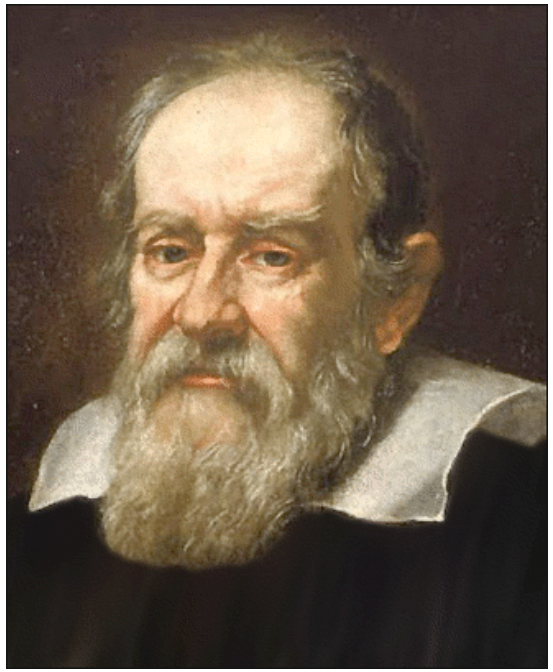
Evaluation Methods and Statistics: Investigating human behavior

Professor Andrew Howes
Dr Ben Cowan

School of Computer Science
University of Birmingham

evidence versus authority

- "The leitmotif which I recognise in Galileo's work is the passionate fight against any kind of dogma based on authority. Only experience and careful reflection are accepted by him as criteria of truth" – Einstein.
- http://www.guardian.co.uk/commentisfree/2011/nov/22/neutrino-revolutionary-image



scientific claims are based on evidence

Here are some headlines from Science Daily (19th Dec 2011).

- Less knowledge, more power: uninformed can be vital to democracy, study finds.
- Traumatic experiences may make you tough.
- New strain of lab mice mimics human alcohol consumption pattern.
- Second-guessing ones decisions leads to unhappiness.

http://www.sciencedaily.com/news/mind_brain/psychology/

claims need evidence

- "E-mail Reveals Your Closest Friends" (Science, 2011)
- "We found that all measures of an ego network structure calculated from the self-reported data correlated significantly ($p<0.05$, ...) with the email derived, ... networks." (Wuchty & Uzzi, 2011)
- http://news.sciencemag.org/sciencenow/2011/11/e-mail-reveals-your-closest-frie.html

Aims of the module

- The aim of the module is to provide an introduction to the use of empirical scientific methods, including experimental design and statistics, for the purpose of investigating human interaction with computers.
- The module is targeted at computer scientists with an interest in
- (i) building systems that support human activities (including Human-Computer Interaction),
- (ii) building computational models of human behaviour, and/or,
- (iii) understanding human behaviour as an inspiration for computational science (Machine Learning and Artificial Intelligence).

Outcomes

- On successful completion of this module, you should be able to:
- Identify and discuss research methodologies for investigating human behaviour.
- Recognise the appropriateness of statistical techniques in data analysis.
- Conduct and report a variety of statistical tests.
- Interpret research findings from a variety of statistical techniques to a high level.
- Discuss issues related to conducting research on human participants (sampling, recruitment etc).

Prerequisites

- There are no formal prerequisites.
- However, a willingness to engage enthusiastically with the mathematics and computation that underpin statistical reasoning will be essential.
- There will be a strong emphasis on individual and group practical work.

Lecturers

- Professor Andrew Howes
- Email:HowesA@bham.ac.uk

- Dr Ben Cowan
- b.r.cowan@cs.bham.ac.uk

What do we mean by human behavior?

- Perceptual-motor control, e.g. moving a mouse, driving a car, flying an airplane.
- Cognitive tasks, e.g. navigating the web, managing a diary.
- Social and collaborative tasks, e.g. maintaining social relationships, collaborating with colleagues.
- Economic tasks, e.g. managing businesses.

The scientific method

- The progress that has been made in understanding human behavior has been due to the application of quantitative scientific methods.
- Methods are used to test scientific theory.
- Methods specify the following components:
  - Constraints on the design of studies (e.g. control conditions, ethics etc.)
  - Tools for the analysis of data.
  - Conventions for reporting studies.
- The content of this module is largely concerned with these methods.

