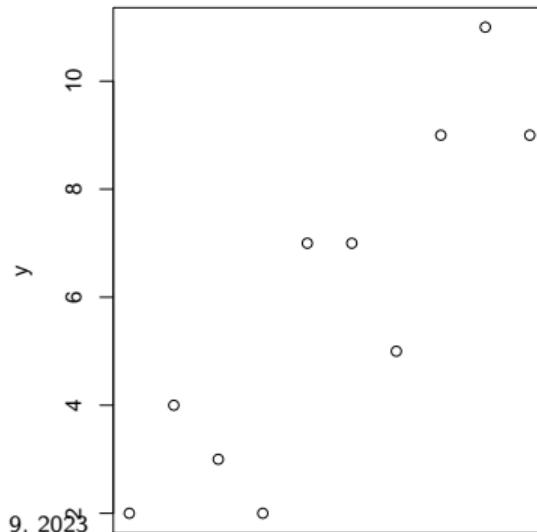
Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
5.8	4.0	1.2	0.2	setosa
5.2	3.5	1.5	0.2	setosa
4.5	2.3	1.3	0.3	setosa
6.5	2.8	4.6	1.5	versicolor

Scatterplot Matrix



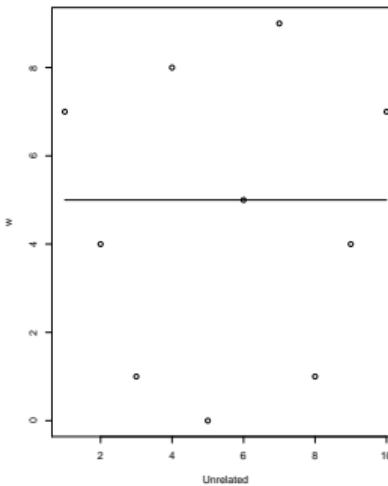
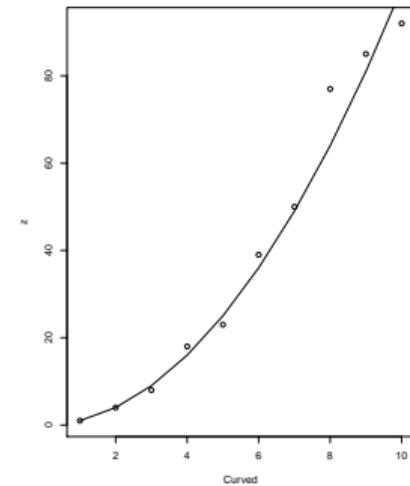
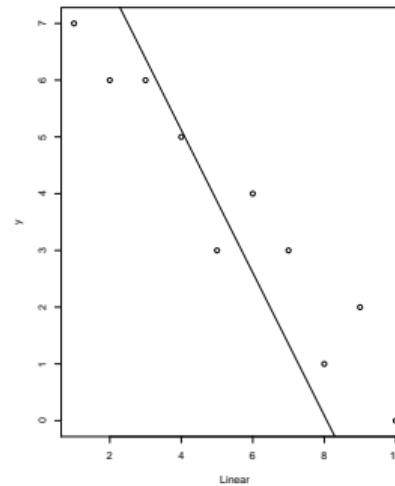
Association

Variables are said to have *positive association* when high values of one result in high values of the other. They have *negative association* if high values in one result in low values of the other.



Best-fit Line

The relationship is said to be *linear* if a straight line is a good approximation of most points. It is *curved (nonlinear)* if a curved line is a good approximation. If the variables are *unrelated*, the best-fit line is horizontal (but not necessarily vice-versa).



Correlation

The statistical correlation between two quantitative variables is a number that indicates the strength and direction of a straight-line relationship.

The **strength** of the relationship is determined by the closeness of the points to a straight line.

- can be **strong** or **weak**.

The **direction** can be **positive** or **negative**.

- Two variables are **positively** correlated if an increase in one is associated to an increase in the other.
- Two variables are **negatively** correlated if an increase in one is associated to a decrease in the other.

Correlation

The correlation coefficient measures strength and direction of linear association.

Correlation

The correlation coefficient measures strength and direction of linear association.

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

R code:

`cor(x,y)`

Correlation

The correlation coefficient measures strength and direction of linear association.

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

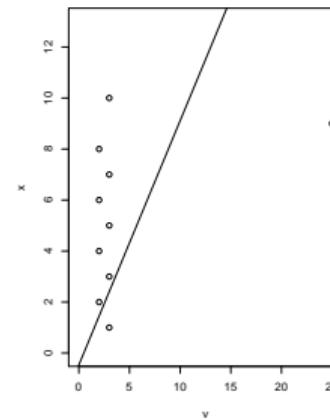
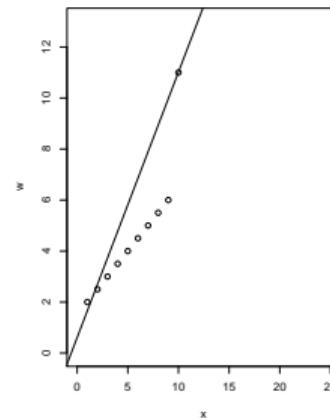
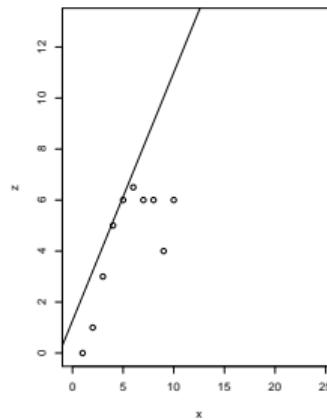
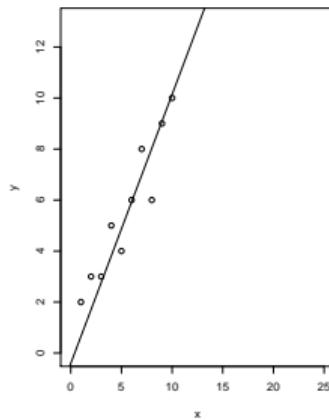
R code:

`cor(x,y)`

- $-1 \leq r \leq 1$
- The sign indicates positive/negative correlation.
- $r \approx 0$ indicates no association. $r = \pm 1$ is a perfect linear association.

Correlation

Be careful, r alone does not tell the whole story.



All these best-fit lines are the same, but the data sets are clearly very different. Figure 3 has a highly influential point and Figure 4 has an outlier.

Motivating Example

- Regression analysis is used to study the relationship between a response variable (Y) and one or more explanatory variables (X_1, X_2, \dots).
 - Do people with higher income (x) spend more on food (y)? Or less?
 - Do people with higher education (x) receive higher income (y)? Or less?
- Predict value of a response variable (y) given the value of explanatory variables (X_1, X_2, \dots).
 - Given the house size (x), how do we predict the expected house price (y)?

Introduction: Regression model

- Let Y be a response variable and X a predictor.
- If the scatter plot of Y vs X indicates that a linear function is appropriate for describing the relationship between Y and X

$$\Rightarrow Y = \beta_0 + \beta_1 X + e \quad (1)$$

can be used to describe the relationship, where

- (i) β_0 and β_1 := unknown parameters
- (ii) $Y - (\beta_0 + \beta_1 X) = e$ = the statistical error.

(1) is called a simple linear regression model.

(Simple) Linear Regression

- The simple linear regression model:

$$Y = \beta_0 + \beta_1 X + e$$

- The **random error** e represents the **unexplained** part $e = Y - (\beta_0 + \beta_1 X)$.
 - If the true relationship between Y and X is linear, e can capture any unexplained variation in Y .
 - e can be the effects of other variables not included in the model
 - e can be the effects of non-linearity in the relationship between y and x .
- There are **two parameters** β_0 and β_1 that need to be estimated given the **sample data**.

Denote the i^{th} observation by (y_i, x_i) , $i = 1, 2, \dots, n$.

If model (1) is appropriate for $\{(y_i, x_i)\}_{i=1,2,\dots,n}$

\Rightarrow we have the following expression:

$$y_1 = \beta_0 + \beta_1 x_1 + e_1$$

$$y_2 = \beta_0 + \beta_1 x_2 + e_2$$

⋮

$$y_n = \beta_0 + \beta_1 x_n + e_n$$

where $\{e_i\}$ **are random errors** (or noise).

Assumptions

$\{e_i\}$ needs to satisfy the following conditions

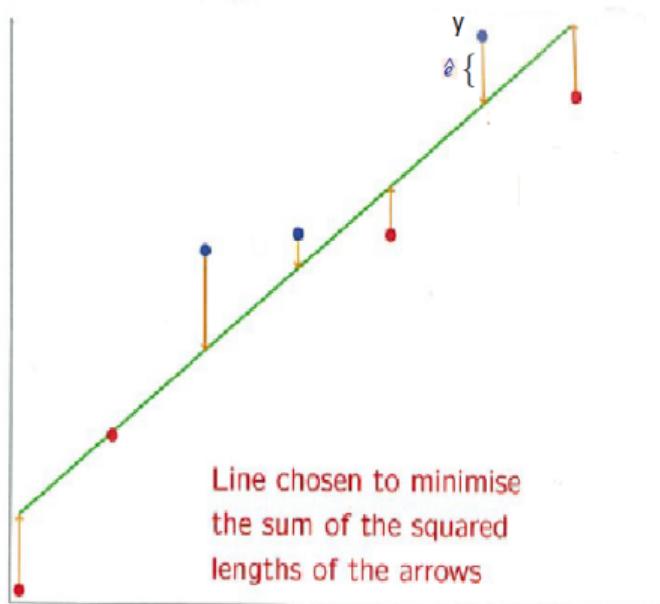
$$E(e_i) = 0$$

$$\text{Var}(e_i) = \sigma^2 \text{ a constant}$$

$$\text{Cov}(e_i, e_j) = 0 \text{ for } i \neq j$$

We are interested in the estimations of the unknown parameters of the model.

Least Square Method



- Least squares method choose a line to minimise sum of square residuals.

Least Squares Estimators

The estimates of β_0 and β_1 should minimize the sum of residuals $\sum \hat{e}_i^2$, where

$$\hat{e}_i = y_i - \hat{y}_i = y_i - (\beta_0 + \beta_1 x_i).$$

Consider the residual sum of squares

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2.$$

If there exist $\hat{\beta}_0$ and $\hat{\beta}_1$ such that

$$RSS(\hat{\beta}_0, \hat{\beta}_1) = \min_{\beta_0, \beta_1} RSS(\beta_0, \beta_1) = \min_{\beta_0, \beta_1} \left\{ \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2 \right\},$$

$\Rightarrow \hat{\beta}_0$ and $\hat{\beta}_1$ are the **least square estimators** of β_0 and β_1 .

Some definitions

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are called the least squares estimators of β_0 and β_1 .
- The **regression** (or prediction) **equation** is defined as

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- The value of \hat{y} given by the i^{th} observation x_i is

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

and is called the i^{th} **fitted value**.

- The **residual** for the i^{th} observation case is defined as

$$\hat{e}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

Influence of β_0 and β_1

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

- If $\hat{\beta}_1 < 0$, then $x \uparrow \implies y \downarrow$
- If $\hat{\beta}_1 > 0$, then $x \uparrow \implies y \uparrow$
- If $\hat{\beta}_1 = 0$, x is independent of y

Example

- The materials are taken from “An Introduction to Statistical Learning, with Application in R (G. James, D. Witten, T. Hastie, and R. Tibshirani (2013))”.
- The library() function is used to load libraries or groups of functions and data sets that are not included in the base R distribution.
- Here we load MASS package, which is a very large collection of data sets and functions.

```
> library(MASS)
```

Example

- The MASS library contains the Boston data set, which records medv (median house value) for 506 neighborhoods around Boston.
- We want to predict medv using 13 predictors such as rm (average number of rooms per house), age (proportion of owner occupied units), and lstat (percent of households with low socioeconomic status).

```
> fix(Boston)
```

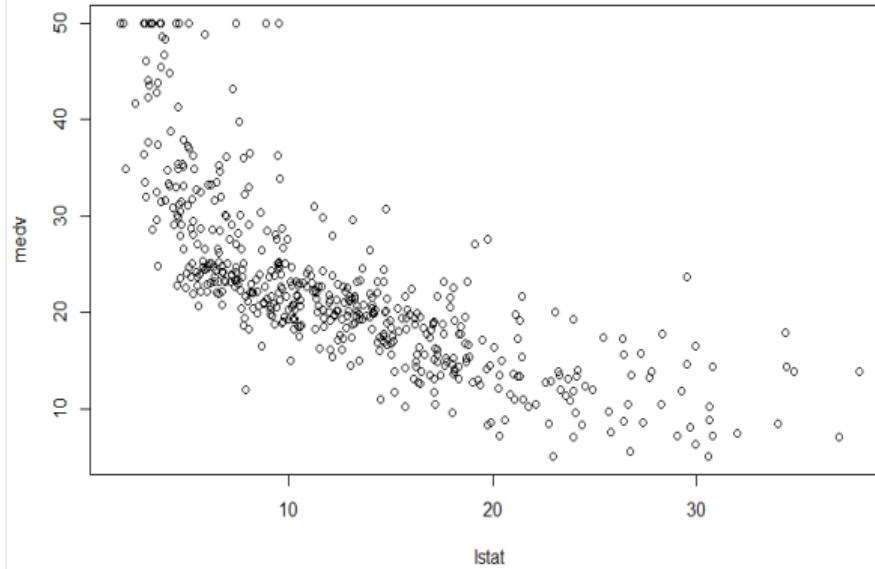
Motivating Example

Data Editor

File Edit Help

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
1	0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	396.9	4.98	24
2	0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.9	9.14	21.6
3	0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
4	0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
5	0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.9	5.33	36.2
6	0.02985	0	2.18	0	0.458	6.43	58.7	6.0622	3	222	18.7	394.12	5.21	28.7
7	0.08829	12.5	7.87	0	0.524	6.012	66.6	5.5605	5	311	15.2	395.6	12.43	22.9
8	0.14455	12.5	7.87	0	0.524	6.172	96.1	5.9505	5	311	15.2	396.9	19.15	27.1
9	0.21124	12.5	7.87	0	0.524	5.631	100	6.0821	5	311	15.2	386.63	29.93	16.5
10	0.17004	12.5	7.87	0	0.524	6.004	85.9	6.5921	5	311	15.2	386.71	17.1	18.9
11	0.22489	12.5	7.87	0	0.524	6.377	94.3	6.3467	5	311	15.2	392.52	20.45	15
12	0.11747	12.5	7.87	0	0.524	6.009	82.9	6.2267	5	311	15.2	396.9	13.27	18.9
13	0.09378	12.5	7.87	0	0.524	5.889	39	5.4509	5	311	15.2	390.5	15.71	21.7
14	0.62976	0	8.14	0	0.538	5.949	61.8	4.7075	4	307	21	396.9	8.26	20.4
15	0.63796	0	8.14	0	0.538	6.096	84.5	4.4619	4	307	21	380.02	10.26	18.2
16	0.62739	0	8.14	0	0.538	5.834	56.5	4.4986	4	307	21	395.62	8.47	19.9
17	1.05393	0	8.14	0	0.538	5.935	29.3	4.4986	4	307	21	386.85	6.58	23.1
18	0.7842	0	8.14	0	0.538	5.99	81.7	4.2579	4	307	21	386.75	14.67	17.5
19	0.80271	0	8.14	0	0.538	5.456	36.6	3.7965	4	307	21	288.99	11.69	20.2

Motivating Example



Motivating Example

- We will start using the `lm()` function to fit a simple linear regression model, with `medv` as the response and the `lstat` as the predictor.
- The basic syntax is `lm(y~x,data)`, where `y` is the response, `x` is the predictor, and `data` is the data set.

```
> lm.fit=lm(medv~lstat,data=Boston)
```

Motivating Example

- We use `summary(lm.fit)` to give us the coefficients, and other statistics.

```
> summary(lm.fit)
```

lm function

```
## Coefficients:  
##                               Estimate     Std. Error      t value    Pr(>|t|)  
## (Intercept)            34.55384     0.56263     61.41 <2e-16 ***  
## x                      -0.95005     0.03873    -24.53 <2e-16 ***
```

Motivating Example

- The estimated model:

$$\widehat{Medv} = 34.55 - 0.95LSTAT$$

- The meaning of the **slope** coefficient is that for every extra one percent of households with low economic status , the median house prices are **expected to decrease** by around \$950.
- The house in the neighborhood with more households with low economic status are expected to cost less.

Prediction

- The fitted regression equation

$$\widehat{Medv} = 34.55 - 0.95LSTAT$$

- We can use this to **predict** the expected median price for a neighborhood with 10% households with low economic status

$$\widehat{Medv} = 34.55 - 0.95(10) = 25.05(\text{in\$'000}).$$

```
> predict(lm.fit,data.frame(lstat=c(10)))
```

1

25.05335

Assumptions (Reminder)

$\{e_i\}$ needs to satisfy the following conditions

$$E(e_i) = 0$$

$$Var(e_i) = \sigma^2 \text{ a constant}$$

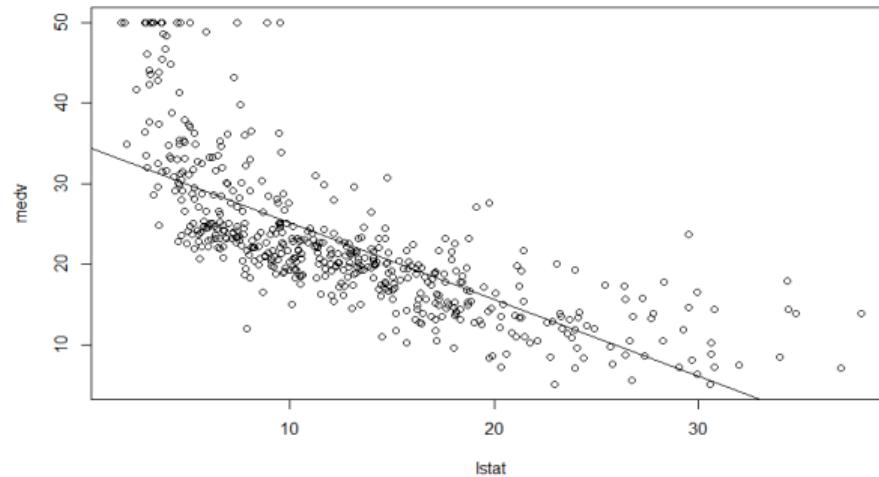
$$Cov(e_i, e_j) = 0 \text{ for } i \neq j$$

Diagnostics

- The residuals from a linear regression fit can be computed using the `residuals()` function.

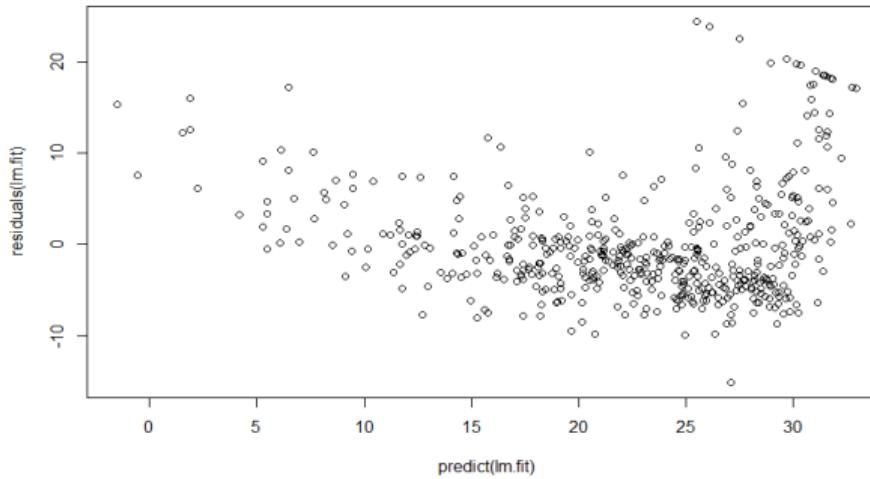
```
> plot(predict(lm.fit), residuals(lm.fit))
```

Motivating Example



There is some evidence for non-linearity in the relationship between lstat and medv.

Diagnostics



There is some evidence for non-linearity in the residual plot.

Multiple Linear Regression

- Simple linear regression is a useful approach for predicting a response on the basis of a single predictor variable.
- However, in practice, we often have more than one predictor.
- How can we include more predictors in the model?

Multiple Linear Regression

Given

- 1 response var.: Y
- p independent (predictors or explanatory) var.: X_1, \dots, X_p
- a sample with size n is observed

These observations may be denoted as follows:

obs.	Y	X_1	X_2	\dots	X_p
1	y_1	$x_{1,1}$	$x_{1,2}$	\dots	$x_{1,p}$
2	y_2	$x_{2,1}$	$x_{2,2}$	\dots	$x_{2,p}$
\vdots					
n	y_n	$x_{n,1}$	$x_{n,2}$	\dots	$x_{n,p}$

Multiple Linear Regression

A multiple linear regression model has the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + e$$

Matrix notation

We usually represent a multiple linear regression model in matrix form:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

where

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

Matrix notation

We usually represent a multiple linear regression model in matrix form:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

where

$$\mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ 1 & x_{2,1} & \cdots & x_{2,p} \\ \vdots & & & \\ 1 & x_{n,1} & \cdots & x_{n,p} \end{pmatrix} := \text{the design matrix}$$

Multiple Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + e \quad (2)$$

- Y : the response variable or the dependent variable to be predicted
- X_1, \dots, X_p : Predictors or covariates or independent variables
- The intercept β_0 and the slope coefficients β_1, \dots, β_p are unknown parameters to be estimated.

Multiple Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \epsilon \quad (3)$$

- We interpret the β_j as the marginal effect on Y of a one unit increase in X_j , holding all other predictors fixed and constant.

Boston Example

- To fit a multiple linear regression using least square method in R, we again use the `lm()` function.
- The syntax `lm(y~x1+x2+x3, data)` is used to fit a model with three predictors x_1 , x_2 , and x_3 .
- The `summary` function outputs all the regression coefficients for all the predictors.

```
lm.fit<-lm(medv~lstat+age,data=Boston)  
summary(lm.fit)
```

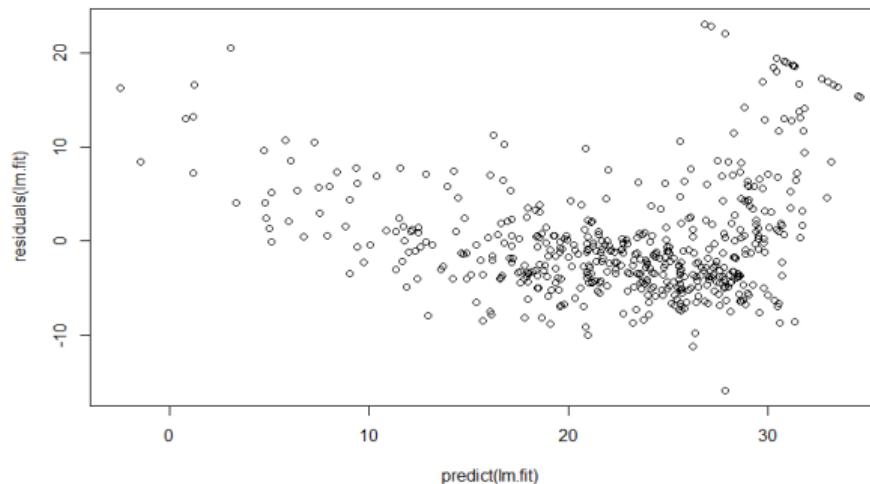
Boston Example

```
## Coefficients:  
##                                     Estimate Std. Error t value Pr(>|t|)  
## (Intercept)                 33.22276   0.73085  45.458 <2e-16 ***  
## lstat                  -1.03207   0.04819 -21.416 <2e-16 ***  
## age                     0.03454   0.01223   2.826  0.00491 ***
```

Boston Example: Interpreting the coefficients

- For a given percentage of households with low socioeconomic status (holding them constant/fixed), a one percent increase in the proportion of owner occupied units leads to an increase in median house price by \$34 on average.
- For a given proportion of owner occupied units (holding them constant/fixed), a one percent increase in the number of households with low economic status leads to a decrease in median house price by \$1032 on average.