Course overview

- How to make an evidence-based argument. How to write scientific reports.
- Basic Statistics (mean, mode, standard deviation, central limit theorem, statistical distributions etc.)
- Experimental design (types of error, the Null Hypothesis, independent/dependent variables)
- Correlation (what does it mean for two variables to be correlated? how to calculate a correlation?)
- How to compare distributions (t-test, ANOVA, 2-way ANOVA).
- Questionnaires and interview methods.

Reading

- Discovering Statistics Using R, **Andy Field, Jeremy Miles, Zoe Field**. Sage, 2012.
- Statistical Methods for Psychology, **David C. Howell**. Duxbury, 2006.
- Statistics in R: An introduction using R, **Michael J. Crawley**. Wiley, 2005.
- Discovering statistics using SPSS, **Andy Field**, Wiley, 2009

Lecture 1

1. Traditional authority: Science is NOT about trusting what scientists claim.
2. Science isn't about trusting what scientists claim are the facts.
    - quality of method?
    - quality of data.
    - mutated facts.
3. Science is about evidence based argumentation.
4. The form of a scientific argument: An example.

PART 1:

Traditional authority: Science is NOT about trusting what scientists claim.

consider these headlines

- "Online networking 'harms health'" (BBC, 2009a)
- "Online risks: from cancer to autism?" (BBC, 2009b)
- "Facebook and Bebo risk 'infantilising' the human mind" from the Guardian (Wintour, 2009).

Professor Baroness Susan Greenfield (Newsnight, 2009)

"... one can look at the features of screen life and see that it is perhaps now mirrored in the behaviour of the upcoming generation if you like. One might argue a shorter attention span, an emphasis on process, on the experience of the moment rather than content, of an identity that needs to be bolstered up with twitter, and perhaps an increased recklessness."

‘infantilised’ mind

- According to Wintour (2009) she told the House of Lords that children's experiences on social networking sites,
- "are devoid of **cohesive narrative** and long-term significance. As a consequence, the mid-21st century mind might almost be infantilised, characterised by short attention spans, sensationalism, inability to empathise and a shaky sense of identity".

“erosion of our identity”

- In the same article Wintour reports that Susan Greenfield had said she found it strange we are "enthusiastically embracing" the possible erosion of our identity through social networking sites, since those that use such sites can lose a sense of where they themselves "finish and the outside world begins".

Claims from qualified scientists

- These articles are based on claims made by two well qualified scientists.
- Dr Aric Sigman is a Fellow of the Royal Society of Medicine, an Associate Fellow of the British Psychological Society, Member of the Institute of Biology and has received the Chartered Scientist award from the Science Council. (Sigman, 2009)
- Baroness Susan Adele Greenfield (CBE 2000, Baroness, 2001; Fullerian Professor of Physiology and Comparative Anatomy, Oxford, 1999-) and Director of the Royal Institution (Royal Institution, 2009).

However... these claims are not evidence based.

- Despite the eminent position of Baroness Greenfield, what she is doing is expressing an opinion.
- Science can be deployed to answer these questions. e.g. how evidence shows that Facebook has helped some build social capital (Ellison et al., 2007).

How can we come to an evidence-based claim?

Rational v Traditional authority

- We can choose to avoid relying on traditional authority and instead rely on rational authority.
  - Driver, Newton, Osborne (2000), for example, have argued for doing so in the classroom.
- Scientists must try and make use of rational authority.
- If someone asks you why you believe something, the answer "because I read it on Wikipedia" or "because Newton said so" (traditional authority), or other sources, is a weak justification.

PART 2

Science is about high quality, quantitative data.

Quality of data

- **Claim**: Virtual Reality provides a better platform for e-commerce than traditional picture and text web sites.
- How might this claim be supported by data?
- **Verbal reports** e.g. 9 out of 10 people said that they preferred QTVR.
- **Performance data** e.g. In a recall study Howes et al. (2001) found that users of Virtual Reality had better memory for products.

as scientists we need to avoid anecdotal evidence everyone has a story...

getting the facts right...