Diagnostics: residual from the multiple regression



```
> plot(predict(lm.fit), residuals(lm.fit))
```

Boston Example

- The Boston dataset contains 13 variables, so it would be cumbersome to have to type all of these to perform the regression using all of the predictors.
- Instead, we can use the following short hand:

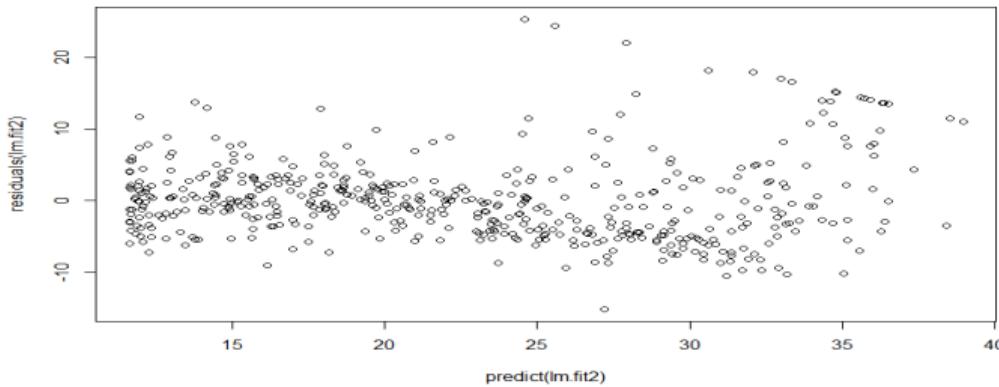
```
lm.fit<-lm(medv~ .,data=Boston)  
summary(lm.fit)
```

Boston Example: Nonlinear transformation of the predictors

- The `lm()` function can also accomodate non-linear transformations of the predictors.
- This following command performs a regression of `medv` with `lstat` and $lstat^2$.

```
> lm.fit2=lm(medv~lstat+I(lstat^2))  
> summary(lm.fit2)
```

Diagnostics: residual from the multiple regression



```
plot(predict(lm.fit2), residuals(lm.fit2))
```

Boston Example: Nonlinear transformation of the predictors

- The `poly()` function is able to create the polynomial within `lm()`.
- For example, the following command produces a fifth-order polynomial fit

```
> lm.fit5=lm(medv~poly(lstat,5))  
> summary(lm.fit5)
```

Boston Example: Nonlinear transformation of the predictors

- We can also use log-transformation.

```
> summary(lm(medv~log(rm),data=Boston))
```

This work is licensed under the Creative Commons Attribution-Noncommercial-Share Alike 2.5 Australia License.

To view a copy of this license, visit

<http://creativecommons.org/licenses/by-nc-sa/2.5/au/>



Distributions



School of Mathematics and Applied Statistics
Faculty of Engineering and Information Sciences

Random Variables

- A random variable is a numerical measurement of the outcome of a random phenomenon.
- An upper-case letter, such as X , refers to a random variable, which cannot be predicted with certainty.
- A lower-case letter, such as x , refers to a particular value of the variable X .

Discrete Probability Functions

A discrete random variable has values restricted to separate points. The *probability function (pf)* of a discrete RV is defined by

$$f(x) = P(X = x).$$

A probability function must satisfy

$$f(x) \geq 0 \quad \forall x, \quad \text{and} \quad \sum_x f(x) = 1.$$

The function may be specified by table or by formula:

x	0	1	2	Total
$f(x)$	0.3	0.55	0.15	1

Cumulative Distribution

The *cumulative distribution function (cdf)* of a discrete RV X , denoted by $F(x)$, is defined by

$$F(x) = P(X \leq x) = \sum_{k \leq x} f(k)$$

To avoid sums with many terms, we use differences of cdfs:

$$P(a < X \leq b) = F(b) - F(a)$$

Be **careful** with $<$ and \leq for discrete variables, eg.:

$$P(20 < X \leq 25) = F(25) - F(20)$$

Cumulative Distribution

We find $F(x)$ by summing values of $f(k)$. To find f from F , we use differences.

$$\begin{aligned}f(x) &= P(X = x) \\&= P(X \leq x) - P(X < x) \\&= F(x) - F(x - 1)\end{aligned}$$

Cumulative Distribution

Example:

x	0	1	2	3
$f(x)$	0.4	0.3	0.2	0.1
$F(x)$	0.4	0.7	0.9	1

$$F(2) = f(0) + f(1) + f(2) = 0.4 + 0.3 + 0.2 = 0.9$$

$$f(2) = F(2) - F(1) = 0.9 - 0.7 = 0.2$$

$$P(0 < X \leq 2) = f(1) + f(2) = F(2) - F(0) = 0.5$$

Expected Value

Def: The *expected value* $E(X)$ of a discrete RV X is defined by

$$E(X) = \sum_x xf(x).$$

$E(X)$ is a weighted average; greater weight is assigned to more likely values of X .

Expected Value

Example: Find the expected value of $f(x) = 0.1(4 - x)$, $x = 0, 1, 2, 3$.

Ans.

x	0	1	2	3
$f(x)$	0.4	0.3	0.2	0.1

$$E(X) = \sum_{x=0}^3 xf(x) = 0(0.4) + 1(0.3) + 2(0.2) + 3(0.1) = 1$$

Similarly, the expected value of any function $g(X)$ is $E[g(X)] = \sum_x g(x)f(x)$. For the above example,

$$E(X^2) = \sum_{x=0}^3 x^2 f(x) = 2$$

Expected Value

Properties of $E(X)$

- $E(a) = a$ for any constant a .
- For a linear transformation, $E(a + bX) = a + bE(X)$.

Note: for a nonlinear transformation $g(X)$, $E[g(X)]$ usually differs from $g(E(X))$, as in the last example $E(X^2) \neq [E(X)]^2$.

Mean and Variance

The *mean* of a discrete RV X is defined as

$$\mu = \mu_x = E(X)$$

For a large sample of observations, we expect the sample mean \bar{x} to be close to the theoretical mean μ .

Recall the sample variance is the average of squared distances from the sample mean:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

The *variance* of a discrete RV X is the expected squared distance from μ :

$$\sigma^2 = \text{Var}(X) = E[(X - \mu)^2] \tag{1}$$

Mean and Variance

A useful alternative representation is

$$\sigma^2 = E(X^2) - \mu^2 \quad (2)$$

Exercise: use properties of $E(X)$ to prove (1) = (2).

The *standard deviation* of X is the positive square root of the variance: $\sigma = \sqrt{\text{Var}(X)}$.

Mean and Variance

Example: find the variance for the previous example.

Ans. Recall that $\mu = 1$ and $E(X^2) = 2$, so $\sigma^2 = E(X^2) - \mu^2 = 1$. Or the longer way:

$$\begin{aligned}\sigma^2 &= E[(X - \mu)^2] = \sum_x (x - 1)^2 f(x) \\ &= (0 - 1)^2 0.4 + (1 - 1)^2 0.3 + (2 - 1)^2 0.2 + (3 - 1)^2 0.1 = 1\end{aligned}$$

Mean and Variance

Properties of Variance

- $\text{Var}(X) \geq 0 \quad \forall X$, and $\text{Var}(X) = 0 \Leftrightarrow X$ is constant.
- $\text{Var}(X + a) = \text{Var}(X) \quad \forall a \in \mathbb{R}$
- $\text{Var}(aX) = a^2 \text{Var}(X) \quad \forall a \in \mathbb{R}$

If a RV has large variance, it means that observations are expected to vary greatly.

Mean and Variance

Example: For $f(x) = 0.1(4 - x)$, we found $\sigma^2 = 1$.

- $\text{Var}\left(\frac{X}{2}\right) = \left(\frac{1}{2}\right)^2 \text{Var}(X) = \frac{1}{4}$
- $\text{Var}(X + 6) = \text{Var}(X) = 1$
- $\text{Var}(6 - 2X) = (-2)^2 \text{Var}(X) = 4$

Binomial Distribution

Binomial Scenario

- Fixed number of independent trials.
- 2 possible outcomes, success and failure.
- Constant probability of success for each trial.
- The quantity of interest is the total number of successes.

Notation:

n = number of trials

p = probability of success for a single trial

$q = 1 - p$ = probability of failure

x = number of successes

Binomial Distribution Function

Binomial Distribution Function

Let X be the number of successes in n independent trials, with constant probability p of success. The X has a binomial probability function

$$f(x) = P(X = x) = \binom{n}{x} p^x q^{n-x}, \quad x = 0, 1, \dots, n, \quad q = 1 - p.$$

R code:

`dbinom(x,n,p)`

Binomial Distribution

An interpretation of binomial distribution is that there are $\binom{n}{x}$ ways of grouping n objects, x of one type (success) and $(n - x)$ of another type (failure). Example:
 $\binom{3}{2} = SSF, SFS, FSS.$

Since the binomial scenario events are independent, the probability of a particular instance of x successes and $(n - x)$ failures in n trials (single path along the tree) is

$$\underbrace{p \cdot p \cdot \dots \cdot p}_{x \text{ times}} \cdot \underbrace{q \cdot q \cdot \dots \cdot q}_{(n-x) \text{ times}} = p^x q^{n-x}$$

The number of such paths is $\binom{n}{x}$, so the probability of x successes is

$$\boxed{\binom{n}{x} p^x q^{n-x}}$$

Binomial Distribution

Note that the sum of all binomial probabilities must be 1. We verify this by the Binomial Theorem.

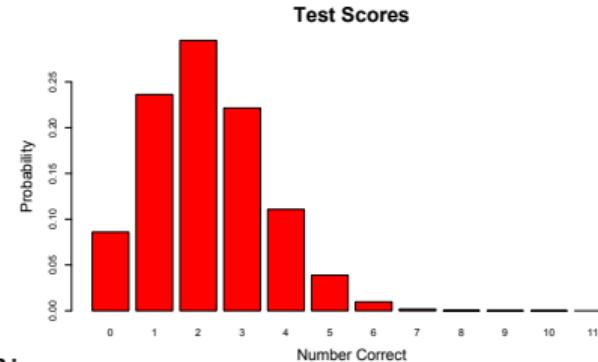
$$\begin{aligned} \binom{n}{0}q^n + \binom{n}{1}pq^{n-1} + \binom{n}{2}p^2q^{n-2} + \cdots + \binom{n}{n}p^n &= \sum_{k=0}^n \binom{n}{k}p^kq^{n-k} \\ &= (q+p)^n \\ &= (1-p+p)^n \\ &= 1 \end{aligned}$$

Binomial Distribution

Example: a multiple choice quiz has 11 questions with 5 possible answers each. What is the probability that a student who guesses at every question gets a score of 4 out of 11?

Ans. $n = 11$, $x = 4$, $p = 0.2$

$$f(4) = P(X = 4) = \binom{11}{4} 0.2^4 0.8^{11-4} = 0.1107$$



One can calculate each outcome and graph them:

So don't guess; you'll most likely get 2/11.

Cumulative Distribution

Cumulative Distribution

R code:

`pbinom(3,11,0.2)`

This gives $P(X \leq 3) = f(0) + f(1) + f(2) + f(3) = 0.8389$ for $n = 11$ and $p = 0.2$.

By hand, this is

$$\binom{11}{0}0.2^00.8^{11} + \binom{11}{1}0.2^10.8^{10} + \binom{11}{2}0.2^20.8^9 + \binom{11}{3}0.2^30.8^8.$$

Cumulative Distribution

Example: In the previous example, what is the probability that the student gets at least 4 out of 11?

Ans. The long way: $P(X \geq 4) = P(X = 4) + P(X = 5) + \cdots + P(x = 11)$.

The short way: $P(X \geq 4) = 1 - P(X < 4) = 1 - 0.8389 = 0.1611$.

Mean and Variance

For a binomial distribution, we find that

$$① \mu = E(X) = np$$

$$② \sigma^2 = Var(X) = np(1 - p)$$

$$③ \sigma = \sqrt{Var(X)} = \sqrt{np(1 - p)}$$

Mean and Variance

Example: Find the mean and standard deviation of the number X of heads obtained in 100 tosses of a fair coin.

Ans. Binomial distribution, $n = 100$, $p = 0.5$.

$$\mu = np = 50$$

$$\sigma = \sqrt{np(1 - p)} = 5$$

This means that although X will be about 50 on average, it would not be uncommon to observe values between 45 and 55 ($\mu - \sigma$ and $\mu + \sigma$).

Geometric Distribution

Geometric Distribution

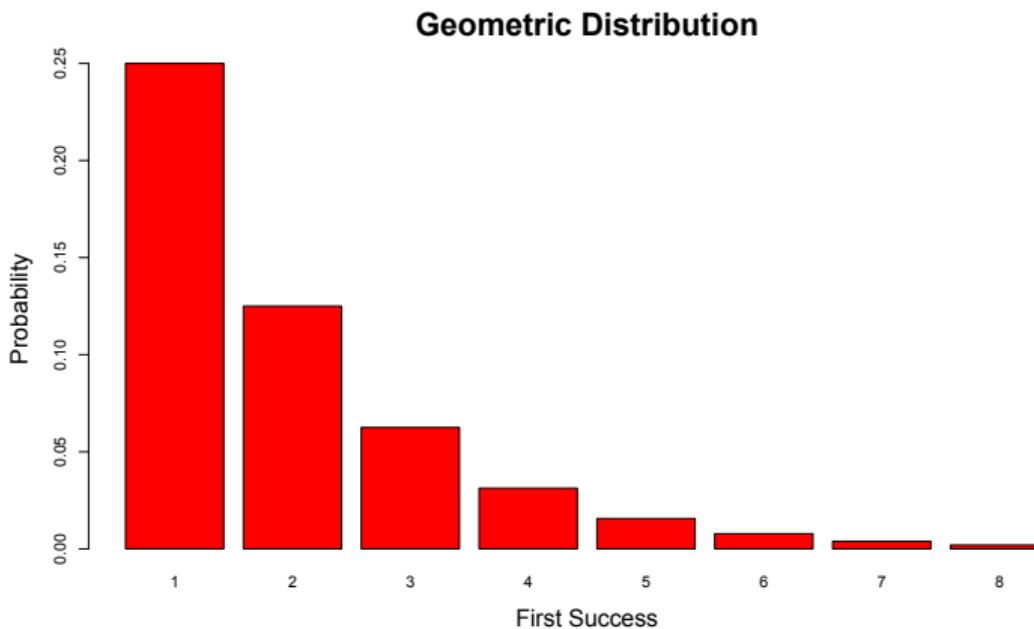
The geometric distribution arises when counting the number X of trials until the first success. The final (successful) trial is also counted, so $X \in \{1, 2, \dots, n\}$. The discrete probability function is

$$f(x) = q^{x-1} p, \quad x = 1, 2, \dots, n$$

Geometric Distribution

Geometric Distribution

The shape of the graph is strongly skewed to the right.



Geometric Distribution

To find the cdf of the geometric distribution, we do the following.

$$\begin{aligned}F(x) &= p + qp + q^2p + q^3p + \cdots + q^{x-1}p \\&= p \frac{1 - q^x}{1 - q} \\&= 1 - q^x, x = 1, 2, \dots, n.\end{aligned}$$

When convenient, you can also use $F(x) = P(X \leq x) = 1 - P(X > x)$, where $P(X > x)$ is the probability of no success in the first x trials.

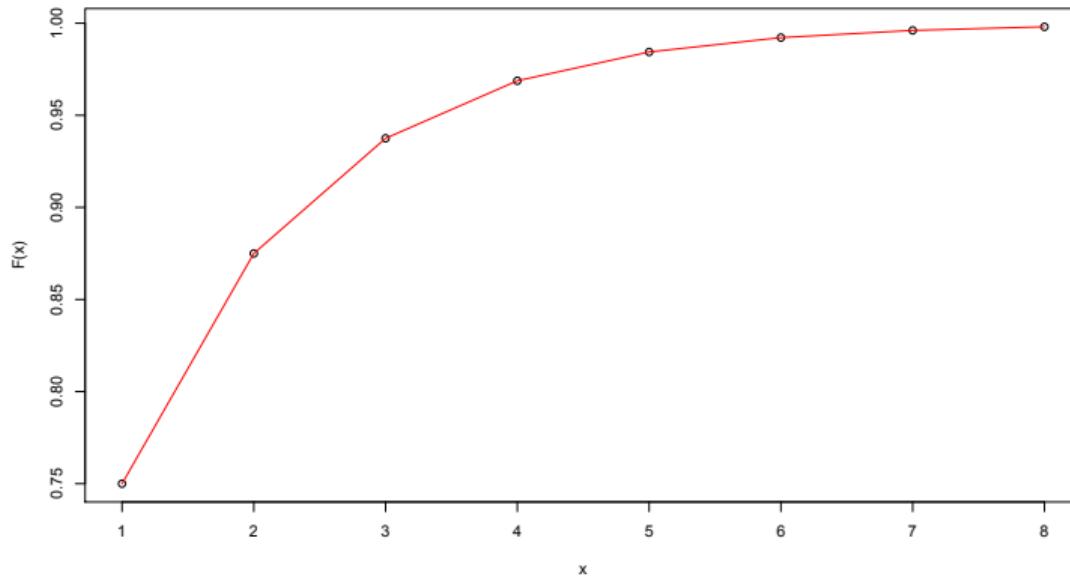
Geometric Distribution

We can check the cdf result by using the difference formula:

$$\begin{aligned}f(x) &= F(x) - F(x - 1) \\&= 1 - q^x - (1 - q^{x-1}) = q^{x-1} - q^x \\&= q^{x-1}(1 - q) = \color{blue}{q^{x-1} p}\end{aligned}$$

Geometric Distribution

the cdf (geometric or not) approaches 1 as x increases.



Geometric Distribution

The mean μ of a geometric distribution involves a summation property that we will not prove here.

$$\begin{aligned}\mu = E(Y) &= \sum_{y=1}^{\infty} yf(y) = \sum_{y=1}^{\infty} y(1-p)^{y-1}p \\ &= p \sum_{k=0}^{\infty} (k+1)(1-p)^k \quad (k = y-1) \\ &= p \frac{1}{(1-(1-p))^2} = 1/p\end{aligned}$$

Geometric Distribution

Example: Students guess at questions with 5 multiple-choice answers each. On average, how long will it take them to get their first right answer? What is the probability that a random student takes longer than average?

Ans. X is geometric with $p = 0.2$, so

$$E(X) = \frac{1}{p} = 5$$

$$\begin{aligned}P(X > 5) &= 1 - P(X \leq 5) = 1 - F(5) \\&= 1 - [1 - (1 - 0.2)^5] = 0.3277\end{aligned}$$

So on average, it will take a student 5 questions to get 1 right, and there's a 33% chance that it will take more than 5 questions.

Poisson Distribution

The *Poisson distribution* is a discrete pdf that applies when we count the number of points in a given time, area, distance or volume. For instance, the number of cars that go through an intersection in 10 minutes, the number of rust spots in a $1m^2$ area of your car, etc.

Let λ be the average rate of occurrences per unit time (or distance, area, volume). Then the expected number μ of occurrences in an interval of length t is

$$\mu = \lambda t$$

Poisson Distribution

- With $\mu = \lambda t$, we obtain the Poisson probability function:

$$f(x) = \frac{\mu^x}{x!} e^{-\mu}, \quad x = 0, 1, 2, \dots$$

There are infinitely many nonzero probabilities, but the sum is still 1:

$$\sum_{x=0}^{\infty} \frac{\mu^x}{x!} e^{-\mu} = e^{-\mu} \left(1 + \mu + \frac{\mu^2}{2} + \frac{\mu^3}{6} + \dots \right) = e^{-\mu} e^{\mu} = 1.$$

For a Poisson RV, $E(X) = \text{Var}(X) = \mu$.

Poisson Distribution

Example: The UOW switchboard receives on average 0.6 calls per minute. Find the probability that in a 4-minute interval there will be (i) exactly 3 calls, (ii) at least 3 calls.

Ans. The rate of calls is $\lambda = 0.6$ per minute, so

$$\mu = \lambda t = 0.6 \cdot 4 = 2.4$$

(i) $P(X = 3) = \frac{2.4^3}{3!} e^{-2.4} \approx 0.209$

R code:

`dpois(3,2.4)`

Poisson Distribution

(ii) $P(X \geq 3) = 1 - P(X < 3)$

$$P(X < 3) = f(0) + f(1) + f(2) = \frac{2.4^0}{0!} e^{-2.4} + \frac{2.4^1}{1!} e^{-2.4} + \frac{2.4^2}{2!} e^{-2.4} = 0.5697$$

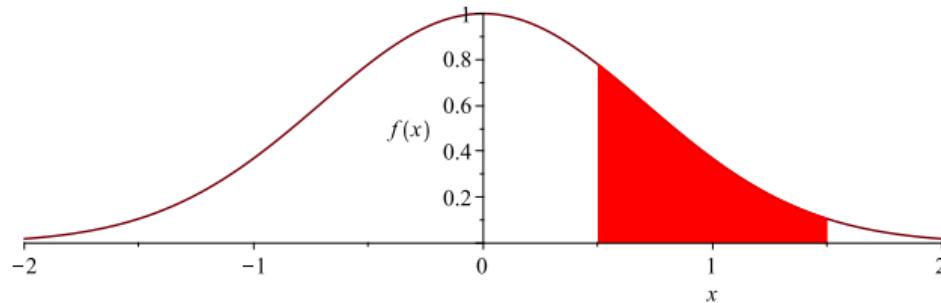
$$\therefore P(X \geq 3) = 1 - 0.5697 = 0.4303$$

R code:

1-ppois(2,2.4)

Continuous RVs

A continuous random variable is one which takes an infinite number of possible values (uncountable). Continuous random variables are usually measurements. Examples include height, weight, the amount of sugar in an orange, the time required to run a mile. The *probability density function (pdf)* of a continuous RV X is the function f such that $P(a < X < b)$ is the area $\int_a^b f(x)dx$ under the curve $y = f(x)$ between $x = a$ and $x = b$.



Continuous Distributions

The total area under the whole curve must be 1.

Properties of Continuous pdf

- For values of x that are never observed, $f(x) = 0$.
- $f(x) \geq 0 \forall x$
- $\int_{-\infty}^{\infty} f(x)dx = 1$

Continuous Distributions

Example: Show that $f(x) = 3x^2$, $x \in (0, 1)$ is a valid pdf. Find $P(0.2 < X < 0.9)$.

Ans. Since $\text{dom } f = (0, 1)$, $f(x) = 0$ for all $x \notin (0, 1)$. We need to check that $f(x) \geq 0 \forall x \in \text{dom } f$, and that the integral over the whole space is 1.

$$3x^2 \geq 0 \quad \forall x$$

$$\begin{aligned}\int_{-\infty}^{\infty} f(x)dx &= \int_{-\infty}^0 f(x)dx + \int_0^1 f(x)dx + \int_1^{\infty} f(x)dx \\ &= 0 + \int_0^1 3x^2 dx + 0 = x^3 \Big|_0^1 = 1\end{aligned}$$

Continuous Distributions

Example: Show that $f(x) = 3x^2$, $x \in (0, 1)$ is a valid pdf. Find $P(0.2 < X < 0.9)$.

$$P(0.2 < X < 0.9) = \int_{0.2}^{0.9} 3x^2 dx = 0.9^3 - 0.2^3 = 0.721$$

Continuous cdf

The *cumulative density function (cdf)* of a continuous RV X is defined by

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t)dt$$

If there are cdfs available, integration can be avoided by using

$$P(a < X < b) = F(b) - F(a).$$

Continuous cdf

Properties of cdf

- As $F(x)$ is a probability, it must lie between 0 and 1.
- As x increases, the event $\{X \leq x\}$ includes more outcomes, so $F(x)$ is an increasing function of x .
- For continuous X , F is continuous. For discrete X , F is a step function.

Continuous cdf

We obtain F from f by integration, so to obtain f from F we use differentiation. By the Fundamental Theorem of Calculus,

$$\frac{d}{dx}F(x) = \frac{d}{dx} \int_{-\infty}^x f(t)dt = f(x),$$

i.e. $f(x) = \frac{d}{dx}F(x).$

Example: If $F(x) = x^3$, $0 < x < 1$, then

$$f(x) = \frac{d}{dx}F(x) = \frac{d}{dx}x^3 = 3x^2, \quad 0 < x < 1.$$

Mean and Variance

Mean and Variance

The expected value of a function of a continuous RV is

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx.$$

. As in the discrete case, we have

- mean $\mu = E(X)$,
- variance $\sigma^2 = E[(X - \mu)^2] = E(X^2) - \mu^2$,
- standard deviation $\sigma = \sqrt{\text{Var}(X)}$.

Mean and Variance

Example: For $f(x) = 3x^2$, $0 < x < 1$, find μ , $E(4X - 2)$, $E\left(\frac{1}{X}\right)$, $\frac{1}{E(X)}$, $E(X^2)$, $\text{Var}(X)$ and σ .

Ans.

$$\mu = E(X) = \int_0^1 xf(x)dx = \int_0^1 3x^3 dx = \frac{3x^4}{4} \Big|_0^1 = \frac{3}{4}$$

$$E(4X - 2) = 4E(X) - 2 = 1$$

$$E\left(\frac{1}{X}\right) = \int_0^1 \frac{1}{x} 3x^2 dx = \frac{3x^2}{2} \Big|_0^1 = \frac{3}{2}, \frac{1}{E(X)} = \frac{4}{3}$$

Mean and Variance

Example: For $f(x) = 3x^2$, $0 < x < 1$, find μ , $E(4X - 2)$, $E\left(\frac{1}{X}\right)$, $\frac{1}{E(X)}$, $E(X^2)$, $\text{Var}(X)$ and σ .

Ans.

$$E(X^2) = \int_0^1 x^2 \cdot 3x^2 dx = \frac{3x^5}{5} \Big|_0^1 = \frac{3}{5}$$

$$\text{Var}(X) = E(X^2) - \mu^2 = \frac{3}{5} - \frac{9}{16} = \frac{3}{80}, \sigma = \sqrt{\frac{3}{80}} = \frac{1}{4}\sqrt{\frac{3}{5}}$$

Median

The median Q_2 of a continuous RV X is found by solving $P(X \leq Q_2) = F(Q_2) = 0.5$.

We want half the area under the pdf to be on either side of $x = Q_2$. Similarly, the upper and lower quartiles satisfy $F(Q_3) = 0.75$ and $F(Q_1) = 0.25$.

Example: Let $F(x) = x^3$, $0 < x < 1$. Find μ and Q_2 . What do these numbers say about f ?

$$f(x) = \frac{d}{dx}F(x) = 3x^2. \text{ From the previous example, } \mu = \frac{3}{4}.$$

$$F(Q_2) = 0.5 \Rightarrow Q_2 = 0.5^{\frac{1}{3}} \approx 0.7937.$$

The median is bigger than the mean, so f is skewed to the left.

Exponential Distribution

The *exponential distribution* is generated from the Poisson process. Let T be the waiting time until the first point with rate λ per unit time. Then

$$P(T > t) = P(\text{no points in } [0, t]).$$



The number of points in $[0, t]$ is a Poisson RV with mean λt :

$$P(T > t) = \frac{(\lambda t)^0}{0!} e^{-\lambda t} = e^{-\lambda t}$$

Exponential Distribution

Def: T has an *exponential distribution* with rate parameter λ if the pdf and cdf, respectively, have the form

$$f(t) = \lambda e^{-\lambda t}, t \geq 0,$$

$$F(t) = 1 - e^{-\lambda t}, t \geq 0.$$

The applications are for inter-arrival times of customers in queue, and time until failure in reliability models.

Exponential Distribution

Exponential Mean and Median

An exponential R has $\mu = \frac{1}{\lambda}$ and $\sigma = \frac{1}{\lambda}$.

To find the median, solve $F(Q_2) = \frac{1}{2}$:

$$-\lambda Q_2 = \ln \frac{1}{2} \Rightarrow Q_2 = \frac{1}{\lambda} \ln 2$$

Exponential Distribution

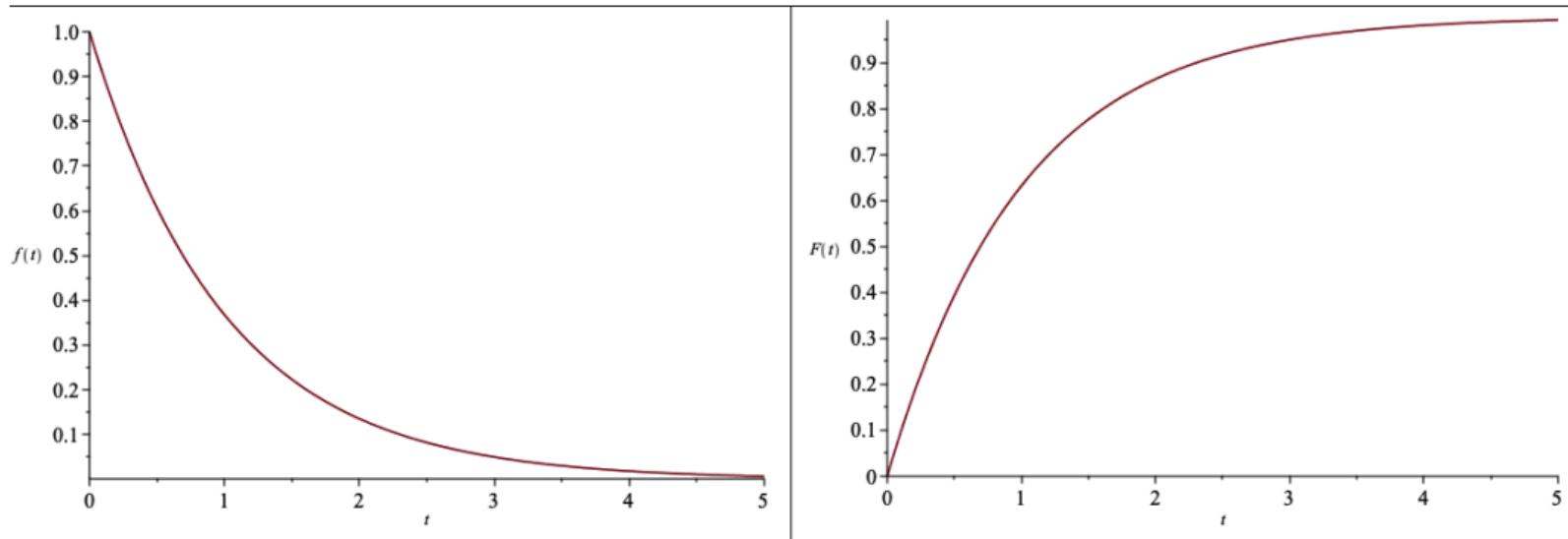


Figure: The plots of $f(t)$ and $F(t)$ for an exponential distribution.

Exponential Distribution

Example: The UOW switchboard receives on average 0.6 calls per minute, according to the Poisson process. The first call of the day arrives at T minutes after 9:00am.

- (i) What is the probability that $T < 2$ min.?
- (ii) What is the median of T ?

Exponential Distribution

Ans. (i) Since T is an exponential RV with $\lambda = 0.6$, we have

$$F(t) = 1 - e^{-0.6t}$$

$$P(T < 2) = F(2) = 1 - e^{-0.6 \cdot 2} \approx 0.6988$$

R code:

`pexp(2,0.6)`

(ii) $F(Q_2) = \frac{1}{2} = 1 - e^{-0.6Q_2} \Rightarrow Q_2 = \frac{\ln 2}{0.6} \approx 1.155$

Notice that the median is less than the mean $\frac{1}{\lambda} = \frac{5}{3}$, so f is skewed to the right.

Normal Distribution

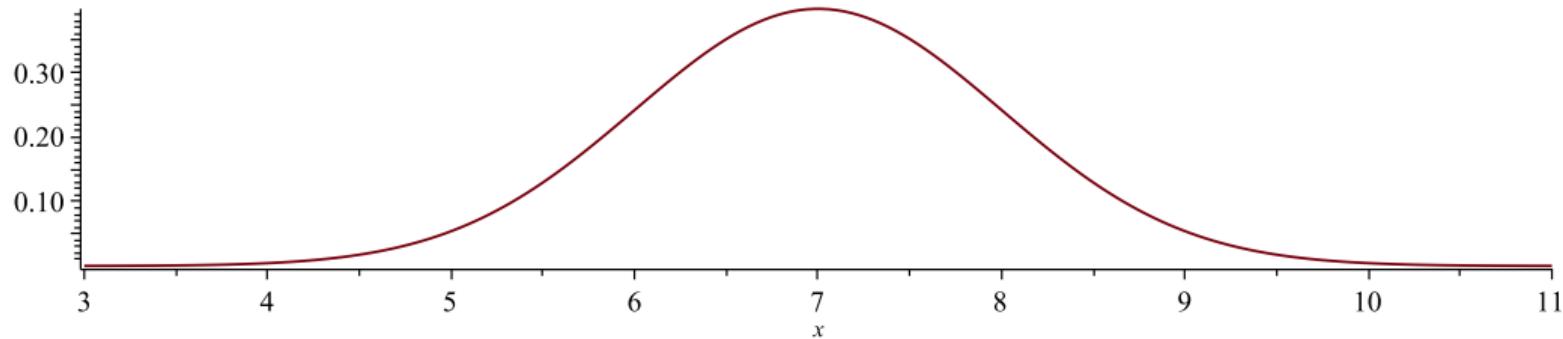
Normal Distribution

A normal (Gaussian) RV has pdf

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty$$

- The notation $N(\mu, \sigma^2)$ is often used. For a *standard* normal RV, $\mu = 0$ and $\sigma = 1$.
- The normal pdf is symmetric, bell-shaped and characterised by μ and σ . There are no upper/lower bounds, but it is very unlikely to observe values more than 3σ away from μ .

Normal Distribution



Analytic expressions do not exist for the normal cdf. Probabilities are found numerically, by tables or software.

R code:

`pnorm`

Normal Distribution

Any normal RV can be standardised with the variable change $Z = \frac{X-\mu}{\sigma}$. Then

$$E(Z) = \frac{E(X) - \mu}{\sigma} = 0 \text{ and } \text{Var}(Z) = \frac{\sigma^2}{\sigma^2} = 1$$

So for $X \sim N(\mu, \sigma^2)$, $Z \sim N(1, 0)$. Therefore, standard normal tables are sufficient for any normal problem:

$$P(X \leq x) \equiv P\left(Z \leq \frac{x-\mu}{\sigma}\right)$$

The Normal Distribution

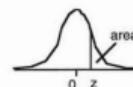
STAT231

Standard Normal Distribution Table

Normal curve areas

Standard Normal probability in right-hand tail

(for negative values of z areas are found by symmetry)

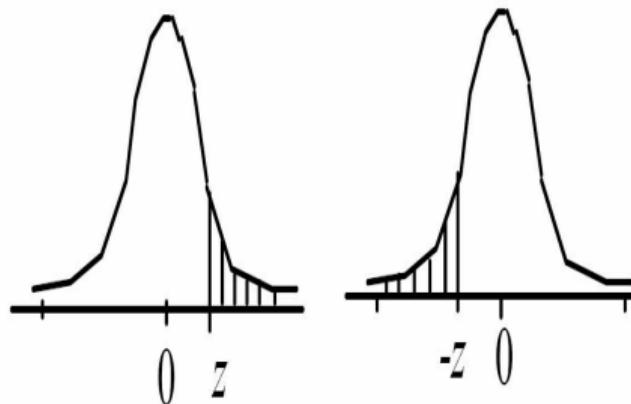


Second decimal place of z

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064

The Normal Distribution

- The table tabulates $P(Z > z) = 1 - F(z)$ (where $Z \sim N(0, 1)$) for positive z .
- We can get $F(z)$ for negative z from this, because $P(Z \leq -z) = P(Z \geq z)$ because distribution is symmetric. So $F(-z) = 1 - F(z)$.



Quantiles of Normal r.v.

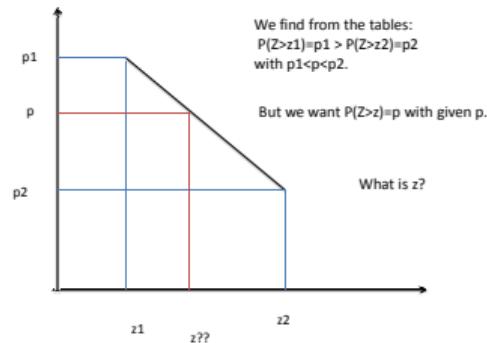
- p -quantile $y_p = F^{-1}(p)$. So can look up the normal probability tables to find the value y_p such that $F(y_p) = p$.
- For example, if $Z \sim N(0, 1)$

$$P(Z \leq z_{0.75}) = 0.75$$

$$\Rightarrow P(Z > z_{0.75}) = 0.25.$$

- From standard Normal table we cannot find a z with $P(Z > z) = 0.2500$. You might round, however better would be to interpolate. We find
 $p_1 = P(Z > 0.67) = 0.2514$ and $p_2 = P(Z > 0.68) = 0.2483$ with $z_1 = 0.67$ and $z_2 = 0.68$. How do we interpolate?

Quantiles of Normal r.v.



Apply equation of line $y = a + b(x - c)$:

$$p = p_1 + \frac{p_2 - p_1}{z_2 - z_1}(z - z_1) = p_1 + \frac{\Delta p}{\Delta z}(z - z_1)$$

Quantiles of Normal r.v.

- We obtain

$$(p - p_1) \frac{z_2 - z_1}{p_2 - p_1} = z - z_1$$

and

$$z = z_1 + \frac{p - p_1}{p_2 - p_1} (z_2 - z_1)$$

In our case $z_2 - z_1 = 0.68 - 0.67 = 0.01$ and we obtain

$$z = 0.67 + \frac{0.25 - 0.2514}{0.2483 - 0.2514} \times 0.01$$

- Hence $z_{0.75} \approx 0.6745$.

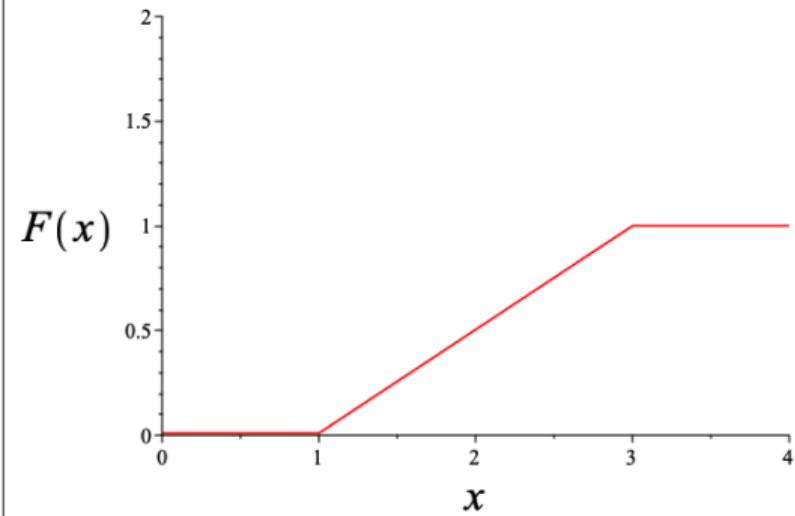
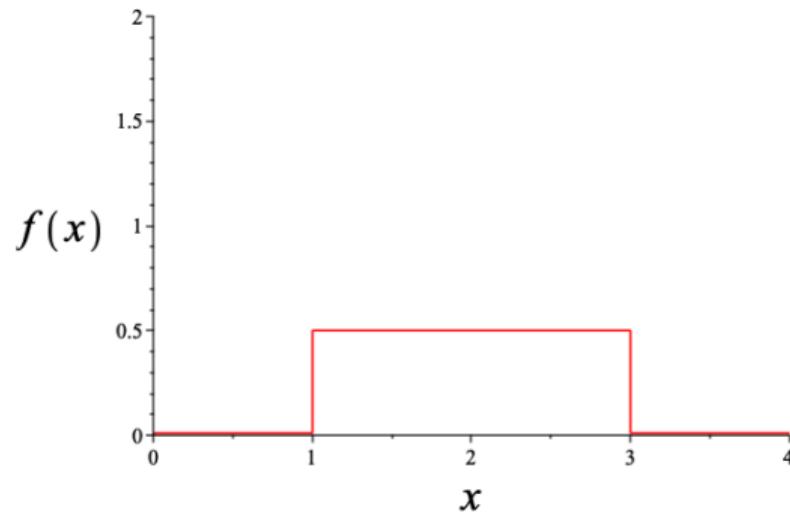
Uniform Distribution

Uniform Distribution

A continuous RV has uniform distribution on (a, b) if its pdf is constant on (a, b) .

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq x \leq b \\ 0, & \text{otherwise} \end{cases} \quad F(x) = \begin{cases} 0, & \text{if } x < a \\ \int_a^x \frac{1}{b-a} dt, & \text{if } a \leq x \leq b \\ 1, & \text{if } x > b \end{cases}$$

Uniform Distribution



Uniform Distribution

Uniform Mean and Variance

$$E(X) = \int_a^b \frac{x}{b-a} dx = \frac{x^2}{2(b-a)} \Big|_a^b = \frac{b^2 - a^2}{2(b-a)} = \frac{b+a}{2}$$

$$E(X^2) = \int_a^b \frac{x^2}{b-a} dx = \frac{b^2 + ba + a^2}{3}$$

$$\Rightarrow \text{Var}(X) = \frac{b^2 + ba + a^2}{3} - \frac{(b+a)^2}{4} = \frac{(b-a)^2}{12}$$

This work is licensed under the Creative Commons Attribution-Noncommercial-Share Alike 2.5 Australia License.

To view a copy of this license, visit

<http://creativecommons.org/licenses/by-nc-sa/2.5/au/>



March 12, 2023

MATH55

Alberto Nettel Aguirre

NIASRA
School of Mathematics and Statistics
University of Wollongong

May 16, 2023

Why do we do Statistics?

- We do not have all the time, nor all the money, nor access to all data, etc.
- to **calculate** population parameters exactly;
- hence, we use samples to **estimate** them.

Now,

- for our specific sample the estimate is exact,
- **but** as an estimate of the population it is subject to sampling error,
- there is variability depending on the sample we get.

Point Estimates

A **point estimate** is a **single number** calculated from our **sample**, that is our “best guess” for the parameter.

E.g.

- Estimate the population mean using the mean from a sample.
- Estimate population proportion using sample proportion.

BUT...

There are many possible samples, we observe one (the one we got) and hence that is just one “point” in the milieu of possible samples.

Sampling Distribution

Sampling Distribution

You may have seen several models for discrete and continuous RVs. How do we know which model to use for a particular data set? How does limited sample size affect the choice?

Consider n observations X_1, \dots, X_n that follow the same distribution, with $E(X_i) = \mu$. By linearity of expected value,

$$E(X_1 + X_2 + \cdots + X_n) = \mu + \mu + \cdots + \mu = n\mu$$

Using the *sample mean* $\bar{X} = \frac{X_1 + \cdots + X_n}{n}$, we find

$$E(\bar{X}) = E\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{1}{n}n\mu = \mu$$

Sampling Distribution

There are 3 different concepts of “mean” here:

- ① distribution of observations within a data set, centred about the sample mean \bar{x} ;
- ② distribution of a RV X , centred about the population mean μ ;
- ③ **sampling** distribution of \bar{x} , describing variation of the sample mean over all possible samples.

If X_1, \dots, X_n are random, then \bar{X} is also a RV.

Sampling Distribution

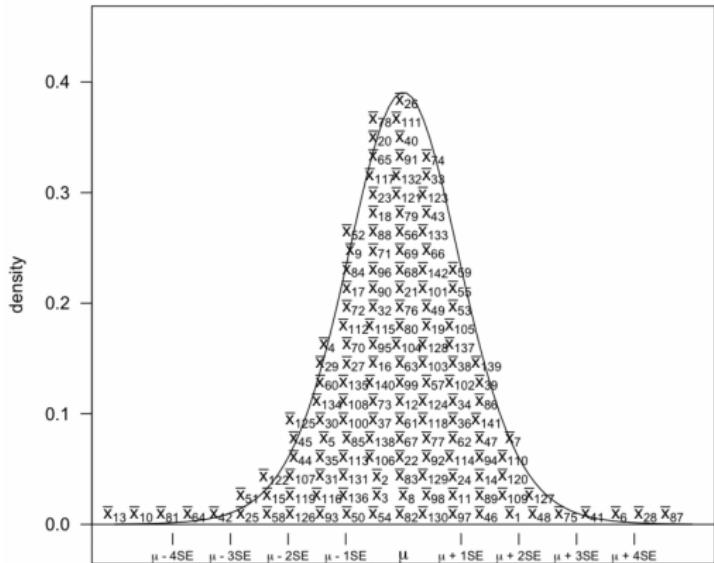
- The sampling distribution of \bar{X} depends on the underlying probability model of a single random variable X .
- The result $E(\bar{X}) = \mu$ says that \bar{X} has the same expected value as a single r.v or observation.
- However, we expect that averaging repeated measurements should increase accuracy. So the sampling distribution of \bar{X} should vary according to the sample size n , with reduced spread as n increases.

Sampling Distribution for the Sample Mean

If \bar{X} is the sample mean of a random sample from a normal $N(\mu, \sigma^2)$

Then \bar{X} is approx. normal with Mean= μ

Sampling distribution of the sample mean \bar{X}



Variability of \bar{X} (standard error)

With n independent observations X_1, \dots, X_n of a RV with μ and σ^2 , the sample mean is $\bar{X} = \frac{X_1 + \dots + X_n}{n}$. This mean has its own expected value, variance and standard deviation.

$$E(\bar{X}) = \frac{1}{n} E(X_1 + \dots + X_n) = \frac{n\mu}{n} = \mu$$

$$\text{Var}(\bar{X}) = \frac{1}{n^2} \text{Var}(X_1 + \dots + X_n) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$$\sigma_{\bar{X}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

$\sigma_{\bar{X}}$ is known as the *standard error*.

Standard Error

- The standard error of the sample mean is $\frac{\sigma}{\sqrt{n}}$.
- As the sample size increases, the standard error decreases.
- Therefore, with large samples, the sample mean is more likely to be close to μ .

If the sample comes from a normal population, then

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right). \text{ Equivalently, } \bar{Z} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

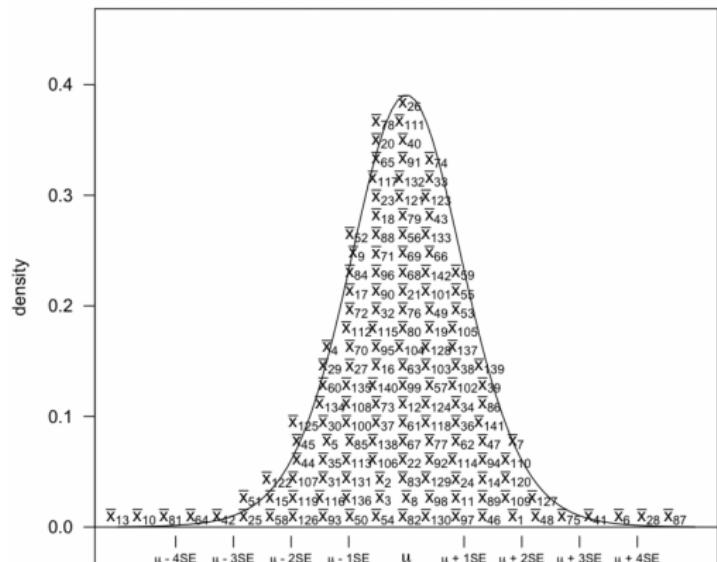
Interestingly, this remains true for samples from any distribution, provided $\sigma < \infty$ and n is large.