- http://www.timesonline.co.uk/tol/comment/columnists/david_aaronovitch/article5834725.ece
- Some data is nothing of the sort. If you look on the web you will find the following statistic:
- "the average Brit is caught on security cameras some 300 times a day"
- For example, this claim was published in The Sunday Times almost exactly two years ago, and referred "to the results of a study by the Government's privacy watchdog" (the Office of the Information Commissioner), which "found people were caught on a national network of 4.2 million CCTV cameras an average 300 times a day".

mutated facts...

- Aaronovitch (2009) shows how the "300 times" had become viral: "It now occurs all over the place, and is the standard statistic used for the number of times Britons may or will be captured by CCTV cameras daily."

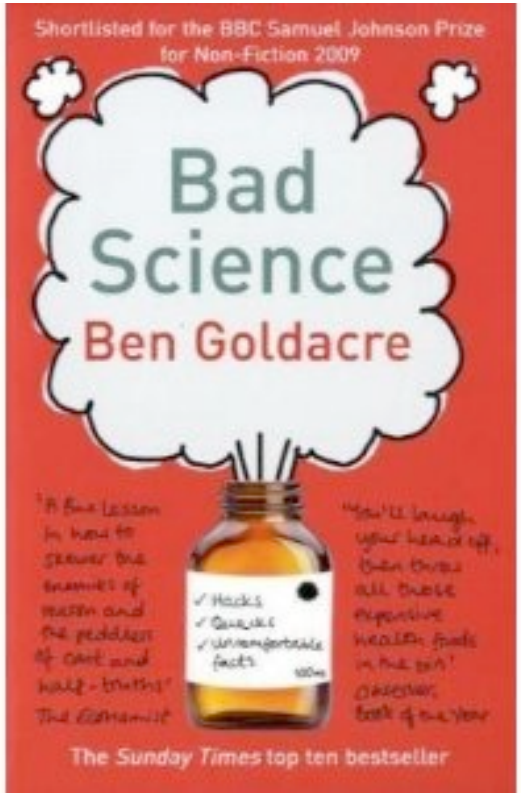
- He also shows the tendency for the statistic to mutate, as in the transformation from "can be captured" to the completely different "the average Briton is captured". A British boy can have a baby at 13. That is clearly not the average age of first fatherhood.

- A New Statesman columnist had it as the "average Londoner going about his or her business… may be monitored by 300 cameras each day", and a Daily Mail report that "it has been calculated that each person is caught on camera an average of 300 times daily".

where does the statistic originate...

- The source was a book "The Maximum Surveillance Society", published in 1999, by two academics, including a C. Norris.
- It wasn't a fact at all, but a fiction. Norris had imagined a "day in the life of Thomas Reams"
- "Reams is a City type who, rather unusually, lives on a drug-infested estate. He manages to visit two schools, the maternity wing of a hospital, goes to work, shops, is caught speeding in his car, crosses a level-crossing, parks in several car parks before switching to public transport. He goes to Heathrow airport, then a football match at Chelsea, after which he drives through London's most notorious red-light district (by mistake, I hasten to reassure the fictional Mrs Reams)."
- by the end of the day the fictional Mr Reams has been observed 300 times.

- This is the study of just one statistic.
- There is plenty of evidence of sloppy reporting and of bad science (e.g. see www.badscience.net for examples).
- The same problems can be found in the academic literature.
- A key contribution of this module will be to help provide you with the intellectual tools to stop you falling into similar traps.



PART 3

Science is about evidence based argumentation.

The problem

- argumentation is difficult.
  - Arguments are subject to confirmation biases.
  - They are inattentive to opposing positions (Kuhn, 2007).

The Layout of Arguments

what not to do...

- consider this example...

"What are the characteristics of a good manager?

A good manager in my view must possess charisma; an individual without charisma is definitely not going to become a good manager. So why is charisma such an important attribute towards a good manager? Well management and leadership are very closely linked; although not the same usually a good manager can easily be seen as a leader. Looking at good leaders through the years it is hard to come across a good leader who was not portrayed as a charismatic character; for example Winston Churchill; Martin Luther King and Barack Obama were or are seen to be charismatic. When we consider the skills and attitudes which create charisma; it becomes very easy to see why charisma plays such a vital role in the formulation of a good manager. "

what not to do..