Sampling Distribution for the Sample Mean

If \bar{X} is the sample mean of a random sample from Normal with $E(X) = \mu$ and $SD(X) = \sqrt{\text{Var}(X)} = \sigma$

Then \bar{X} is approx. normal with Mean = μ and SD = $\frac{\sigma}{\sqrt{n}}$

Sampling distribution of the sample mean \bar{X}



Sampling Distribution

Example. Weights of tiles are normally distributed with $\mu = 1\text{kg}$ and $\sigma = 20\text{g}$. Find the probability that a pack of 12 tiles has average weight below 995g.

Ans. $\bar{X} \sim N(1000, 20^2/12)$.

$$\begin{aligned}P(\bar{X} < 995) &= P\left(\bar{Z} < \frac{995 - 1000}{20/\sqrt{12}}\right) \\&= P(\bar{Z} < -0.866) \approx 0.1933.\end{aligned}$$

Central Limit Theorem

Central Limit Theorem: For random sampling with a large sample size n , the sampling distribution of the sample mean is approximately normal with mean μ and standard error $\frac{\sigma}{\sqrt{n}}$. This is true no matter the type of probability distribution that provides the samples.

- The sampling distribution of the sample mean takes more of a bell shape as n increases.
- The more skewed the population distribution, the larger n must be before the shape is close to normal.
- In practice, $n \geq 30$ usually starts getting close to normal.
- A nice animated demonstration

Central Limit Theorem

Example. 1000 random observations were simulated from the pdf $f(x) = 2x$, $0 < x < 1$. This was repeated 16 times to form a table of 16 columns and 1000 rows. For each row, averages were calculated for the first 2, first 4, and all 16 observations.

The expected value of a single observation is

$$\mu = E(X) = \int_0^1 x \cdot 2x dx = \frac{2x^3}{3} \Big|_0^1 = \frac{2}{3}$$

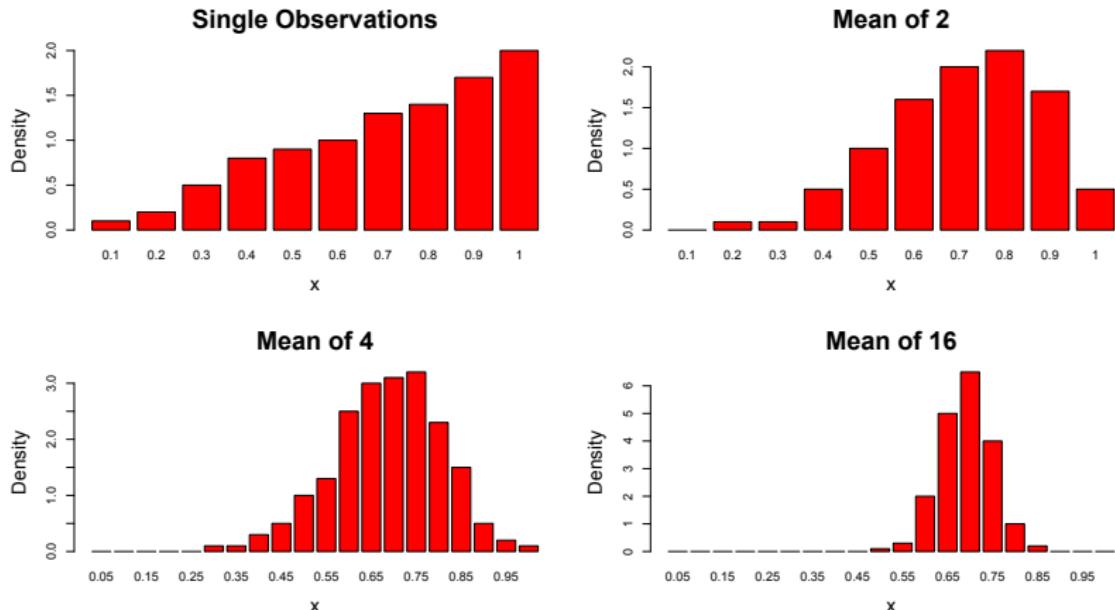
The variance of a single observation is

$$E(X^2) - \mu^2 = \int_0^1 x^2 \cdot 2x dx - \frac{4}{9} = \frac{x^4}{2} \Big|_0^1 - \frac{4}{9} = \frac{1}{2} - \frac{4}{9} = \frac{1}{18}$$

So the variance for an average of n observations is

$$\frac{\sigma^2}{n} = \frac{1}{18n}$$

Central Limit Theorem



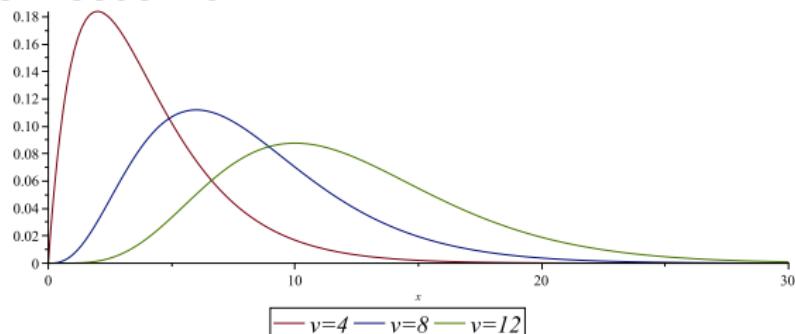
Note that as n increases,

- the shape becomes more symmetric and bell-like,
- the centre remains about the same,
- the spread becomes smaller.

Sampling distribution of s^2

Consider a sample X_1, \dots, X_n of independent $N(\mu, \sigma^2)$ observations. The *sample variance* $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ varies among samples. The sampling distribution of $\frac{(n-1)s^2}{\sigma^2}$ is called a chi-squared (χ^2) distribution with $n - 1$ degrees of freedom.

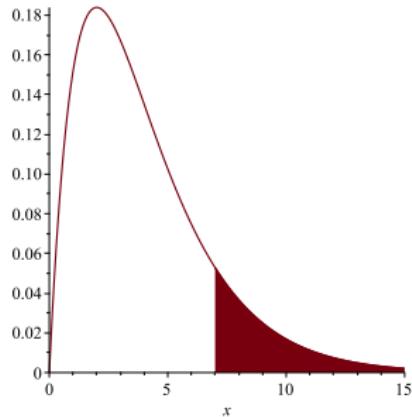
χ^2 is a continuous model with many applications. The minimum possible value is 0; there is no maximum. The shape, mean ν and variance 2ν depend on a parameter known as the degrees of freedom df .



χ^2 Distribution

Tables usually list χ^2 values for right-tail probabilities α . Some tables include left-tail areas $1 - \alpha$.

	α	0.10	0.05	0.025
df	$1 - \alpha$	0.90	0.95	0.975
1		2.706	3.841	5.024
3		6.251	7.815	9.348
5		9.236	11.070	12.833



Shaded region is α , non-shaded region is $1 - \alpha$.

χ^2 Distribution

Example. For a sample of size 6 from a normal population with $\mu = 70$ and $\sigma^2 = 45$, look up a χ^2 table with $6 - 1 = 5 \text{ df}$ to find

$$P\left(\frac{5s^2}{45} > 11.070\right) = 0.05$$

$$P(s^2 > 99.63) = 0.05$$

$$P(s > 9.981) = 0.05$$

WHy do we care about χ^2 distribution?

When we have a data, a sample

- Most likely we will not know the population's σ
- we use the sample sd, s to estimate it.
- that estimate carries some error in it
- our estimation of variability may be off

Student's t Distribution

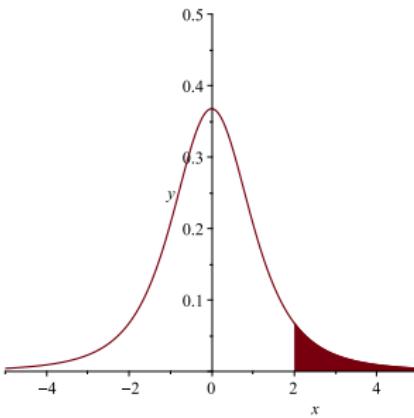
Using $\bar{Z} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ on a random sample from a $N(\mu, \sigma^2)$ distribution leads to the $N(0, 1)$ distribution. Using $T = \frac{\bar{X}-\mu}{s/\sqrt{n}}$ instead, one obtains the student's t distribution with $n - 1$ df, written $T \sim t_{n-1}$.

- The t distribution is bell-shaped and symmetric about 0.
- The t distribution has thicker tails and is wider spread than the standard normal distribution.
- The probabilities depend on the degrees of freedom.
- For a t score based on a single sample of size n ,
 $df = n - 1$.
- As $df \rightarrow \infty$, the t distributions approaches standard normal.

Student's t Distribution

Tables list values of $t_{df;\alpha}$ (right-tail) and some tables have $1 - \alpha$ left-tail areas.

	α	0.10	0.05	0.025
df	$1 - \alpha$	0.90	0.95	0.975
1		3.078	6.314	12.606
3		1.638	2.815	3.182
5		1.476	2.025	2.571



Example. For a sample of size 6 from a normal population with $\mu = 70$, look up a t table with $6 - 1 = 5$ df to find

$$P\left(T = \frac{\bar{X} - 70}{s/\sqrt{6}} < 1.476\right) = 0.90 \Rightarrow P(T < -1.476) = 0.10.$$

Point and Interval Estimates

A **point estimate** does not tell us anything about how close the estimate may be to the parameter.

An **interval estimate** is more useful, giving us an idea of precision, as it incorporates a margin of error.

Point estimates are the most common form reported by the mass media and rarely do we get interval estimates.

Point estimates alone are harder to conclude from.

In general, whatever the parameter of interest is (could be a mean, a proportion, etc.)

An **interval estimator** is a method for calculating **the** 2 numbers, *Lower_limit* and *Upper_limit*, that enclose the interval.

Lower_limit and *Upper_limit* exist theoretically within a sample distribution,

And we estimate them by as functions of our observed sample X_1, \dots, X_n , therefore they are random variables themselves.

ideally, we would like this interval to possess 2 properties:

- The interval contains the population parameter, so
 $Lower_limit \leq parameter \leq Upper_limit$
 - **Expected Coverage**
- The interval is as narrow as possible
 - **Precision**

In practice, there is always a tradeoff between these goals.

Confidence intervals

What is the whole tradeoff about? **Intuitively...**

- One can be 100% confident that their final grade in PGPH901 will be between 0% and 100%
- Yet, this is not giving us really any information (no precision on what it will be).
- if you want more precision, would you be 100% confident (as in , put your money on it) that your grade will be between 90% and 95%?
- most likely your confidence is reduced
- the more information you get, the more confident you may be about being tighter around a certain number

Introduction to Confidence Intervals

Scenario

Caz and Carole have to meet at the station to catch a train for a shopping trip to Sydney.

Carole: "What time will the train be at the station?"

Caz: "3pm, give or take 5 minutes."

What does Caz mean by her answer? How else could she say it?

How likely is it that the train will arrive when she says it will? Caz has checked that the trains are very punctual and that 9 times out of 10, they do arrive when the schedule says it will.

What could go wrong?

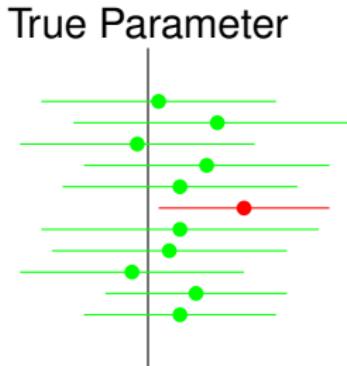
Confidence Intervals

The probability that a confidence interval (C.I.) we build will enclose the parameter is called the **confidence level**:

For example, setting $\alpha = 0.05$ produces a 95% C.I.

In the long run, we expect 95% (or $100 \times (1 - \alpha)\%$) of the CIs to contain θ .

A *wider* interval has a *greater* chance of ‘capturing’ the true parameter, but gives a vaguer indication of the likely value.



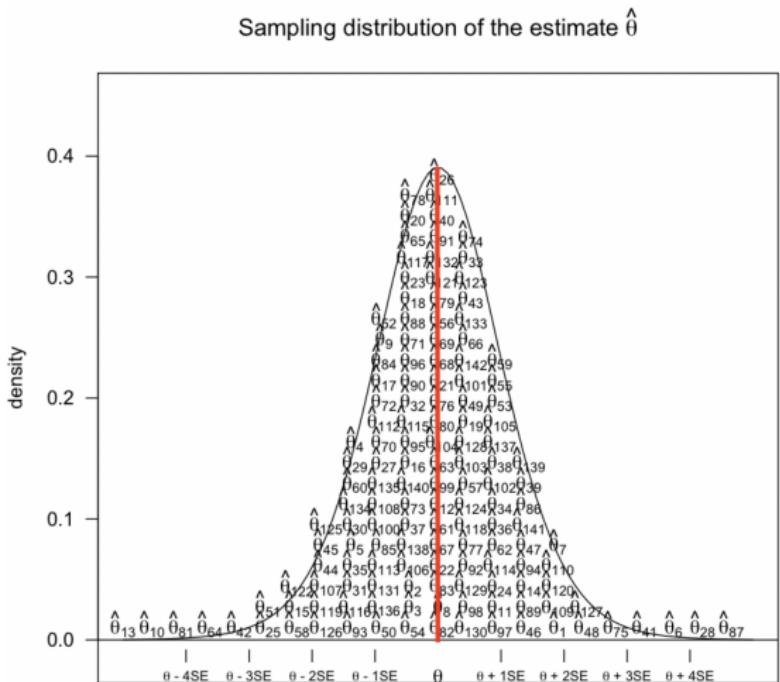
Confidence intervals

What is the whole tradeoff about? **Intuitively...**

- One can be 100% confident that their final grade in HAS 945 will be between 0% and 100%
- Yet, this is not giving us really any information (no precision on what it will be).
- if you want more precision, would you be 100% confident (as in , put your money on it) that your grade will be between 90% and 95%?
- most likely your confidence is reduced
- the more information you get, the more confident you may be about being tighter around a certain number

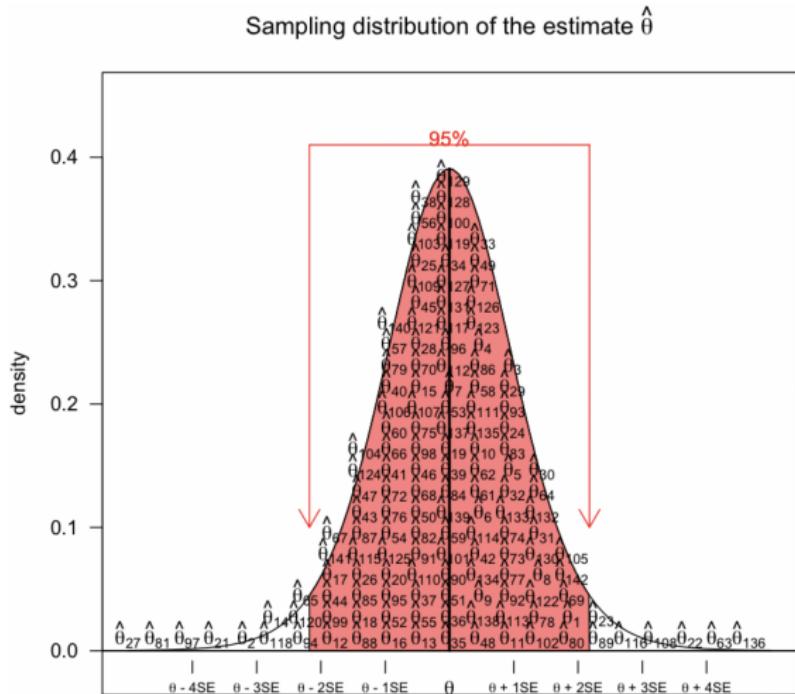
Understanding the Method

Our **one** sample estimate sits somewhere in the sampling distribution of the estimate.



Say $\alpha = 0.05$

Hence we are looking at $P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) = .95$

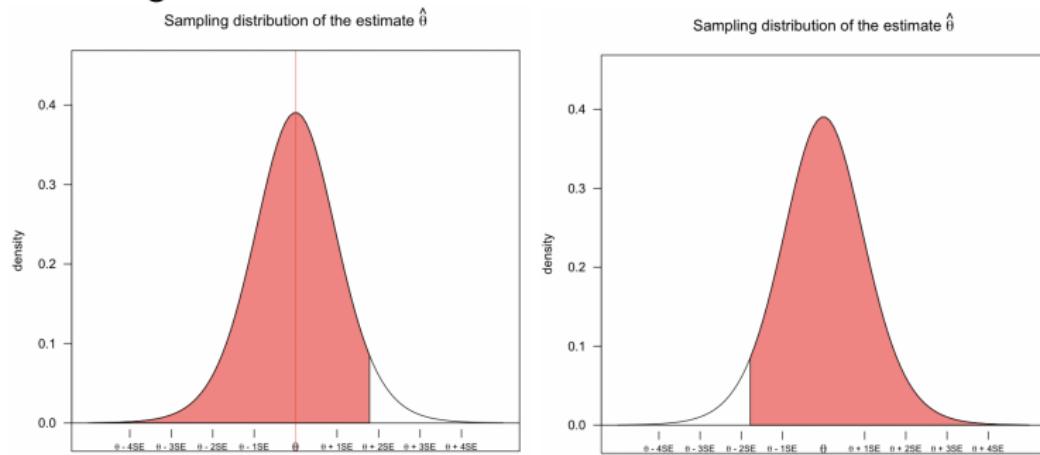


Why those 2?

You may have asked “*why the 2 that enclosed that shaded area and not others?*”

Recall, the **precision** idea is to have the **narrowest** possible interval

In previous example, the sampling distribution is symmetric, hence symmetric values give the narrowest, consider the following



What does a confidence interval construction look like?

Calculation of **confidence interval limits** usually has the form

Point Estimate \pm Margin of Error

and presented as

$$[\hat{\theta}_L, \hat{\theta}_U]$$

where $\hat{\theta}_L$ = Point Estimate - Margin of Error,
 $\hat{\theta}_U$ = Point Estimate + Margin of Error

Note that the CI has **two** components:

- The **precision** or uncertainty is indicated by the margin of error (MOE), the “give or take” or the range of predictions.
- The **expected coverage** is indicated by the confidence coefficient/level.

Use of Confidence Intervals in Reports

Reports for some standardised tests provide confidence intervals

Example: NAPLAN 2011 Report p317

Table R1.3_5: Gain in Reading Achievement for Students from Year 3 to Year 5, by State and Territory, 2008–2010 and 2009–2011.

	NSW	Vic	Qld	WA	SA	Tas	ACT	NT	Aust
2008–2010 Average gain (with 95% confidence interval)	83.9 ± 8.2	82.3 ± 8.1	97.6 ± 8.4	90.8 ± 8.8	76.0 ± 9.0	83.4 ± 10.7	87.6 ± 11.1	105.5 ± 27.7	86.9 ± 7.9
2009–2011 Average gain (with 95% confidence interval)	73.1 ± 9.4	73.3 ± 9.4	83.5 ± 9.5	84.7 ± 10.0	79.0 ± 10.1	81.2 ± 11.8	82.7 ± 12.6	81.1 ± 27.9	77.3 ± 9.2

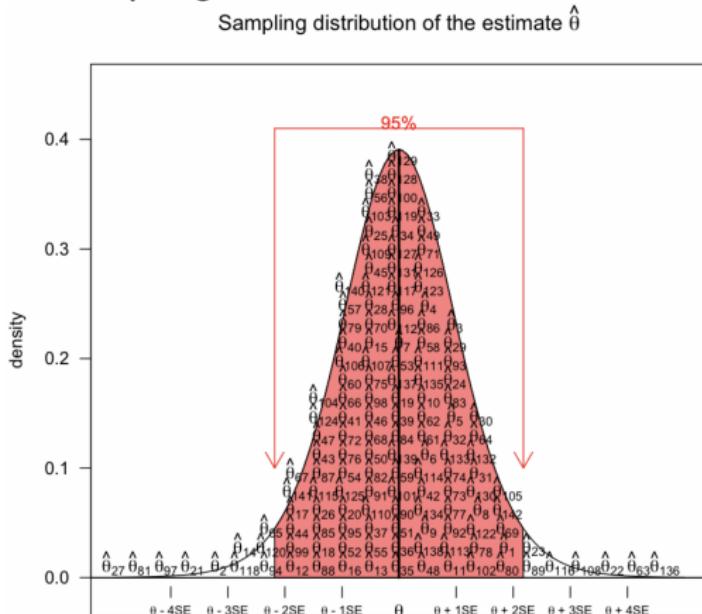
The confidence interval provided is for the specific jurisdictional gain and should not be used for comparisons between jurisdictions.

NAPLAN 2011 Report p317

Where do we get Margin of error from?

How will we know that the **chosen** MOE will get us the $1 - \alpha$ confidence? That is, that $P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) = 1 - \alpha$

Well, from the sampling distribution of the estimate $\hat{\theta}$. Think of



MOE from?

Pick one of the points in the graph that cover the desired % , say θ_L and see how far it is **in standard errors** from the $\mathbb{E}(\hat{\theta}) = \theta$

How far in SEs?...Well, if you know I live 5km away from uni, and blocks in Wollongong are in general 0.5km each you would do

$$\frac{\text{distance}}{\text{length of block}} = \# \text{blocks} \text{ so } \frac{5}{0.5} = 10$$

right? Then

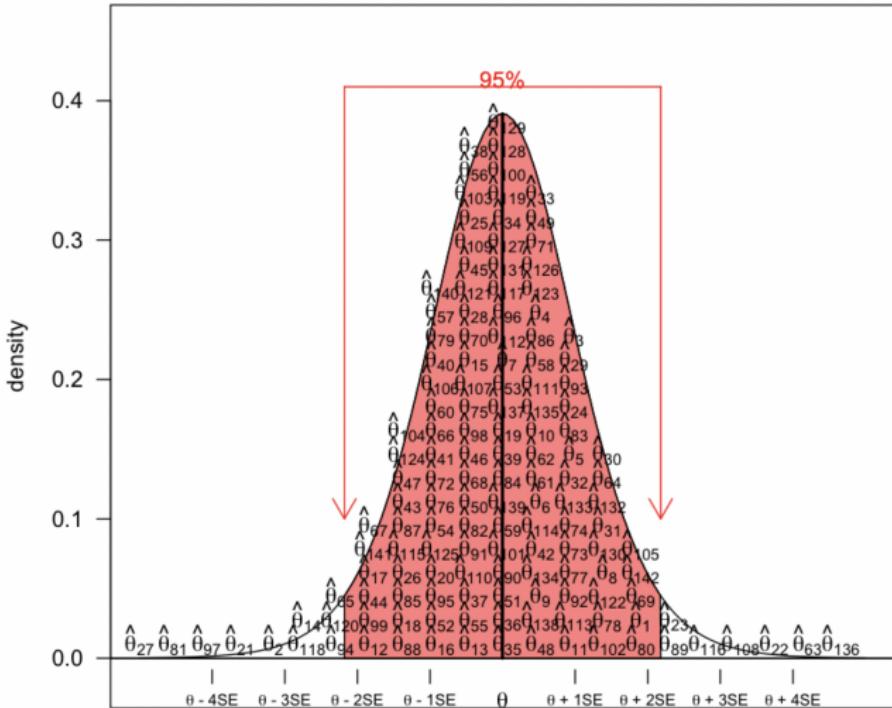
$$\frac{\text{distance}}{\text{length of block}} = \frac{\theta_L - \theta}{SE_{\text{estimator}}} = \# \text{SEs} = Distr_quantile_{\alpha/2}$$

the actual # of SEs will depend on the actual form of the sampling distribution

Note: we could have picked θ_U too, as they are symmetric to θ

How does it work on the Interval?

Sampling distribution of the estimate $\hat{\theta}$



Final form of MOE

So now we can see that if we are $\pm Distr_quantile_{\alpha/2}$ times the length of standard error $SE_{\hat{\theta}}$ away from θ we accumulate 95% of the distribution.

And hence if we are then $\pm Distr_quantile_{\alpha/2} \times SE_{\hat{\theta}}$ from $\hat{\theta}$ we can be 95% **confident** the interval covers θ

and hence the form for the Margin of error to use in calculating limits is

$$MOE = Distr_quantile_{\alpha/2} \times SE_{\hat{\theta}}$$

and then the limits are estimated by

$$\hat{\theta} \pm Distr_quantile_{alpha/2} \times SE_{\hat{\theta}}$$

Understanding the Method

We can label the 95% as the level of **confidence** because it specifies **how confident we can be that a CI includes θ**

We cannot talk of probability anymore because that applied to $P(\theta_L \leq \theta \leq \theta_U)$ knowing the values,

The estimate we obtained from **our** sample can be any one of the many many in the sampling distribution.

Since we do not know θ , we don't know whether our interval does or does not capture θ .

In the long run we expect 95% of the CIs will include θ ,

About 5% of CI in the long run will not include θ .

We **hope/bet** on our CI being one of the 95%.

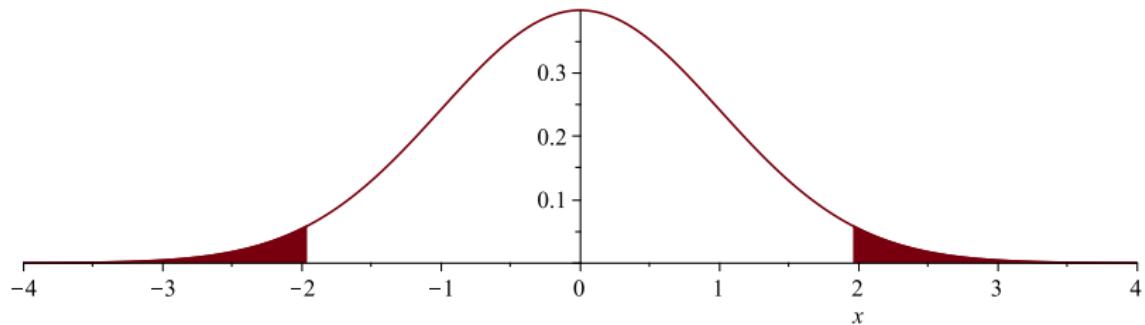
We generally take a random sample from a population to get some information. We estimate the mean μ by the sample mean \bar{x} . But \bar{x} is a single number (point estimate) and is almost certainly not exact. Often, we prefer an interval estimate, such as $\mu \in [3.4, 5.6]$.

Confidence Intervals

- A confidence interval contains plausible values for a parameter.
- How sure we are that the parameter is contained in the interval is the *confidence level*, most often 0.95.
- Many confidence intervals are of the form (point estimate) \pm (margin of error).

Confidence Intervals

The simplest case is when σ is known and μ is unknown. From standard tables, we find $P(-1.96 < z < 1.96) = 0.95$.



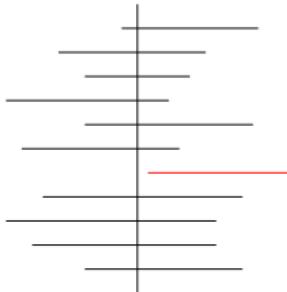
So for \bar{x} from $N(\mu, \sigma^2)$, we have

$$\begin{aligned} 0.95 &= P\left(-1.96 < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < 1.96\right) \\ &= P\left(\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}\right) \end{aligned}$$

Confidence Intervals

In other words, the interval contains μ with “probability” 0.95. This is the 95% confidence interval for μ .

In the long run, if 95% intervals are used for many samples, about 95% of the intervals will contain the population parameter.



- By the central limit theorem, we can apply this method to non-normal data as well, if n is large.
- Using a larger confidence interval, such as 0.99, gives a larger margin of error and a wider interval.

Confidence Intervals

So a $100(1 - \alpha)\%$ confidence interval for μ is $\bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$.

However, when μ is unknown, σ is usually unknown as well and estimated by the sample standard deviation s . Then for confidence intervals, we use the student's t distribution.

For instance with $df = 6$, we find on the t table that

$$P(T > 19.43) = P(T < -1.943) = 0.05 \\ \Rightarrow P(-19.43 < T < 1.943) = 0.90$$

Confidence Intervals

In general, when σ is unknown,

$$\begin{aligned}1 - \alpha &= P\left(-t_{n-1; \frac{\alpha}{2}} < \frac{\bar{x} - \mu}{s/\sqrt{n}} < t_{n-1; \frac{\alpha}{2}}\right) \\&= P\left(\bar{x} - t_{n-1; \frac{\alpha}{2}} \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{n-1; \frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right)\end{aligned}$$

So a $100(1 - \alpha)\%$ confidence interval for μ is $\bar{x} \pm t_{n-1; \frac{\alpha}{2}} \frac{s}{\sqrt{n}}$.

So now we can see that depending on what we know we have different $\pm Distr_quantile_{\alpha/2}$ multiplying the standard error $SE_{\hat{\theta}}$ away from θ we accumulate 95% of the distribution.

And hence if we have:

- Knowledge about the true σ we use the Normal distribution and hence $z_{\alpha/2}$ multiplier
- NO knowledge about the true σ we use s , and then the t distribution ($n-1$ df) and hence $t_{\alpha/2,n-1}$ multiplier

Confidence Intervals

Example. 8 samples of the benzene concentration in the air in mg per m³ are 2.2, 1.8, 3.1, 2.0, 2.4, 2.0, 2.1, 1.2. Thus, $n = 8$, $\bar{x} = 2.1$ and $s = 0.5372$. Assuming a normal population, construct a 90% confidence interval for μ .

Ans. From the t table with 7 df , we find $t_{7;0.05} = 1.895$.

$$\text{Lower bound: } 2.1 - 1.895 \frac{0.5372}{\sqrt{8}} = 1.74$$

$$\text{Upper bound: } 2.1 + 1.895 \frac{0.5372}{\sqrt{8}} = 2.46$$

The 90% confidence interval for μ is [1.74, 2.46]mg/m³.

Confidence Intervals

Example. A sample of 169 fish is randomly selected from a large population. Fish length X is distributed with $\mu = 50\text{cm}$ and $\sigma = 26\text{cm}$. Find the probability that the sample's mean is between 46 and 48cm.

Ans. $n = 169$, so what kind of distribution is the sample mean? (almost) normal. In that case, what are $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$?

$$\mu_{\bar{x}} = \mu = 50, \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{26}{\sqrt{169}} = \frac{26}{13} = 2$$

Confidence Intervals

$$\begin{aligned}P(46 \leq \bar{X} \leq 48) &= P\left(\frac{46 - 50}{2} \leq \frac{\bar{X} - 50}{2} \leq \frac{48 - 50}{2}\right) \\&= P(-2 \leq \bar{Z} \leq -1)\end{aligned}$$

From Z tables, $P(Z \leq -1) = 0.1587$ and $P(Z \leq -2) = 0.0228$.

$$\Rightarrow P(-2 \leq \bar{Z} \leq -1) = 0.1587 - 0.0228 = \textcolor{blue}{0.1359}.$$

other CI with similar form

When estimating proportions, say like for the p in a binomial, where X is the number of successes from n trials

The estimate of p , is $\hat{p} = \frac{X}{n}$ with standard error $se_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

and the sampling distribution of \hat{p} happens to be Normally distributed, hence use $Z_{\alpha/2}$ as the quantile for the CI.

the CI for a proportion is

$$\hat{p} \pm Z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

NOTE: This CI has **coverage** issues, hence a quick “fix” is to use

$\tilde{X} = X + 2$ and $\tilde{n} = n + 4$ hence $\tilde{p} = \frac{\tilde{X}}{\tilde{n}}$ instead of \hat{p} and n

Proportion of babies > 3kg

A doctor has interest in knowing the overall proportion of babies born > 3kg. Using the “babies” data.

Frequencies of Over3kg

Levels	Counts	% of Total	Cumulative %
Over	232	90.625 %	90.625 %
not over	24	9.375 %	100.000 %

$$\hat{p} \pm z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} = 0.90625 \pm 1.96 \times \sqrt{\frac{0.90625 \times 0.09375}{256}}$$

gives the 95% CI (0..8705, 0.9420)

Using $\tilde{X} = X + 2$ and \tilde{n} , we get $\tilde{X} = 232 + 2$ and $\tilde{n} = 256 + 4$, and $\tilde{p} = \frac{234}{260} = 0.9$

$$\tilde{p} \pm z_{\alpha/2} \times \sqrt{\frac{\tilde{p}(1 - \tilde{p})}{\tilde{n}}} = 0.9 \pm 1.96 \times \sqrt{\frac{0.9 \times 0.1}{260}}$$

gives the 95% CI (0.8635, 0.9365)

Confidence Intervals for a Population Mean, σ known

A $(1 - \alpha)\%$ confidence interval for the mean, μ is of the form

sample estimate $\pm Distr_quantile_{\alpha/2} \times$ standard error

A Confidence Interval for a population mean μ , with σ known, is given by

$$\bar{x} \pm z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

- the margin of error (MOE) = $z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$
- $z_{\alpha/2}$ is the $\alpha/2$ percentile of the standard normal distribution
- α is level of significance, commonly 0.05 or 5%
- Then $C = 100(1 - \alpha)\%$, so if $\alpha = 0.05$,
 $C = 100 \times (1 - 0.05) = 95\%$
- Other popular choices for C are 90%, 98%, and 99%

Confidence Intervals for a Population Mean, σ unknown

The population standard deviation σ is usually **unknown**, the standard error of the mean is usually estimated as

$$se(\bar{X}) \approx \frac{s}{\sqrt{n}}$$

where s is the sample standard deviation.

A $(1 - \alpha)\%$ **confidence interval** for the **mean**, μ is of the form

sample estimate $\pm Distr_quantile_{\alpha/2} \times$ standard error

and we know that when we estimate the σ by s we have a t_{n-1} distribution

Confidence Intervals for a Population Mean, σ unknown

A Confidence Interval for a population mean μ , with σ **unknown**, is given by

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

- the *margin of error* (MOE) = $t_{n-1,\alpha/2} \times \frac{s}{\sqrt{n}}$
- $t_{n-1,\alpha/2}$ is the $\alpha/2$ percentile of the t_{n-1} distribution.
- the value of the *multiplier* $t_{n-1,\alpha/2}$ depends on the level of confidence $C = 100(1 - \alpha)\%$ and the sample size n
- the *degrees of freedom* $df = n - 1$

Babies data

Data from a study on babies' birthweight `bweight` by sex gender.

```
babies.df <- read.csv('babies.csv', header=T, stringsAsFactors = T)

by(babies.df$bweight, babies.df$gender, function(x) c(mean=mean(x), sd=sd(x)) )

babies.df$gender: Male
  mean      sd
3.4430252 0.3305685
-----
babies.df$gender: Female
  mean      sd
3.5316423 0.4284853
```

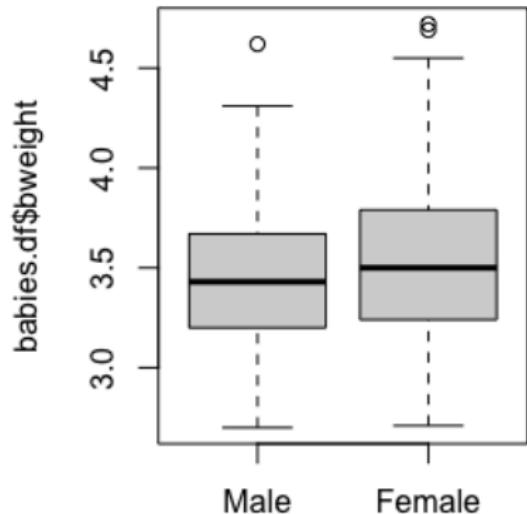
Exercise:

- 1 Calculate a 95% CI for the population mean birth weight for
 - Male babies;
 - Female babies.
- 2 Plot the two CIs using a common scale.

Descriptive graphs of birthweight by sex

Always plot the data: in R:

```
plot(babies.df$bweight babies.df$gender, ylab='Bweight', xlab='Gender', las=1)
```



95% CI for means of birthweights

Male

$$\bar{x} = 3.443, s = 0.331, n = 119$$

$$\bar{x} \pm t_{118,\alpha/2} * 0.331 / \sqrt{119}$$

$$\alpha = 0.05$$

$$t_{118,0.025} = -1.98$$

Female

$$\bar{x} = 3.531, s = 0.428, n = 137$$

$$\bar{x} \pm t_{136,\alpha/2} * 0.428 / \sqrt{137}$$

$$\alpha = 0.05$$

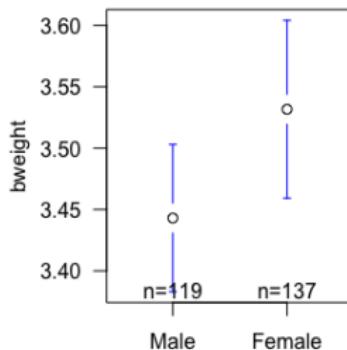
$$t_{136,0.025} = -1.978$$

The CI (3.3832, 3.503)

The CI (3.459, 3.604)

in R:

```
by(babies.df$bweight, babies.df$gender, gmodels::ci)  
gplots::plotmeans(bweight ~ gender, data=babies.df, connect=F, las=1)
```



CI for Difference of Two Population Means

$$\text{estimate} \pm \text{quantile} \times \text{se}_{\text{estimate}}$$

If we are happy to assume equal variances (i.e. $\sigma_1^2 = \sigma_2^2$) then we can use the formula

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$\bar{x}_1 - \bar{x}_2 \pm t_{n_1+n_2-2,\alpha/2} \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

where s_p is the 'pooled' standard deviation.

Example

Some researchers were interested in the effect of hangovers amongst college students. Students were asked whether their parents suffered from alcohol problems and asked to rate the severity and duration of their own hangovers on a 13-point scale, with 13 being the most severe. 1227 students were contacted and the data are shown below. Calculate a 95% confidence interval.

Group	Sample size	Mean	Standard deviation
Parental alcohol problems	$n_1 = 282$	$\bar{x}_1 = 5.9$	$s_1 = 3.6$
No parental alcohol problems	$n_2 = 945$	$\bar{x}_2 = 4.9$	$s_2 = 3.4$

We have the pooled standard deviation as:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(282 - 1)3.6^2 + (945 - 1)3.4^2}{282 + 945 - 2} = 3.45^2$$

So

$$\text{s.e.}(\bar{x}_1 - \bar{x}_2) = \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} = \sqrt{3.45^2 \left(\frac{1}{282} + \frac{1}{945} \right)} = 0.23$$

Finally we have $t_{n_1+n_2-2,\alpha/2} = t_{282+945-2,2.5\%} = 1.96$ and the 95% confidence interval is:

$$5.9 - 4.9 \pm 1.96 \times 0.23 = (0.54, 1.46) \text{ symptoms}$$

Note that this interval is entirely above zero so we might conclude that those with parental alcohol problems tend to have more severe hangovers.

CI for difference between 2 proportions

Just like for 2 means, we can compare proportions in 2 groups

From the babies example: Over 3kg males X_m and females X_f , and total per sex n_m , and n_f

we are interested to see whether or not proportions differ by sex.

Estimate of p_m and p_f , are $\hat{p}_m = \frac{X_m}{n_m}$ and $\hat{p}_f = \frac{X_f}{n_f}$ respectively

the difference $\hat{p}_m - \hat{p}_f$ has standard error

$$se_{\hat{p}_m - \hat{p}_f} = \sqrt{\frac{\hat{p}_m(1 - \hat{p}_m)}{n_m} + \frac{\hat{p}_f(1 - \hat{p}_f)}{n_f}}$$

CI diff 2 props (cont.)

The sampling distribution of $\hat{p}_m - \hat{p}_f$ happens to be Normally distributed, hence use $Z_{\alpha/2}$ as the quantile for the CI.

the CI for a difference in proportions is

$$\hat{p}_m - \hat{p}_f \pm Z_{\alpha/2} \times \sqrt{\frac{\hat{p}_m(1-\hat{p}_m)}{n_m} + \frac{\hat{p}_f(1-\hat{p}_f)}{n_f}}$$

NOTE: This CI has **coverage** issues, hence a quick “fix” is to use

$$\tilde{X}_m = X_m + 1 \text{ and } \tilde{n}_m = n_m + 2 \text{ hence } \tilde{p}_m = \frac{\tilde{X}_m}{\tilde{n}_m}$$

$$\text{and } \tilde{X}_f = X_f + 1 \text{ and } \tilde{n}_f = n_f + 2 \text{ hence } \tilde{p}_f = \frac{\tilde{X}_f}{\tilde{n}_f}$$

example for babies birthweight

$$\hat{p}_m = \frac{109}{119} = 0.916 \text{ and } \hat{p}_f = \frac{123}{137} = 0.8978$$

$$\hat{p}_m - \hat{p}_f \pm Z_{\alpha/2} \times \sqrt{\frac{\hat{p}_m(1 - \hat{p}_m)}{n_m} + \frac{\hat{p}_f(1 - \hat{p}_f)}{n_f}}$$

$$0.916 - 0.8978 \pm 1.96 \times \sqrt{\frac{0.9159(0.0841)}{119} + \frac{0.8978(0.1022)}{137}}$$

And the 95% CI for the difference $p_m - p_f$ is
 $(-0.05296, 0.08927)$

Using the fix for each group $\tilde{X}_m = X_m + 1$ and $\tilde{X}_f = X_f + 1$ and adding 2 to each n_m, n_f

the 95% CI is $(-0.05569, 0.08970)$

Not all CIs are built the same way

if you recall, if s is the estimate of the standard deviation σ , then

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$$

Hence if we have the 2 quantiles $\chi_{n-1,\alpha/2}^2$ and $\chi_{n-1,1-\alpha/2}^2$ then

$$P\left(\chi_{n-1,\alpha/2}^2 \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi_{n-1,1-\alpha/2}^2\right) = 1 - \alpha$$

doing some algebra

$$P\left(\frac{(n-1)s^2}{\chi_{n-1,1-\alpha/2}^2} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi_{n-1,\alpha/2}^2}\right) = 1 - \alpha$$

CI for one variance (or standard deviation)

Hence

the $100 \times (1 - \alpha)\%$ confidence interval for Variance σ^2

$$\left(\frac{(n-1)s^2}{\chi_{n-1,1-\alpha/2}^2}, \frac{(n-1)s^2}{\chi_{n-1,\alpha/2}^2} \right)$$

and

the $100 \times (1 - \alpha)\%$ confidence interval for Standard deviation σ

$$\left(\sqrt{\frac{(n-1)s^2}{\chi_{n-1,1-\alpha/2}^2}}, \sqrt{\frac{(n-1)s^2}{\chi_{n-1,\alpha/2}^2}} \right)$$

Common misinterpretation of CIs

There are many **misinterpretations** of Confidence Intervals.

Two of the most common are:

- “95% of the observed values will be within the limits”;
- “There is a 95% chance that the population parameter is within the interval”

Discuss why these are incorrect interpretations.

DO NOT fall into this trap!!!

More on interpretation of CIs

More generally, in terms of parameters (μ , p , σ , $\mu_a - \mu_b$, $p_a - p_b$)

- we are 95% **confident** that our CI contains the true parameter
- our CI is a **range of plausible values** for μ (p , σ , etc.). Values outside the CI are considered relatively implausible.
- the choice of 95% sets a standard that says we will be right 0.95 of the time, or **19 times out of 20**.
- the CI is random in the sense that if we repeated the entire study, we would have a **different** random sample and hence a **different CI**.
- if the experiment was repeated many times and a CI calculated for each, in the long run 95% of the intervals would include μ (p , σ , etc.)
- The interval helps us **make a decision** as we get an idea of the magnitude and we can be $100 \times (1 - \alpha)$ **confident**.

What happens to the MOE when:

- the required confidence level increases?
- the overall variability increases?
- the sample size changes?

The *margin of error* (the part after \pm) MOE for a CI for parameters μ and for a proportion p is

$$MOE = Distr_quantile_{1-\alpha/2} \times SE_{estimator}$$

For a mean: $MOE = Distr_quantile_{1-\alpha/2} \times \sqrt{\frac{\sigma}{n}}$.

For a proportion: $MOE = z_{1-\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

So far, we've calculated \bar{x}, \hat{p} for a given sample of size n , and evaluated MOE . Now, what if we are designing a study and need to decide on a sample size?

Sample Size For Estimating Means

If standard deviation σ is known, the margin of error MOE of a mean CI is $MOE = z_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}}$. Suppose we set target amount of precision MOE and level of confidence $(1 - \alpha) \times 100\%$ that we want. Then,

$$MOE = z_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

$$\frac{MOE}{\sigma z_{1-\alpha/2}} = \frac{1}{\sqrt{n}}$$

$$n \geq \frac{\sigma^2 (z_{1-\alpha/2})^2}{MOE^2}$$

- There is no “default” value for σ .
- Use prior literature, expert opinion, pilot study, etc.

Example: Length of Trips

We wish to estimate the population mean length of vacation trips. We believe that the population standard deviation $\sigma \approx 30$ days, and we want to estimate it to the nearest 10 days at 90% confidence. How big a sample do we need?

$$n \geq \frac{\sigma^2(z_{1-\alpha/2})^2}{MOE^2} = \frac{30^2 \times 1.645^2}{10^2} = 14.803$$

Sample Size Calculations for a proportion

The *margin of error* (the part after \pm) MOE of a proportion CI is

$$MOE = z_{1-\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

So far, we've calculated \hat{p} for a given sample of size n , and evaluated MOE . Now, what if we are designing a study and need to decide on a sample size?

- Too small a sample size will give us too little precision.
- Too large a sample size will be wasteful.

A little bit of algebra

Suppose we set target amount of precision d and level of confidence $(1 - \alpha) \times 100\%$ that we want. Then,

$$MOE = z_{1-\alpha/2} \times \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

$$\frac{MOE}{z_{1-\alpha/2}} = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

$$\left(\frac{MOE}{z_{1-\alpha/2}}\right)^2 = \frac{\hat{p}(1 - \hat{p})}{n}$$

$$n = \frac{\hat{p}(1 - \hat{p})}{\left(\frac{MOE}{z_{1-\alpha/2}}\right)^2} = \frac{\hat{p}(1 - \hat{p})(z_{1-\alpha/2})^2}{MOE^2}$$

Sample Size Formula for Estimating Proportions

$$n \geq \frac{p(1-p)(z_{1-\alpha/2})^2}{d^2}$$

We don't know p before we gather the data. What do we do?