- This is a more subtle example.
- It is a highly cited paper by a respected author in a scientific conference.
- The citation appears to offer evidence for a claim but the nature of this evidence is not described and may not be present at all.
- Instead a metaphor follows the citation.

Brignull and Rogers (2003): "Social embarrassment has been identified as a key factor, especially in determining whether people will interact with a public display in front of an audience (Rogers & Brignull, 2002). We draw an analogy here with a street performer in a public place, who invites a participant from the audience to 'help out' with their show. Such a person can often be wary of volunteering, not knowing what exactly will be required from them, especially if it entails making them look foolish in the eyes of the on-looking audience."

- Brignull & Rogers (2003). Enticing people to interact with large public displays in public spaces.. Proceedings of Interact, 3, 17-24.

What is a good argument?

- Toulmin (1958) provides a framework that can be used to help distinguish well-formed from poorly formed or incomplete arguments.
- If you make a claim that is challenged then you will need to make an argument to support the claim. What form should that argument take?

Claim (C) Data (D) Warrant (W)

- Claim: Virtual Reality could increase online sales.
- Data: Howes et al. (2001) observed that people tend to remember more about the range of available products when using virtual reality.
- Warrant: People who remember more about the contents of a store are more likely to return and therefore more likely to purchase more.

Exercise

- identify the claim, data, and warrant in the following paragraph.

Claim, data, warrant

- The internet has beneficial effects for social connectivity. The Pew internet surveys between 2000 and 2003 asked hundreds of people about the role of email in family communication and a majority of respondents said that it increased frequency of communication. Generally it is thought that increased frequency of contact is associated with higher social connectivity.

Claim, data, warrant

- The internet has beneficial effects for social connectivity. The Pew internet surveys between 2000 and 2003 asked hundreds of people about the role of email in family communication and a majority of respondents said that it increased frequency of communication. Increased frequency of contact is associated with higher social connectivity.

Claim, data, qualification, warrant

- The internet may have beneficial effects for social connectivity. The Pew internet surveys between 2000 and 2003 asked hundreds of people about the role of email in family communication and a majority of respondents said that it increased frequency of communication. Generally it is thought that increased frequency of contact is associated with higher social connectivity.

Practical

- practical class in the R programming language starts next week.
-

The R programming language

- R is a programming language for statistical computing.
- It was first implemented in the 1990s and has millions of users.
- It is possible to get started using R with a simple command-line syntax.
- For example,

```
> 1+2
[1] 3
> data <- c(1,2,3,4)
> mean(data)
[1] 2.5
>
```

end.

Evaluation Methods and Statistics
Lecture 2
Basic statistics

Professor Andrew Howes
Dr Ben Cowan

School of Computer Science
University of Birmingham

Previous week

- The structure of argumentation.
- Traditional versus rational authority.
- Claims, Data, Warrants, and Qualifiers.
- Social capital and facebook intensity of use example.
- Statistical modeling with R.

This week

- Work towards modeling variability, distributions and central tendency.
- Frequency plots
- Density plots
- Central Limit Theorem

What did the crocodile swallow in Peter Pan?

Which is the only mammal that can't jump?

- Your mind has now primed "Google" (Sparrow et al., 2011).

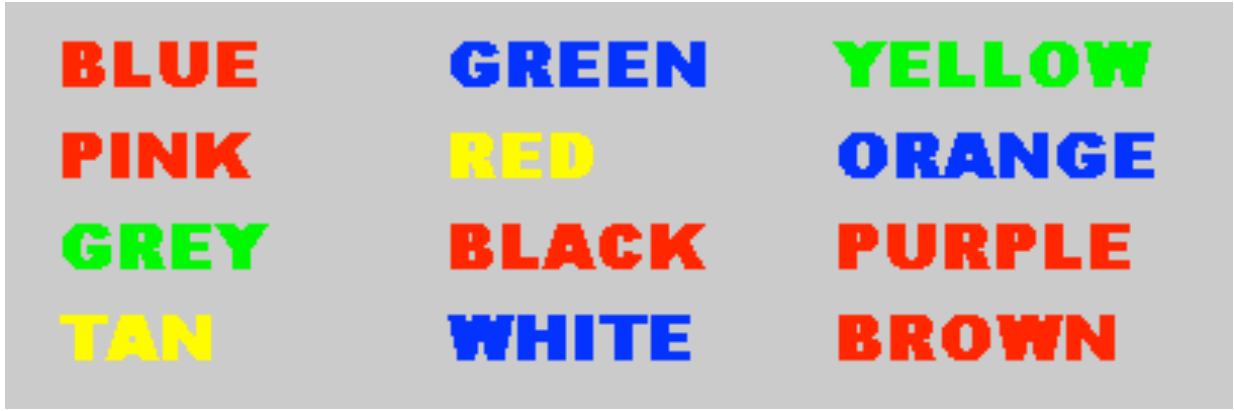
- Sparrow, B., Liu, J. & Wegner, D.M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science, 333*, 776-778.