- Make an educated guess (e.g., prior literature, expert opinion, pilot study, etc.).
- Use the conservative value of $\frac{1}{2}$:

$$n \geq \frac{\frac{1}{2}(1 - \frac{1}{2})(z_{1-\alpha/2})^2}{d^2} = \frac{(z_{1-\alpha/2})^2}{4d^2}.$$

Example: Alzheimer's Test Sensitivity

The estimated sensitivity of the Alzheimer's test was 96.89%, and the 95% CI margin of error was 1.60 percentage points. Suppose that we want a new study to estimate it to within 1% instead with 95% confidence, i.e., $MOE = 0.01$. Then, using the result of the previous study, we need

$$n \geq \frac{p(1-p)(z_{1-\alpha/2})^2}{d^2} = \frac{(0.9689 * (1 - 0.9689) * 1.96^2}{0.01^2} = 1157.94$$

If we had no idea what p was, we could use the conservative value of $\frac{1}{2}$:

$$n \geq \frac{(z_{1-\alpha/2})^2}{4d^2} = \frac{1.96^2}{4 \cdot 0.01^2} = 9603.64$$

Sample Size Calculation: Summary

We are trading-off among four variables:

σ : population standard deviation

MOE : margin of error

n : sample size

$(1 - \alpha) \times 100\%$: level of confidence

Holding other things fixed,

- $MOE \propto 1/\sqrt{n}$: doubling the sample size reduces the margin of error by a factor of 1.414.
- $n \propto 1/MOE^2$: to halve the margin of error, the sample size needs to be *quadrupled*.
 - There is a diminishing returns effect.
- $\sigma \propto MOE$: if the population standard deviation is doubled, the margin of error is also doubled.
 - $n \propto \sigma^2$: if population standard deviation is doubled, to keep the margin of error the same, the sample size must be *quadrupled*.

Software packages can often provide more accurate results with more sophisticated calculations.

MATH55

Alberto Nettel Aguirre

NIASRA
School of Mathematics and Statistics
University of Wollongong

May 19, 2023

Introduction to Null Hypothesis Significance Testing

- Statistical inference uses **sample statistics** to make decisions and predictions about **population parameters**.
- The aim of **null hypothesis significance testing** (NHST) is to decide whether there is **sufficient evidence** from a sample to support the posed hypothesis and hence infer conclusions about the population
- Some examples:
 - testing if the mean heart rate in two different groups differs
 - testing if socioeconomic status is associated with academic performance

Intuitive Hypothesis testing

Humour me for a minute.

- Say I start class complaining that I earn at most **\$1,000** a month and you believe me,
- Class finishes and I offer you a lift,
- we get to my car and it is the latest **BMW fully loaded model**.

I tell you that on the way to drop you off I need to run some errands, these being:

- Go to my **private hangar** to check on
- my **private jet**, because
- I am taking a break at my **beach house** in Lord Howe Island.
- I show you the deeds for all the above.

Intuitive Hypothesis testing

What would be the first thing you think?

Don't be shy....

Big fat liar!!!

right?

You just performed, though not “formally”, your own **hypothesis test**

Intuitive Hypothesis testing

You are thinking: *What??? I did what??*

- You **assumed** I did earn only **\$1,000** a month,
- you **observed** the BMW, hangar, jet and photos of the beach house (and all deeds)
- Your head went into shock and thought:

"Yeah right!!, you own all of these and you only earn \$1,000 a month???... Highly unlikely"

You **inferred** I must be lying about the salary...right?

Well, that is all that **hypothesis testing** is.

Introduction to Null Hypothesis Significance Testing

Has anyone ever been called to jury duty?

Consider an everyday example of a person who has been charged for committing a crime and is being tried in court.

- Based on the available evidence, the jury will make one of two possible decisions:
 - ➊ **Declare** the person as not guilty
 - ➋ **Declare** the person as guilty
- At the start of the trial, the person is presumed **NOT** guilty.
- It is up to the prosecutor to **give evidence** that refutes the assumption of NOT guilty.

Two Hypotheses, Formally

In **inferential statistics**,

- we call the first option (declare not guilty) the *null hypothesis*
- we call the second option (declare guilty) the *alternative hypothesis*

Hypotheses are statements about the parameter(s) or distribution or some quality (such as independence) of the **population**:

The **null hypothesis** is a claim or statement about a population parameter that is assumed to be true until it is not believed to be so.

The **alternative hypothesis** is a claim about a population parameter that will be taken as conclusion when null hypothesis is not supported.

We find support (or lack of it) for the hypothesis using information from the **sample**

Null Hypotheses

The **null hypothesis**, H_0 ,

- is a neutral statement of the form “ = a default value” or “no difference”, “no change”, “no improvement” etc;
- Eg.s about one parameter: $\mu = 100g$; $p = 0.5$;
- Eg.s about two parameters: $\mu_1 = \mu_2$; $p_1 - p_2 = 0.2$;
- Eg.s about the underlying population distribution such as Normal or Uniform
- Some quality about the relationship or association occurring in the population - such as independence
- is retained unless there is sufficient evidence to reject it (analogous to NOT guilty unless declared guilty).

Alternative Hypotheses

The **alternative hypothesis**, H_1 (or H_A),

- specifies departure from H_0 , possibly in a particular direction;
- typically involves not a specific value; $<$ or $>$ a specific value; or a change, or a difference;
- e.g. about one parameter: $\mu \neq 100g$; $\mu > 100g$; $p \neq 0.5$;
- e.g. about two parameters: $\mu_1 - \mu_2 \neq 0$; $p_1 - p_2 > 0.2$;
- e.g. about the underlying population distribution such as deviation from Normal or NOT Uniform
- Some quality about the relationship or association NOT occurring in the population - such as NOT independent
- is usually what a researcher is trying to justify.

5 Steps in Testing Hypotheses

Testing Hypotheses (5 steps):

- ① Set up the Null and alternative hypotheses
- ② Select a level of significance
- ③ Select the appropriate statistical test and state decision rule
- ④ Perform experiment - collect data, check assumptions of test and calculate test statistic, interval estimates and p -value
- ⑤ Make decision and draw conclusions

Set up the Null and alternative hypotheses

There are always two hypotheses.

- The first is the *null hypothesis* (H_0). This usually says that nothing “different” is happening.
 - e.g. That there is no relationship , that the difference between 2 parameters is 0, etc.
 - This one most always carries **the equality part** “=” or “ \leq ” or “ \geq ”
 - or some specific statement,
 - as we need to assume something specific to then set stage and compare.
- The *alternative hypothesis* (H_A) is the research hypothesis. The researcher suspects that the status quo belief is incorrect
 - e.g.that there is **indeed** a relationship, that there is a difference, etc.

Set up the Null and alternative hypotheses

The researcher needs to be quite sure that the observed evidence **does not support the null**, before they reject the null hypothesis in favour of the alternative. Lack of support for the null hypothesis is obtained by **showing how unlikely the observed is under the null**.

e.g. In a trial situation the hypotheses are:

H_0 : Defendant is NOT guilty. vs. H_A : Defendant is guilty.

In my silly \$1,000 salary example

H_0 : I earn $\leq \$1,000$ (not a liar) vs. H_A : I earn more than \$1,000 (a liar).

Level of Significance α

The **significance level**, α , is the **tolerated** conditional probability of rejecting the null hypothesis when it is true.

This is our cutoff for deeming “**unlikely**”

The **confidence level**, $1 - \alpha$, is the conditional probability of not rejecting the null hypothesis when it is true.

- α is typically set to 0.05 or smaller.
- α is typically set ahead of time, prior to data collection. At time of statistical planning.

Statistical Test

When using the significance level approach

- The decision in a hypothesis test is based on a summary of the observed data contrasting with the hypothesis. This summary is called the *test statistic*.
- the *test statistic* is used to find how (un)likely the observed would be under the null.
- There are many different test statistics and the one used depends on the situation.
- However, many of them take the form

$$\frac{\text{observed} - \text{hypothesised}}{\text{standard error of observed}}$$

Decision Rule

- In order to decide if the results observed could be still likely or unlikely under the null, the following question is asked:

Assuming the null hypothesis is true, how likely are we to observe a test statistic at least as extreme as the one we have observed ?

p-value

The *p*-value is computed by *assuming the null hypothesis is true*, and then asking how likely we would be to observe results at least as extreme as we have under that assumption.

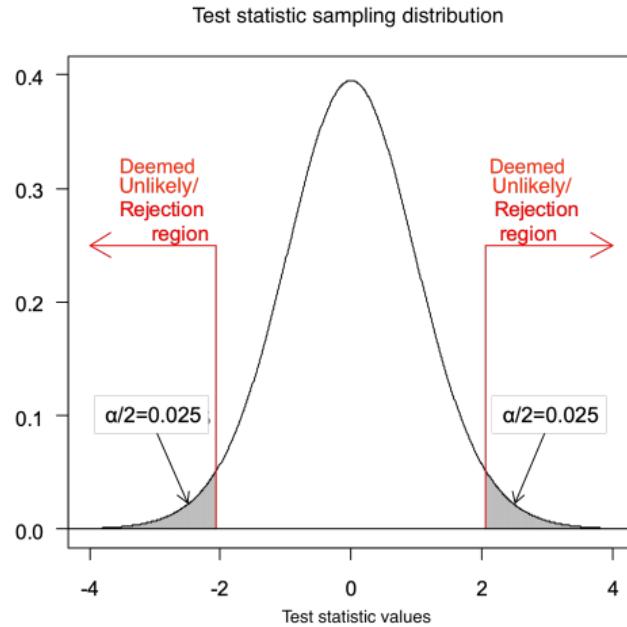
- p*-value will be compared to the significance level α .
- The p*-value does not give the probability that the null hypothesis is true

Test Statistics and p -value

- The test statistic is obtained based on the observed data
- p -value is the conditional probability having a test statistic at least as extreme as the observed one.
 - Such conditional probability is based on the sampling distribution of the test statistic
 - which is informed by the observed data.

Test Statistics and p -value

One example could be



Make Decision

- Once we know how (un)likely the observed test statistic is, we face two choices.

Choice 1: The p -value is not small enough to convincingly rule the observed as “unlikely”, so we *fail to reject the null hypothesis*.

Choice 2: The p -value is small enough to convincingly rule the observed as “unlikely”, so we *reject the null hypothesis* and *conclude along* what the alternative hypothesis poses.

- A p -value of less than the cut off point (the *level of significance* (α)) is considered small enough for “unlikely”, and hence to reject the null hypothesis.

Make decision

- Courtroom example:

Choice 1: There is **not enough** evidence to rule out that the defendant is NOT guilty so defendant is **declared not guilty**

Choice 2: There is **enough** evidence to rule out the possibility the defendant is NOT guilty so defendant is **declared guilty.**

- My silly example:

Choice 1: There is **not enough** evidence to rule out that I earn only (not more than) \$1,000 so you believe me and feel sorry for me, **declared not a liar.**

Choice 2: There is **enough** evidence to rule out the possibility that I only earn \$1,000 and I am **declared a liar.**

Video: What is a p -value?

Video: Data demystified:

Statistical Significance and p -Values Explained Intuitively (8:56 mins)

<https://youtu.be/DAkJhY2zQ3c>



Type I and Type II Errors: Definition

When we draw a conclusion there are two scenarios and hence two possible errors we can make:

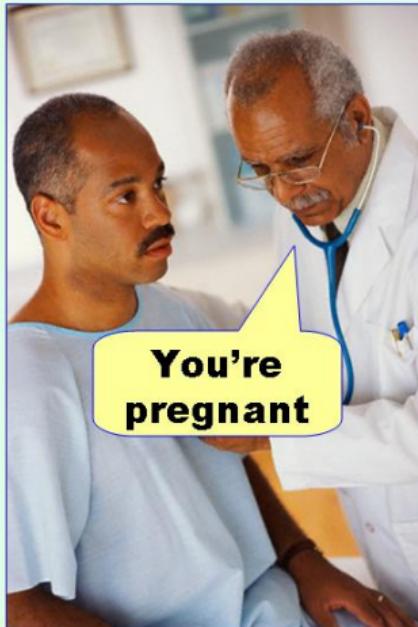
Decision	Reality /Truth	
	H_0 is true	H_0 is false
Reject H_0	Type I Error False Positive	Correct Decision
Do not reject H_0	Correct Decision	Type II Error False Negative

$$\begin{aligned} P(\text{Type I error}) &= P(\text{Reject } H_0 | H_0 \text{ true}) \\ &= \alpha \\ &= \text{level of significance} \end{aligned}$$

$$\begin{aligned} P(\text{Type II error}) &= P(\text{Do Not Reject } H_0 | H_0 \text{ false}) \\ &= \beta \end{aligned}$$

Type I and Type II Errors: Comic

Type I error
(false positive)



Type II error
(false negative)



Power

When the alternative hypothesis is true, then probability of making the correct decision is called the power of a test

$$\text{Power of the test} = 1 - \beta$$

Remember complement probability: $P(A^c) = 1 - P(A)$

Since $\beta = P(\text{Do Not Reject } H_0 | H_0 \text{ false})$

then $1 - \beta = 1 - P(\text{Do Not Reject } H_0 | H_0 \text{ false}) = P(\text{Reject } H_0 | H_0 \text{ false})$

Basically the power of a test is the probability that the test will lead to rejection of the null hypothesis (correctly).

We can make $\alpha = 0$ by never rejecting H_0 but then $\beta = 1$ and vice versa.

Properties

The relationship between n , α and β can be explained as follows (considering other issues like variability, and value for alternative fixed)

n	α	β
↑	fixed	↓
fixed	↑	↓
↑	↓	fixed

- Type I and Type II errors are related. A decrease in the probability of one type generally results in an increase in the other.
- For tests with the same level of significance α , β decreases as n increases.
- Equivalently, for tests with the same α , the power increases with n : more information gives more power.
- An increase in the sample size, n , will reduce both α and β simultaneously.

Population Mean

It is very common to want to test whether the mean of a population is equal to a specified value.

or compare between 2 or more groups.

How?

We use our knowledge of the **sampling distribution of \bar{X}** to test the proposed hypothesis.

Quick reminder

If X is a random variable,

$$\mathbb{E}[X] = \mu$$

and the **sample estimate** of μ is $\bar{X} = \frac{\sum X_i}{n}$

Central Limit Theorem

The sampling distribution of the mean has a *special* form, and the **central limit theorem (CLT)** tells us which.

Let X_1, \dots, X_n be a random sample of size n from a population of interest.

Each X_i has expected value $\mathbb{E}[X_i] = \mu$ and variance given by σ^2 .

The distribution of the sample mean \bar{X} **approaches a normal distribution** as n (sample size) becomes large.

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Normal Distribution

If a variable X follows a normal distribution with a mean of μ and a standard deviation of σ , then **standardisation** of a normal distribution, or to convert a normal distribution to the standard normal distribution.

$$\frac{X - \mu}{\sigma}$$

follows a standard normal distribution: $\frac{X - \mu}{\sigma} \sim N(0, 1)$

Therefore, since $\bar{X} \sim N(\mu, (\sigma/n)^2)$

$$Z := \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Why do we care?

Recall

In hypothesis testing we need to get a sense of how likely (under the null) it is to observe what we observe (or more extreme)

If we are to test means, we will need a way to know how likely our observed mean is under the null...

See where I am going?

Test of the Mean if σ known

We use a **one-sample z-test** to test whether the population mean is (or is not) significantly different from some hypothesised value.

$H_0: \mu = \mu_0$ against one of

$$H_{1a}: \mu \neq \mu_0, \quad \text{or} \quad H_{1b}: \mu < \mu_0 \quad \text{or} \quad H_{1c}: \mu > \mu_0.$$

NB: μ_0 means the value of μ specified by H_0 .

Based on a random sample X_1, \dots, X_n from $N(\mu, \sigma^2)$,
where σ is **known**, we use the sample mean \bar{X} as our observed, and hence the
test statistic

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

tells us how far the observed is from the expected and allows us to calculate the probability of observing it under the null.

Test of the Mean if σ known

Assuming H_0 true, then $\mu = \mu_0$, and then the sample mean approximately follows a Normal distribution with expected value $\mathbb{E}(\bar{X}) = \mu$ and Variance σ^2/n ,

that is $\bar{X} \sim N(\mu, \sigma^2/n)$

then the test statistic Z approximately follows a standard normal distribution:

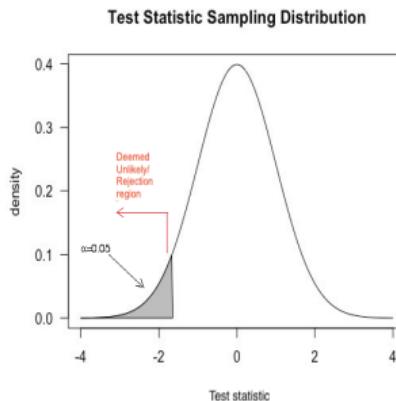
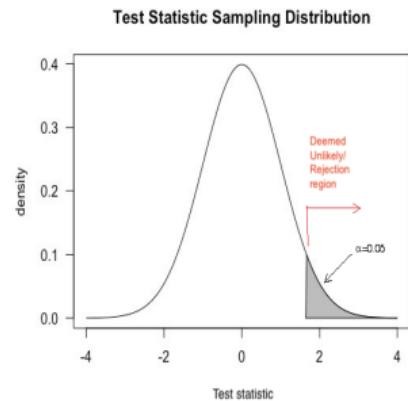
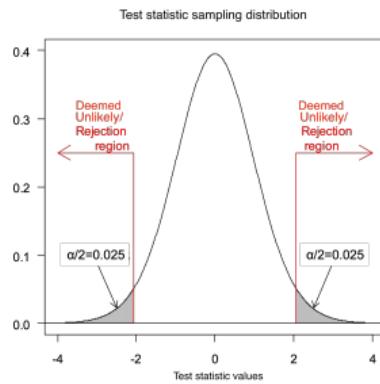
$$Z \sim N(0, 1) \quad \text{under } H_0$$

If H_0 is **not** true, then Z will not follow a standard normal distribution.

Decision (i.e. how likely?)

Our **decision rule** will depend on the choice of H_1

- We reject H_0 if Z or something more extreme is unlikely to happen under H_0
- ‘extreme’ gets defined by the direction(s) suggested by H_1
- remember we look for $P(Z \text{ as observed or more extreme})$
- Which direction on graph?



'Too extreme'

- If $H_0: \mu = \mu_0$ is true, we expect \bar{X} to be close to μ_0 , and Z should be close to 0.
- If H_1 says
 - $\mu \neq \mu_0$, then Z is **more extreme** if it is too much above 0 or too much below 0.
 - $\mu < \mu_0$, then Z is **more extreme** if it is too much below 0.
 - $\mu > \mu_0$, then Z is **more extreme** if it is too much above 0.

p-value

The *p*-value describes how (un)likely the observed data would be if H_0 were true.

For a **two-tailed test** the p-value is

- the probability of getting a test statistic as extreme (either end of the distribution) as this statistic

For a **one-tailed test** the p-value is

- the probability of getting this statistic or greater; OR
- the probability of getting this statistic or smaller

p-value

Let z be the observed test statistic.

If $H_1 : \mu \neq \mu_0$ (two-tailed),

$$p\text{-value} = P(Z \geq |z| \text{ or } Z \leq -|z|) = 2 \times P(Z \geq |z|)$$

If $H_1 : \mu < \mu_0$ (one-tailed -lower),

$$p\text{-value} = P(Z \leq z)$$

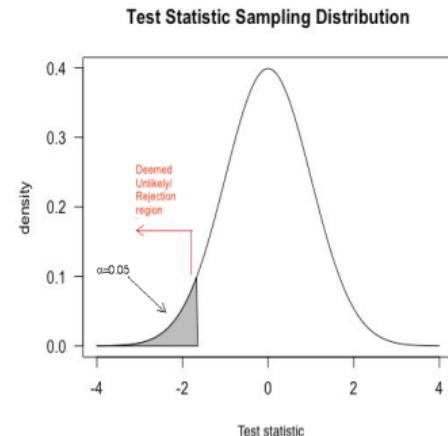
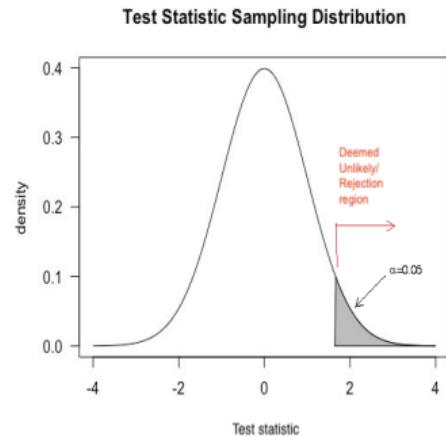
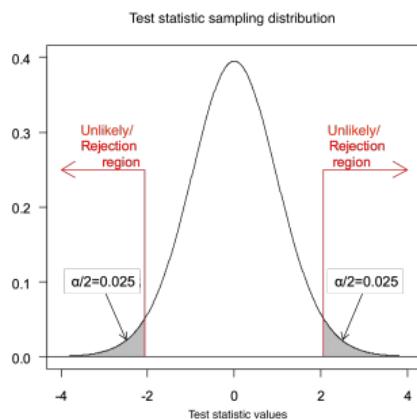
If $H_1 : \mu > \mu_0$ (one-tailed -upper),

$$p\text{-value} = P(Z \geq z)$$

Decision Rule

Given a **significance level** α , we reject $H_0 : \mu = \mu_0$ if $p\text{-value} < \alpha$.

The calculation of the p -value depends on if H_1 is two-tailed, one-tailed lower or one-tailed upper.



Keep in mind

We are pretty much modelling how $\mathbb{E}(X)$ is behaving,
we are saying

$$\mathbb{E}(X) = \mu_0$$

where μ_0 is a constant.

so we could be saying that the xs we observe are the result of the following model

$$X = \mu_0 + \varepsilon$$

Where $\varepsilon \sim N(0, \sigma^2)$, and recalling expected value properties (e.g. lab6)

$$\mathbb{E}(X) = \mu_0 + \mathbb{E}(\varepsilon) = \mu_0$$

OR, if $X - \mu_0 = \varepsilon$ then $\mathbb{E}(X - \mu_0) = \mathbb{E}(\varepsilon) = 0$

Test of the Mean if σ unknown

In **most** applications, the population standard deviation σ is **unknown**, so it must be estimated from the data.

In this case, the test statistic becomes

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

where s is the sample standard deviation.

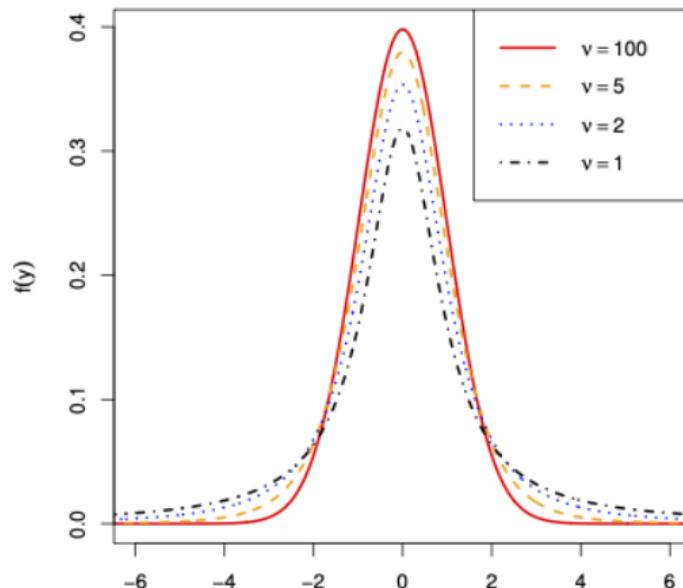
It can be shown that T approximately follows a **student's t distribution** with $n - 1$ degrees of freedom:

$$T \sim t_{n-1}$$

Student's t distribution

The **student's t distribution** is a symmetric distribution with one parameter ν called the degrees of freedom.

As the degrees of freedom ν increases, the t -distribution approaches the normal distribution with mean 0 and variance 1.



Student's t distribution

The t -distribution has mean 0 for $\nu > 1$ (otherwise undefined) and variance $\frac{\nu}{\nu+2}$ for $\nu > 2$ (otherwise undefined).

The t -distribution has heavier tails than the standard normal distribution, meaning that:

- it is more likely to produce values that fall far from its mean.
- it puts less weight in the centre of the distribution
- this accounts for the extra uncertainty on the standard deviation estimation

Test of the Mean if σ unknown

We use a **one-sample t-test** to test

$H_0: \mu = \mu_0$ against one of

$H_1: \mu \neq \mu_0$, or $H_1: \mu < \mu_0$ or $H_1: \mu > \mu_0$.

NB: μ_0 means the value of μ specified by H_0 .

Based on a random sample X_1, \dots, X_n from $N(\mu, \sigma^2)$,
where σ is **unknown**, we use the **test statistic**

$$T = \frac{(\bar{X} - \mu_0)}{S/\sqrt{n}} \sim t_{n-1}$$

Under H_0 , $T \sim t_{n-1}$.

Our **decision rule** will depend on the choice of H_1

We reject H_0 if T is too extreme in the direction(s) suggested by H_1

'Too extreme'

- If $H_0: \mu = \mu_0$ is true, we expect \bar{X} to be close to μ_0 , and T should be close to 0.
- If H_1 says
 - $\mu \neq \mu_0$, then T is **too extreme** if it is too much above 0 or too much below 0.
 - $\mu < \mu_0$, then T is **too extreme** if it is too much below 0.
 - $\mu > \mu_0$, then T is **too extreme** if it is too much above 0.

p-value

Let t be the observed test statistic.

If $H_1 : \mu \neq \mu_0$ (two-tailed),

$$p\text{-value} = P(T \geq |t| \text{ or } T \leq -|t|)$$

If $H_1 : \mu < \mu_0$ (one-tailed -lower),

$$p\text{-value} = P(T \leq t)$$

If $H_1 : \mu > \mu_0$ (one-tailed -upper),

$$p\text{-value} = P(T \geq t)$$

5-step Significance Tests

Testing Hypotheses (5 steps):

- ① Set up the Null and alternative **hypotheses** about μ and hypothesised value μ_0
- ② Select a **level of significance**, α
- ③ Select the appropriate **statistical test** and state **decision rule**
- ④ Perform experiment - **do calculations**
 - calculate test statistic, determine p -value, decision rule is:
Reject H_0 if p -value $< \alpha$
- ⑤ Make decision and **draw conclusions**
 - always state your conclusion in the **context of the problem**

Example: One-sample t-test

Exercise 1: Test hypothesis about the population mean

Consider the manufacturer's claim that, on average, a certain processor cpu has a core-voltage requirement of 1.15V . In reality, this claim may or may not be true.

Let μ be the mean core-voltage requirement. A sample of 20 processors is randomly chosen from a production batch. The sample values in Volts are given below.

1.03	1.18	1.26	0.92	1.19	1.20	1.09	1.10	1.09	1.06
1.10	1.05	1.07	1.16	1.25	1.14	1.10	1.06	1.07	1.39

- (a) Write down the null and alternative hypotheses.
- (b) Test the manufacturer's claim.
- (c) Calculate a 95% CI for the population mean core-voltage requirement.

Exercise 1: solution

Step 1: State Hypotheses:

$H_0: \mu = 1.15V$ i.e The mean core-voltage requirement of these processors is 1.15 V

$H_1: \mu \neq 1.15V$ i.e. The mean core-voltage requirement of these processors is not 1.15V

Step 2: Assign $\alpha = 0.05$

R Summary statistics:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd
0.92	1.07	1.10	1.13	1.18	1.39	0.10

Step 3: Determine appropriate statistical test

Step 3:

- Use T-test as σ is unknown
- Decision Rule: Reject H_0 if $p < \alpha$ i.e. $p < 0.05$

Step 4: Carry out the One sample t-test

Step 4: Carry out T-test

Can do this in **R**:

```
> procvolts <- c(1.03, 1.18, 1.26, 0.92, 1.19, 1.20, 1.09,  
+ 1.10, 1.09, 1.06, 1.10, 1.05, 1.07, 1.16, 1.25, 1.14,  
+ 1.10, 1.06, 1.07, 1.39)  
> t.test(procvolts, mu=1.15)
```

One Sample t-test

```
data: procvolts  
t = -1.0846, df = 19, p-value =  
0.2917  
alternative hypothesis: true mean is not equal to 1.15  
95 percent confidence interval:  
 1.078221 1.172779  
sample estimates:  
mean of x  
 1.1255
```

Step 5: Conclusions

Step 5: Conclusions

We consider the manufacturer's claim that, on average, a **processors** have a 1.15V core-voltage requirement:

Since p-value greater than 0.05, we do not reject H_0

hence, we conclude that there is not enough evidence to dispute the manufacturer's claim that, on average, a **these processors** have a core-voltage requirement of 1.15V.

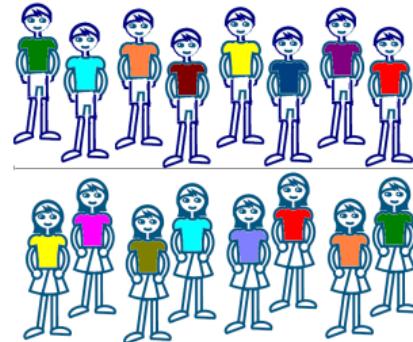
Two-independent-groups design

- People or units are randomly assigned to two separate groups (i.e. different people in each group) or sampling units belong to two separate groups.
- e.g. two different materials for components.
- e.g. Randomised Clinical trial of new drug: control vs experimental group
- e.g. Year 3 NAPLAN results in spelling for M compared with F
- e.g. Results of TIMSS Science Achievement in Australia compared with the U.S.

Comparing Two Independent Groups

The two populations are represented by **two independent samples** such as

- **two categories** of a nominal variable
 - such as M & F
 - or Australian vs Non-Australian
- or **two different treatment groups** in an experiment
 - components were heated
 - control (no heat applied)



We are interested in testing whether **the population means are equal**, i.e. is there a difference between the means of the two groups?

Samples from Two Populations

Examples:

- whether on average, boys and girls perform equally on a certain test
- whether the average nicotine content in one brand of cigarette exceeds that of another brand
- whether the average heat dissipation of cores is different in two different positions within the hardware.
- whether the average weekly food expenditure of families in one city is lower than that of families in another city.
- whether people on Drug A have lower blood pressure than those on drug B

two pops (cont.)

Let X_g be a characteristic of interest, where g denotes which group X comes from

Then $X_{a1}, X_{a2}, \dots, X_{an_a}$ comes from a $N(\mu_a, \sigma_a^2)$ distribution,
and $X_{b1}, X_{b2}, \dots, X_{bn_b}$ be from a $N(\mu_b, \sigma_b^2)$ distribution.

We wish to **make inferences about** $\mathbb{E}(X_a)$ **and** $\mathbb{E}(X_b)$, **say** $\mu_a - \mu_b$.

It is reasonable to expect that this will involve looking at $\bar{X}_a - \bar{X}_b$.

Hypotheses for Comparing Two Means

- The **null** hypothesis: population means differ by an exact amount D_0 :

$$H_0: \mu_a = \mu_b + D_0 \text{ or } \mu_a - \mu_b = D_0$$

- The **alternative** hypothesis may be one- or two-sided:

- if two-sided it will be

$$H_1: \mu_a \neq \mu_b + D_0 \text{ or } \mu_a - \mu_b \neq D_0$$

- if one-sided it could be either

$$H_1: \mu_a - \mu_b > D_0 \text{ or } H_1: \mu_a - \mu_b < D_0$$

if $D_0 = 0$ we are testing for them to be equal to each other

Consider $\bar{X}_a - \bar{X}_b$

We know that $\bar{X}_a \sim N(\mu_a, \frac{\sigma_a^2}{n_a})$, and $\bar{X}_b \sim N(\mu_b, \frac{\sigma_b^2}{n_b})$.

Since a **linear combination** of normal random variables is normal,

$$\bar{X}_a - \bar{X}_b \sim N\left(\mu_a - \mu_b, \frac{\sigma_a^2}{n_a} + \frac{\sigma_b^2}{n_b}\right)$$

We can **standardise**, to obtain

$$Z = \frac{\bar{X}_a - \bar{X}_b - (\mu_a - \mu_b)}{\sqrt{\frac{\sigma_a^2}{n_a} + \frac{\sigma_b^2}{n_b}}} \sim N(0, 1)$$

Consider $\bar{X}_a - \bar{X}_b$ (cont.)

If the values of σ_a^2 and σ_b^2 are known, we could **test hypotheses**:

$$H_0: \mu_a - \mu_b = D_0 \text{ against}$$

$$H_1: \mu_a - \mu_b \neq D_0$$

using the **Test statistic**:

$$Z = \frac{\bar{X}_a - \bar{X}_b - D_0}{\sqrt{\frac{\sigma_a^2}{n_a} + \frac{\sigma_b^2}{n_b}}}$$

If $\sigma_a^2 = \sigma_b^2$

But usually we **don't know** the values of σ_a^2 and σ_b^2

but we believe that they **are equal**; i.e., $\sigma_a^2 = \sigma_b^2 = \sigma^2$ (say).

Then

$$Z = \frac{\bar{X}_a - \bar{X}_b - (\mu_a - \mu_b)}{\sqrt{\left(\frac{\sigma_a^2}{n_a} + \frac{\sigma_b^2}{n_b}\right)}} = \frac{\bar{X}_a - \bar{X}_b - (\mu_a - \mu_b)}{\sqrt{\sigma^2 \left(\frac{1}{n_a} + \frac{1}{n_b}\right)}}$$

If we could find an estimator of σ^2 , this would contain just one unknown parameter.

From the two samples we have S_a^2 and S_b^2 which are both estimators of σ^2 , and so some combined, or **pooled**, estimator might be even better.

How do we combine them?

Pooled Sample Variance

We take a weighted average of S_a^2 and S_b^2 ,
where the weights are the degrees of freedom
associated with each estimator.

Hence

$$S_p^2 = \frac{(n_a - 1)S_a^2 + (n_b - 1)S_b^2}{(n_a + n_b - 2)}$$

So we replace the unknown σ^2 in the $N(0, 1)$ Z variable by the known S_p^2 and get a t -variable instead.

Summary: Independent Samples t -test

This test is often called the **independent samples t -test** and in the case of assuming equal variances, then **pooled variance two-sample t -test**.

Let $X_{a1}, X_{a2}, \dots, X_{an_a}$ be drawn from a $N(\mu_a, \sigma_a^2)$ distribution, and $X_{21}, X_{b2}, \dots, X_{bn_b}$ be drawn from a $N(\mu_b, \sigma_b^2)$ distribution.

- **Assume** independent samples from Normal distributions with $\sigma_a^2 = \sigma_b^2 = \sigma^2$.
- **Hypotheses:** $H_0: \mu_a - \mu_b = D_0$

$$H_1 : \begin{cases} H_1: \mu_a - \mu_b \neq D_0 & \text{two-tailed} \\ H_1: \mu_a - \mu_b < D_0 & \text{one-tailed -lower} \\ H_1: \mu_a - \mu_b > D_0 & \text{one-tailed -upper} \end{cases}$$

Summary: Independent Samples t -test cont.

- **Test statistic:**

$$T = \frac{\bar{X}_a - \bar{X}_b - D_0}{\sqrt{S_p^2 \left(\frac{1}{n_a} + \frac{1}{n_b} \right)}}$$

where $S_p^2 = \frac{(n_a-1)s_a^2 + (n_b-1)s_b^2}{(n_a+n_b-2)}$

- Under H_0 , the test statistic T follows a t -distribution with $n_a + n_b - 2$ degrees of freedom:

$$T \sim t_{n_a+n_b-2}$$

p-value

Let t be the observed test statistic.

If $H_1 : \mu_a - \mu_b \neq D_0$ (two-tailed),

$$p\text{-value} = P(T \geq |t| \text{ or } T \leq -|t|)$$

If $H_1 : \mu_a - \mu_b < D_0$ (one-tailed -lower),

$$p\text{-value} = P(T \leq t)$$

If $H_1 : \mu_a - \mu_b > D_0$ (one-tailed -upper),

$$p\text{-value} = P(T \geq t)$$

Decision Rule

Given a **significance level** α , we reject $H_0 : \mu_a - \mu_b = D_0$ if **p-value** $< \alpha$.

The calculation of the *p*-value depends on if H_1 is two-tailed, one-tailed lower or one-tailed upper.

Example

Exercise: Comparing two population means

- Test at the 5% level of significance to see if there is a difference in the mean level of blood pressure for men and women
- for males: 123.37, 131.97, 172.3, 144.32, 140.23, 138.41, 131.35, 123.6, 131.09, 148.5, 128.2, 129.43, 142.42, 126.74, 149.5, 143, 140.39, 134.13, 115.98, 185.19, 154.75, 121.53, 150.22, 118.65, 123.23, 160.32, 141.05, 145.91, 141.12, 126.6, 145.06, 141.15, 179.99.
- for Females 129.91, 128.44, 113.26, 112.53, 133.65, 127.2, 147.26, 166.86, 147.68, 146.58, 135.58, 149.7, 130.87, 120.37, 155.26, 174.06, 109.81, 115.73, 151.19, 130.63, 134.89, 136.71, 136.06, 133.24, 156.14, 139.55, 140.56, 116.03, 137.66, 99.761, 129.81, 157.9, 140.8, 131.87, 140.42, 95.763, 119.27, 152.17

Example: R output- making sense of data

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd
Males	115.98	128.20	140.39	140.29	145.91	185.19	16.53
Females	95.76	127.51	135.24	134.87	147.09	174.06	17.29

Females have a higher mean blood pressure reading than males and possibly a lower spread.

Are these differences sufficiently small to be considered to be likely if there is no real difference?

Example: Formal Test

Here groups a, b are m, f

Step 1: State hypotheses

$$H_0: \mu_m - \mu_f = 0$$

$$H_1: \mu_m - \mu_f \neq 0$$

Step 2: Assign α

Step 3: Decide statistic

Use t with equal variances and $df=n_m + n_f - 2$

Step 4: Calculations: Determine t and Decision rule

Step 5: Make decision & Conclude

Example: R Output: Carry out t-test

Assuming population variances can be equal:

```
> bps <- data.frame(bp=c(123.37, 131.97, 172.3, 144.32, 140.23, 138.41, 131.35,
+ 123.6, 131.09, 148.5, 128.2, 129.43, 142.42, 126.74, 149.5, 143, 140.39,
+ 134.13, 115.98, 185.19, 154.75, 121.53, 150.22, 118.65, 123.23, 160.32,
+ 141.05, 145.91, 141.12, 126.6, 145.06, 141.15, 179.99, 129.91, 128.44,
+ 113.26, 112.53, 133.65, 127.2, 147.26, 166.86, 147.68, 146.58, 135.58,
+ 149.7, 130.87, 120.37, 155.26, 174.06, 109.81, 115.73, 151.19, 130.63,
+ 134.89, 136.71, 136.06, 133.24, 156.14, 139.55, 140.56, 116.03, 137.66,
+ 99.761, 129.81, 157.9, 140.8, 131.87, 140.42, 95.763, 119.27, 152.17),
+ sex=c(rep('M',33),rep('F',38)))
>
```

Example: R Output: Carry out t-test

```
> t.test(bp~sex, data=bps, var.equal=T)
```

Two Sample t-test

data: bp by sex

t = -1.3449, df = 69, p-value =

0.1831

alternative hypothesis: true difference in means between group F and group M is not equal to 0

95 percent confidence interval:

-13.462050 2.620171

sample estimates:

mean in group F mean in group M

134.8730 140.2939

Example: R Output: Carry out t-test cont.

Given variances can be assumed equal use $t = -1.349$ with $df=69$ ($=38+33-2$)

Verify value of $t = -1.344$:

$$T = \frac{\bar{X}_m - \bar{X}_f - D_0}{\sqrt{S_p^2 \left(\frac{1}{n_m} + \frac{1}{n_f} \right)}}$$

$$S_p^2 = \frac{(n_m - 1)s_m^2 + (n_f - 1)s_f^2}{(n_m + n_f - 2)} = \frac{(38 - 1)(17.289)^2 + (33 - 1)16.526^2}{38 + 33 - 2} = 286.9517$$

$$T = \frac{134.873 - 140.293 - 0}{\sqrt{286.97 \left(\frac{1}{38} + \frac{1}{33} \right)}}$$

Example: R Output: Carry out t-test cont.

Decision rule: Reject H_0 if $p\text{-value} < 0.05$

Or p-value is 0.183 - so probability of getting a t as extreme as this is ie
 $0.183 > 0.05$ under the null hypothesis of no difference in means

Decision: therefore there is not enough evidence to reject H_0 .

Conclude

Given the evidence, we have no reason to reject H_0

We have no evidence to suggest that there is any difference in blood pressure for males and females

Again, as a model

We can see the two sample t-test as saying that the X s we observe are the result of the following model

$$X = \mu_a + \beta G + \varepsilon$$

Where $\varepsilon \sim N(0, \sigma^2)$, and let G be an indicator for group where

$$G = \begin{cases} 0 & \text{if } X \text{ from group } a \\ 1 & \text{if } X \text{ from group } b \end{cases}$$

$$\mathbb{E}(X|G=g) = \mu_a + \beta(0) + E(\varepsilon) = \mu_a \text{ if } G=a$$

$$\mathbb{E}(X|G=g) = \mu_a + \beta(1) + E(\varepsilon) = \mu_a + \beta \text{ if } G=b$$

And so if we think of β as $-D_0$ then, we had $\mu_a - \mu_b = D_0$ hence $\mu_b = \mu_a - D_0$ or $\mu_b = \mu_a + \beta$.

What if $\sigma_a \neq \sigma_b$?

In the case where assuming $\sigma_a = \sigma_b$ seems far fetched

There exists the **Two independent sample t-test with different variances** also known as **Welch's t-test**

- then no S_p is used
- the SE used assumed the original $\sqrt{\frac{\sigma_a^2}{n_a} + \frac{\sigma_b^2}{n_b}}$
and we use its estimate $\sqrt{\frac{S_a^2}{n_a} + \frac{S_b^2}{n_b}}$
- the degrees of freedom are not that easy to remember
- but R (and other software) will do for you

Standard Errors (unpooled)

Unpooled: If we *do not* assume that $\sigma_1 = \sigma_2$

$$\text{Unpooled standard error } \hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} =$$
$$\sqrt{\frac{16.5263291710898^2}{33} + \frac{17.2891618042024^2}{38}}$$

The t distribution has either,

- Welch's df = $\frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)}$ = 68.3418942359236
(more accurate),
- or a conservative estimate df = $\min(n_1 - 1, n_2 - 1) = 32$ (easier to compute).

Same example $\sigma_a \neq \sigma_b$ with R

Say we think 17.289 is very different from 16.526, then.

```
> t.test(bp~sex, data=bps, var.equal=F)
```

Welch Two Sample t-test

data: bp by sex

t = -1.3492, df = 68.342, p-value =

0.1817

alternative hypothesis: true difference in means between group F and group M is

95 percent confidence interval:

-13.437564 2.595685

sample estimates:

mean in group F mean in group M

134.8730 140.2939

Studies involving more than one population or group

- Most of the time we are not only interested in parameter of one population
- a lot of research deals with comparing 2 groups,
 - Two different populations
 - one population but separated into 2 independent groups
 - 2 groups each receiving different treatments
- sometimes 2 sets of measurements on one same group

So far

our observations have been independent of each other

When dealing with comparisons we may have the following 2 scenarios

Comparing Means: Two *Independent* Samples

Data: Two separate groups (e.g., treatments) of individuals, and one measurement on each individual (be it numerical continuous or categorical). we'll start with the continuous numerical scenario

Research Questions:

- Is there sufficient evidence to believe that the population means for the two treatments, say, μ_1 and μ_2 differ? $\mu_1 \stackrel{?}{=} \mu_2$
- How do the population means for the two treatments differ? $\mu_1 - \mu_2 = ?$

Comparing Means: *Paired* Samples

Data:

- One group of individuals, and two measurements:
 - before and after a treatment
 - or measured for two treatments on each individual (think cross over design).
- Two groups of individuals *matched up* before they are sampled.

Research Questions:

- Is there sufficient evidence to believe that the population means for the two treatments (say, μ_1 and μ_2) differ? $\mu_1 \stackrel{?}{=} \mu_2$
 - How do the population means for the two treatments differ? $\mu_1 - \mu_2 = ?$
- ⇒ “Same” question, different kinds of data.
⇒ Different statistical methods.

Paired Data

Since we have looked at the one population problem, let's start there. At the end of the day we can think of the paired data as one population of pairs or differences.

- To assess effectiveness of a weight-loss regimen, weigh each participant before and weigh them after.
- To compare grip strength of dominant hand vs. the “off” hand, measure the grip strength of each hand on each subject.
- *Blocking* in a randomised experiment.
- *Twin studies*.

Generally,

- Whether data are paired depends on the study design.
- “Pairedness” is a property of data, not population.
- We **must** take this into account, sometimes it simplifies our inference and makes it more powerful.
- sometimes, if more than 2 measurements it does not simplify but we **must not** ignore.

Example: A Twin Study

A researcher wanted to measure the effect of home environment on academic achievement of 12-year-olds. Because genetic differences may also affect it, the researcher performed a *twin study*, with 30 pairs of identical twins (i.e., same genes) who had all been adopted prior to turning 1, with one twin placed in a home where academics were emphasised and the other in a home in which they were not.

Twin Study data

Their scores on a test, out of 100 points, are given here:

Pair		Academic		Pair		Academic		Pair		Academic	
ID	Yes	No	ID	Yes	No	ID	Yes	No	ID	Yes	No
1	78	71	11	67	63	21	81	76	22	89	78
2	75	70	12	55	52	23	82	78	24	70	62
3	68	66	13	49	48	25	68	73	26	74	73
4	92	85	14	66	67	27	85	75	28	97	88
5	55	60	15	75	70	29	95	94	30	78	75
6	74	72	16	90	88						
7	65	57	17	89	80						
8	80	75	18	73	65						
9	98	92	19	61	60						
10	52	56	20	76	74						

What are the two populations being compared?

1

Identical twins separated before age 1 and raised in households with different degree of academic emphasis from the other twin, raised in a home in which academics were emphasised.

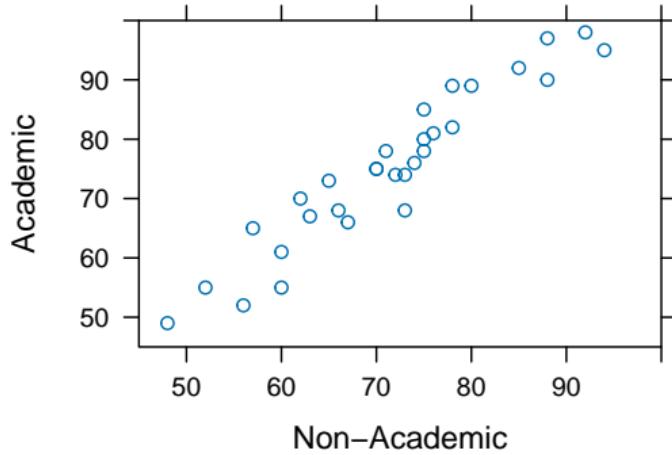
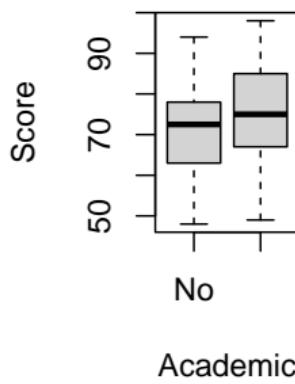
2

Same, but raised in a home in which academics were not emphasised.

Call their population mean test scores μ_Y and μ_N .

Twin Data Summaries

Each pair of twins has the same “nature” - twins with the same genetic makeup *may* have correlated test scores:



Academic	n	\bar{x}	s	r
Yes	30	75.23	13.29	0.953210039849017
No	30	71.43	11.42	

Idea

The observations **within** pair are **not** independent (same genetics, and is the basis of the experiment)

The pairs are independent of each other (different parents, different genetics)

Focus on the **within pair** difference for each pair of twins.

Then we have a set of n independent observations and we take into account the pairing

- Let $d_i = x_{1,i} - x_{2,i}$ (or $d = x_1 - x_2$):

ID	Yes	No	diff=Yes-No
1	78	71	7
2	75	70	5
:	:	:	:
30	78	75	3

Idea (cont.)

- Average of differences (\bar{d}) is (numerically) the same as difference between averages ($\bar{x}_1 - \bar{x}_2$):

$$\bar{d} = \frac{\sum_{i=1}^n d_i}{n} = \frac{\sum_{i=1}^n (x_{1,i} - x_{2,i})}{n} = \bar{x}_1 - \bar{x}_2 = 3.8$$

- Now, the sample standard deviation of differences d_i as one sample, takes the form:

$$s_d = \sqrt{\frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n - 1}} = 4.21$$

Careful!! it is **not** the difference of Standard deviations s_1 and s_2

$$s_d \neq s_1 - s_2$$

and it is **not** the sum $s_1 + s_2$

From this point on, we can proceed as if we just had one sample: d , and compute an interval for $\mu_d = \mu_1 - \mu_2$ using the techniques for one sample we have already seen.

Paired Samples t -Test for Difference of Means

0. Data collection: Test scores for $n = 30$ pairs of twins.

1. Hypotheses: Let $\mu_d = \mu_1 - \mu_2$

$H_0: \mu_d \leq 0$ against $H_A: \mu_d > 0$

$H_0: \mu_d = 0$ against $H_A: \mu_d \neq 0$

$H_0: \mu_d \geq 0$ against $H_A: \mu_d < 0$

$H_0:$

$H_A:$

2. Assumptions and Test Statistic:

- samples paired, pairs independent random sample
- either the population of *differences* is normal or the sample is large
- test statistic: for $d_i = x_{1,i} - x_{2,i}$ (or $d = x_1 - x_2$): 7

$$t = \frac{\bar{d}}{s_d / \sqrt{n}} = \frac{3.8}{4.21 / \sqrt{30}} = 4.9438140582414$$

where

$$s_d = \sqrt{\frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n - 1}} = 4.21$$

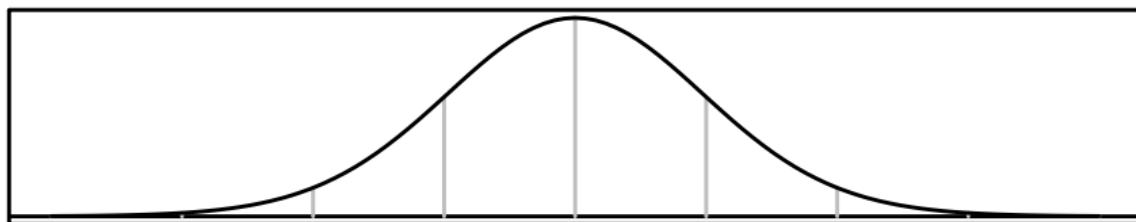
- it has $df = n - 1 = 30 - 1 = 29$

3. Calculate the p -value:

H_0	H_A	p -value
$\mu_d \leq 0$	$\mu_d > 0$	$P(T_{\text{df}} \geq t)$
$\mu_d \geq 0$	$\mu_d < 0$	$P(T_{\text{df}} \leq t)$
$\mu_d = 0$	$\mu_d \neq 0$	$P(T_{\text{df}} \geq t)$
or $2 \times \min(P(T_{\text{df}} \leq t), P(T_{\text{df}} \geq t))$		

$$p\text{-value} = P(T_{\text{df}} \geq t) =$$

\approx



4. Statistical decision: Since $2.9e - 05 < \alpha$, reject H_0 in favour of H_A .

5. Conclusion: There _____ sufficient evidence that twins raised in an environment where academics were emphasised score, on average, _____ than those who are not.

R: Paired Samples t -Interval

```
> datas<-c(1, 78, 71, 11, 67, 63, 21, 81, 76, 2, 75, 70, 12, 55, 52, 22, 89,
+      78, 3, 68, 66, 13, 49, 48, 23, 82, 78, 4, 92, 85, 14, 66, 67, 24,
+      70, 62, 5, 55, 60, 15, 75, 70, 25, 68, 73, 6, 74, 72, 16, 90, 88,
+      26, 74, 73, 7, 65, 57, 17, 89, 80, 27, 85, 75, 8, 80, 75, 18, 73,
+      65, 28, 97, 88, 9, 98, 92, 19, 61, 60, 29, 95, 94, 10, 52, 56, 20,
+      76, 74, 30, 78, 75)
> twins <- setNames(as.data.frame(matrix(datas,ncol=3,byrow=TRUE)),
+      c("ID","Yes","No"))
> twins <- twins[order(twins$ID),]
>
>
```

R: Paired Samples *t*-Interval

```
> t.test(twins$Yes, twins$No, paired=T)
   Paired t-test

data:  twins$Yes and twins$No
t = 4.9496, df = 29, p-value =
2.918e-05
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 2.229795 5.370205
sample estimates:
mean difference
            3.8
```

Other Notes on Paired t Intervals and Tests

- Pairing is a function of how the data are collected!
- Paired t procedures can be used to “control” for some confounding factors, if the “pairing” is based on that factor.
- Make sure you know what is being paired with what. For example, suppose that we have two dieting regimens (A and B) that we want to compare, so we assign each subject to one of the two regimens, and measure their “before” and “after” weights.
 - Before and after weights are paired.
 - Regimens A and B are not.
- Just because $n_1 = n_2$ doesn’t mean the data are paired!

Confidence Intervals: Recap

Most confidence intervals generally have the form

$$\text{estimate} \pm \underbrace{\text{multiplier}_{1-\alpha/2}^* \times \text{standard error}}_{\text{margin of error}}$$

parameter	estimate	multiplier	standard error
pop. proportion p	$\hat{p} = \frac{\# \text{ success}}{n}$	$z_{1-\alpha/2}^*$	$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
pop. mean μ	\bar{x}	$t_{n-1, 1-\alpha/2}^*$	$\frac{s}{\sqrt{n}}$
diff. pop. means, paired samp. $\mu_d = \mu_1 - \mu_2$	$\bar{d} = \bar{x}_1 - \bar{x}_2$	$t_{n-1, 1-\alpha/2}^*$	$\frac{s_d}{\sqrt{n}}$
diff. pop. means, ind. samp. $\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$t_{df, 1-\alpha/2}^*$	$\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$ or $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

$\hat{\square}$ = estimate of \square

standard error of \square = how much \square varies from sample to sample

Why bring that back?

When looking at these confidence intervals in particular, remember that :

- they are derived from their sampling distributions
- Hence we are looking, for θ at $P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) = 1 - \alpha$

Say for a standardised sample mean \bar{x} , with known σ under its sampling distribution. We have

$$\begin{aligned}.95 &\approx P\left(-1.96 < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < 1.96\right) \\&\approx P\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)\end{aligned}$$

which is the form of the known confidence interval

CIs and 2-tailed Hyp Tests

But the general form would be

$$1 - \alpha \approx P \left(z_{\alpha/2} < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < z_{1-\alpha/2} \right)$$

by symmetry of Z distribution

$$1 - \alpha \approx P \left(z_{\alpha/2} < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < -z_{\alpha/2} \right)$$

$$\approx P \left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

CLs and 2-tailed Hyp Tests

and even more general

$$1 - \alpha \approx P \left(distr_quant_{\alpha/2} < \frac{\hat{\theta} - \theta}{std_error_{\hat{\theta}}} < distr_quant_{1-\alpha/2} \right)$$

by symmetry of Z and t distributions

$$1 - \alpha \approx P \left(distr_quant_{\alpha/2} < \frac{\hat{\theta} - \theta}{std_error_{\hat{\theta}}} < -distr_quant_{\alpha/2} \right)$$

$$\approx P \left(\hat{\theta} - distr_quant_{\alpha/2} std_error_{\hat{\theta}} < \theta < \hat{\theta} + distr_quant_{\alpha/2} std_error_{\hat{\theta}} \right)$$

Check that most of our test statistics so far are of the form

$$\text{test_statistic} = \frac{\text{estimate} - \text{observed}}{\text{Std_error}_{\text{estimate}}}$$

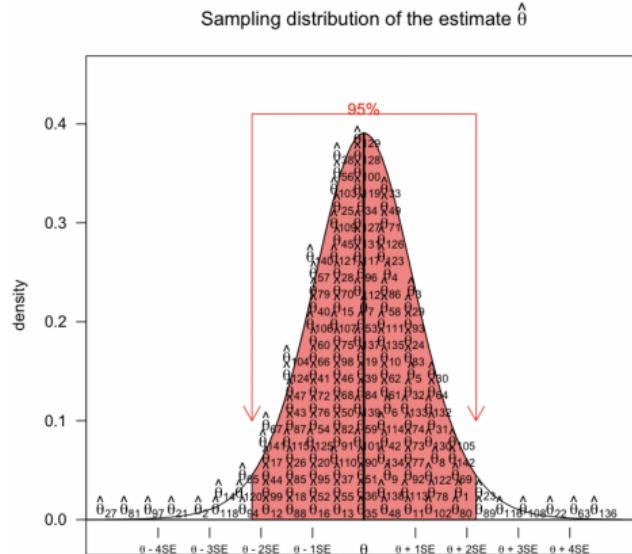
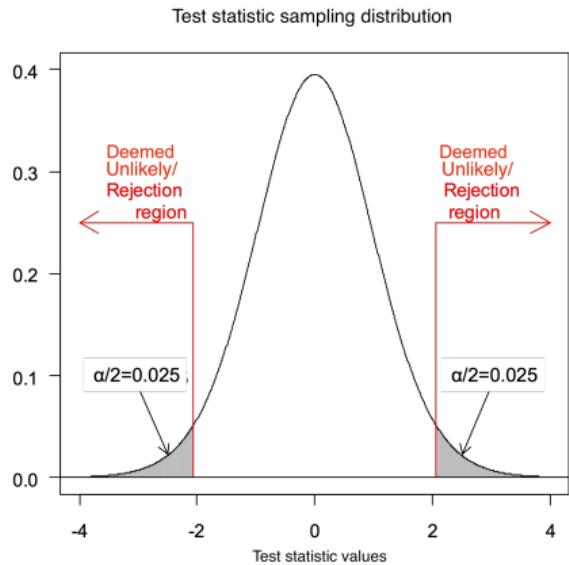
And hence:

- the critical value of the test for two tailed test, the one giving rejection region, is that one that has:
- $\alpha/2$ area on the tails
- and $1 - \alpha$ in the middle.

So hence most of the time we will conclude the same with a two tailed test as we would, based on a confidence interval.

CI and 2-tailed hypotheses

Think of it:



Aspirin and Heart Attacks: The Physicians' Health Study

More details at <http://phs.bwh.harvard.edu/phs1.htm>

- In late 1940s, Dr. Lawrence Craven, MD, a Californian GP, observed that aspirin prevented blood clotting, and suspect that it may reduce the risk of heart attack. He prescribed it to several thousands of his patients, claiming success. More rigorous studies in the 1960s and 1970s did not produce clear conclusions.
- In 1981, a team from Harvard Medical School invited 261,248 American male physicians aged between 40- 84 to participate in a double-blind study of aspirin's effects. After filtering out unqualified participants and screening, a total of 22,071 were ultimately randomised into two groups: aspirin (11,037) and placebo (11,034).

Data

By 1987, the following was the incidence of myocardial infarction in the two groups:

		Heart Attack		
		Yes (H)	No (\bar{H})	Total
Aspirin (A)		104	10933	11037
Placebo (\bar{A})		189	10845	11034

Aspirin group: $P(H|A) \approx \frac{104}{11037} = 0.00942$

Placebo (baseline) group: $P(H|\bar{A}) \approx \frac{189}{11034} = 0.0171$

Strictly, these **are** conditional probabilities, but we **can** refer to them as p_A and $p_{\bar{A}}$ where $p = P(\text{Heart attack= yes})$ and subscript A, \bar{P} identifies the group in which they were measured.

Just like we named μ_a and μ_b for 2 groups a and b

Test of one proportion

We use a **one-sample z-test** to test whether the population proportion is (or is not) significantly different from some hypothesised value.

$H_0: p = p_0$ against one of

$$H_{1a}: p \neq p_0, \quad \text{or} \quad H_{1b}: p < p_0 \quad \text{or} \quad H_{1c}: p > p_0.$$

NB: p_0 means the value of p specified by H_0 .

Based on a random sample X_1, \dots, X_n from $\text{Bin}(n, p)$,
we use the sample proportion $\hat{p} = \frac{X}{n}$ as our observed estimate, and hence the
test statistic

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p(1-p)}{n}}}$$

tells us how far the observed is from the expected and allows us to calculate the probability of observing it under the null.

Test of a proportion

Assuming H_0 true, then $p = p_0$, and then the sample mean approximately follows a Normal distribution with expected value $\mathbb{E}(\hat{p})=p_0$ and Variance $\frac{p(1-p)}{n}$,

that is $\hat{p} \sim N(p_0, \frac{p(1-p)}{n})$

then the test statistic Z approximately follows a standard normal distribution:

$$Z \sim N(0, 1) \quad \text{under } H_0$$

If H_0 is **not** true, then Z will not follow a standard normal distribution.

Differences of Proportions

- How does the five-year survival proportion after Treatment A differ from that after Treatment B?
- How much larger is the risk of early onset of bipolar disorder for those with a family history of mood disorders than those without?
- Is the proportion of stunted under 5 year olds in Rural Uganda higher or lower than it was 10 years ago?

Aspirin and Heart Attack Data

The following was the incidence of myocardial infarction in the two groups:

		Heart Attack	
		Yes	No
		Total	
Aspirin	104	10933	11037
Placebo	189	10845	11034

Note that **Incidence** is about new “happenings” within a population considered at risk)over a certain period).

different from **Prevalence** that is the number in a “*state*” or with a specific characteristic during a specific time frame.

Define our Populations

There are two **populations** with two proportions:

Placebo group: Call p_P the proportion who would have a heart attack out of those who participated in this experiment treated with the placebo.

Aspirin group: Call p_A the proportion who would have a heart attack out of those who participated in this experiment treated with aspirin.

What we have (our sample): Size of (n_P) and numbers of heart attacks X_P in the placebo group and, respectively n_A and X_A in the aspirin group.

What we want: Difference $p_A - p_P$: the amount of change in population proportion of heart attack due to taking aspirin.

$$\hat{p}_A - \hat{p}_P = 0.0094 - 0.0171 = -0.0077$$

2-Sample z -test for Difference of Population Proportions

- ① Check that
 - samples are independent, random (or close)
- ② set Hypothesis for p_1 vs p_2
- ③ Find test statistic z_{test}
- ④ Calculate p-value based on decision rule
- ⑤ Conclude

Aspirin and Heart Attacks

	Heart Attack		
	Yes	No	Total
Aspirin (Group 1)	104	10933	11037
Placebo (Group 2)	189	10845	11034

Does Aspirin Reduce the Rate of Heart Attacks?

		Heart Attack		Total
		Yes	No	
Aspirin (Group 1)	Yes	104	10933	11037
	Placebo (Group 2)	189	10845	11034

We already have that:

$$\hat{p}_A = \frac{X_A}{n_A} = \frac{189}{11034} = 0.0171 \text{ and } \hat{p}_P = \frac{X_P}{n_P} = \frac{104}{11037} = 0.0094.$$

and that

95% CI for $p_A - p_P$ is between -1.07 and -0.47 percentage points.

Is there sufficient evidence to believe that aspirin reduces heart attack probability?

2-Proportion z -Test

0. Data: To assess whether the probability (population proportion) of heart attack under the Aspirin treatment is lower than that under the Placebo treatment, a randomised, controlled trial was conducted among 22071 male physicians aged 40- 84 with no prior history of heart disease or stroke. Outcomes (myocardial infarction or not) in $n_P = 11034$ physicians in the placebo group and $n_A = 11037$ in the experimental group.

1. Hypotheses: Notice that $p_A \square p_P$ is equivalent to $p_A - p_P \square 0$:

$$H_0: p_A - p_P \leq 0 \text{ vs. } H_A: p_A - p_P > 0$$

$$H_0: p_A - p_P = 0 \text{ vs. } H_A: p_A - p_P \neq 0$$

$$H_0: p_A - p_P \geq 0 \text{ vs. } H_A: p_A - p_P < 0$$

$$H_0:$$

$$H_A:$$

2. Assumptions and Test Statistic:
- Counts X_A and X_P are independent (and things being counted are a random (or close) sample from their population).
 - X_A and X_P are well-approximated by a normal distribution: $(n_A + n_P)\bar{p} \geq 10$ (or 15, or 20 to be safe) and $(n_A + n_P)(1 - \bar{p}) \geq 10$, with \bar{p} defined below.
 - We use a “pooled” \hat{p} for the standard error calculation and assumptions checking:

$$\begin{aligned}\bar{p} &= \frac{x_A + x_P}{n_A + n_P} = \frac{n_A \hat{p}_A + n_P \hat{p}_P}{n_A + n_P} \\ &= \frac{104 + 189}{11037 + 11034} = \frac{293}{22071} = 0.0133.\end{aligned}$$

$$(11037 + 11034) \times 0.0133 = 293 > 20$$

$$(11037 + 11034) \times (A - 0.0133) = 21778 > 20$$

Why? If $H_0 : p_A = p_P$ is, in fact, true, then we only have one p to estimate, so we combine \hat{p}_A and \hat{p}_P .

2. Assumptions and Test Statistic: (continued)

- Then, the standard error of $\hat{p}_A - \hat{p}_P$ under H_0 is

$$\begin{aligned}\hat{\sigma}_{\hat{p}_A - \hat{p}_P} &= \sqrt{\frac{\bar{p}(1-\bar{p})}{n_A} + \frac{\bar{p}(1-\bar{p})}{n_P}} = \sqrt{\bar{p}(1-\bar{p}) \left(\frac{1}{n_A} + \frac{1}{n_P} \right)} \\ &= \sqrt{0.0133(1-0.0133) \left(\frac{1}{11037} + \frac{1}{11034} \right)} = 0.0015.\end{aligned}$$

- The test statistic is then

$$\begin{aligned}z &= \frac{(\hat{p}_A - \hat{p}_P) - (p_A - p_P)}{\hat{\sigma}_{\hat{p}_A - \hat{p}_P}} \\ &= \frac{(0.0094 - 0.0171) - 0}{0.0015} = -5.0014.\end{aligned}$$

3. *p*-value:

H_0	H_A	<i>p</i> -value
$p_A \leq p_P$	$p_A > p_P$	$P(Z \geq z)$
$p_A - p_P \leq 0$	$p_A - p_P > 0$	
$p_A \geq p_P$	$p_A < p_P$	$P(Z \leq z)$
$p_A - p_P \geq 0$	$p_A - p_P < 0$	
$p_A = p_P$	$p_A \neq p_P$	$P(Z \geq z)$ or
$p_A - p_P = 0$	$p_A - p_P \neq 0$	$2 \times \min(P(Z \leq z), P(Z \geq z))$

$$P(Z < -5.001) \approx 0.000000285.$$

4. Statistical decision:

Since

5. Conclusion:

There _____ sufficient evidence to believe that aspirin reduces the rate of heart attacks, at least among American male physicians aged 40- 84.

- In R we can a vector of counts of successes, a one-dimensional table with two entries (for oneproportion)
- or a two-dimensional table (or matrix) with 2 columns, giving the counts of successes and failures, respectively
- or if we had the raw data we would have the variable Heart attack and treatment and we would have for each subject the values.

in our case right now it looks this way:

```
> use <- matrix(c(104,10933,189,10845),nrow=2,byrow=T)
> prop.test(use,cor=F)
```

```
2-sample test for equality of
proportions without continuity
correction
```

```
data: use
X-squared = 25.014, df = 1, p-value
= 5.692e-07
alternative hypothesis: two.sided
95 percent confidence interval:
-0.010724297 -0.004687751
sample estimates:
prop 1      prop 2
0.00942285 0.01712887
```

MATH55

Alberto Nettel Aguirre

NIASRA
School of Mathematics and Statistics
University of Wollongong

May 19, 2023

Very Important

- I hope you have noticed how nuanced my interpretations of our hypotheses tests have been.
- The whole machinery of statistics is about **inference**
- as such there needs to be care in how these tests are used and interpreted
- statistical tests have a history of being misused and misinterpreted
- in the last 10 years there have been harsher and more vocal criticisms to statistical hypothesis methods
- with extremes in journals **banning** the use of p-values.
- The American Statistical Association convened a panel of experts to shed more light and bring things back on track.

ASA Statement on p-values

6 principles:

- ① P-values can indicate how incompatible the data are with a specified statistical model.
- ② P-values **do not** measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
- ③ Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
- ④ Proper inference requires full reporting and transparency.
- ⑤ A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
- ⑥ By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

American Statistical Association (March 7, 2016)

<https://www.amstat.org/newsroom/pressreleases/P-ValueStatement.pdf>.

Null Hypothesis Significance Testing

Given a **model** and a set of **assumptions**, we evaluate our observed data, which we take as evidence *comparing with*:

H_0 : The null hypothesis

- Often the null hypothesis postulates the absence of an effect, such as no difference between two groups, or the absence of a relationship between a factor and an outcome.
- Not always a lack of effect, remember it states an equality or a model that allows us to set a stage.
- The smaller the p-value, the greater the statistical incompatibility of the data with the null hypothesis, *if the underlying assumptions used to calculate the p-value hold*.
- This incompatibility can be interpreted as casting doubt on or providing evidence against the null hypothesis or the underlying assumptions.

ASA principle 2

*P-values do **not** measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.*

- Researchers often wish to turn a p-value into a statement about the truth of a null hypothesis, or about the probability that random chance produced the observed data.
- The p-value is neither.
- It is a statement about data in relation to a specified hypothetical explanation, and is not a statement about the explanation itself.
- Remember I have emphasised it is a **conditional** probability
 - Conditional on assumptions
 - Conditional on hypothesis
- always referred to as how likely is what we have observed under the hypothesis (remember salary, jet, BMWs, etc...)

ASA principle 3

Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.

Researchers should bring many contextual factors into play to derive scientific inferences, including:

- the design of a study,
- the quality of the measurements,
- the external evidence for the phenomenon under study, and
- the validity of assumptions that underlie the data analysis.
- CIs as potential measures of magnitude of “effect”.

The widespread use of “statistical significance” (generally interpreted as $p \leq 0.05$) as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process.

ASA principle 4

Proper inference requires full reporting and transparency.

- P-values and related analyses should not be reported selectively.
- Conducting multiple analyses of the data and reporting only those with certain p-values (typically those passing a significance threshold) renders the reported p-values essentially uninterpretable.

There is also the issue of the “file drawer” problem: if a study doesn’t produce $p \leq 0.05$, often the results are not published.

ASA principle 5

A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.

- Statistical significance is not equivalent to scientific, human, or economic significance.
- Smaller p-values do not necessarily imply the presence of larger or more important effects, and larger p-values do not imply a lack of importance or even lack of effect.
- Any effect, no matter how tiny, can produce a small p-value if the sample size or measurement precision is high enough.
- remember you can always care more or less about the errors and hence move α
- a test with $p = 0.003$ is not more important than one with $p = 0.008$

ASA principle 6

By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Consider alternative approaches:

- confidence, credibility, or prediction intervals.
- Bayesian methods.
- alternative measures of evidence, such as likelihood ratios or Bayes Factors.
- decision-theoretic modelling.
- false discovery rates.

Remember

- Assumptions are important
- Stay away from using “*results were due to chance*” kind of wording.
 - what does that even mean?
 - there are probability distributions which could be thought of as “chance”
 - hence it is an empty statement
- Stay away from using any wording around *Truthfulness & proving*
- To always think if conclusions/findings make sense