Limitations on top-down control.

- People have a limited ability to control information processing.
- One study that supports this claim was reported by Stroop (1935). In a typical Stroop study participants are asked to name the ink colour of colour name words. Congruent colour words are printed in the same colour as the meaning of the word, e.g. the word green is printed in green ink. Noncongruent words are printed in a different colour, e.g. the word blue in red ink. Many studies have observed a significant effect of incongruence on reaction times. It takes longer for people to identify the ink colour of incongruent words.
- The Stroop effect provides some, though limited, evidence that the processing of words in the brain interferes with the colour naming task despite the explicit intention to do otherwise.

Example: The Stroop Effect

- named after J. Ridley Stroop.
- the task is to report the colour of the ink as quickly as possible without reading the words.



- Stroop claimed that, on average, it takes longer to report colours of incongruent stimuli than those that are congruent.

the original paper

- Stroop, J.R. (1935). Studies of interference in serial verbal reactions. J. Exp. Psychol., 18:643-662.
- Stroop (1935) is one of the most cited studies in psychology research.

experimental stimuli

- congruent = the word is the word for the colour of the ink.

GREEN

- incongruent = the word is the word of a different colour.

GREEN

theory

- why is the effect thought to occur?
- top-down control over information processing is limited.
- humans appear incapable of entirely switching-off word reading when words are presented in the visual field.
- words are sometimes read more quickly than colours can be reported.
- people are more experienced at reading words than at reading their colours.

experimental design

- The hypothesis concerns the relative effect of congruent and incongruent colour words on **Reaction Time** (**RT**).
- The hypothesis concerns a **population**. Sometimes, this is all normal humans.
- From the population we take a **sample** of size **N participants**.
- Stroop experiments typically use a **within-participant design**.
- In a within-participant design all participants take part in all **conditions**.
- The Stroop experiment has two conditions: One with congruent stimuli and the other with incongruent stimuli.

how is the experiment conducted?

- typically...
- with sequentially presented stimuli.
- multiple participants
- each participant receives both congruent and incongruent stimuli (a within-subject design).

GREEN

GREEN

what do the data look like?

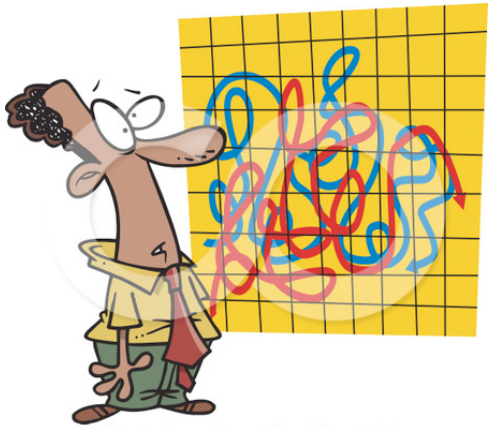
UserID	T	Cond	Word	Color	Response	Time
74229	15	IncW	YELLOW	G	G	896
74229	16	IncW	GREEN	B	B	1472
74229	17	IncW	YELLOW	R	R	1008
74229	18	IncW	BLUE	Y	B	1023
74229	19	IncW	GREEN	R	R	1056
74229	20	IncW	BLUE	Y	Y	1040
74229	21	ConW	YELLOW	Y	Y	1548
74229	22	ConW	RED	R	R	840
74229	23	ConW	YELLOW	Y	Y	640
74229	24	ConW	RED	R	R	752
74229	25	ConW	GREEN	G	G	815
74229	26	ConW	YELLOW	Y	Y	800
74229	27	ConW	RED	R	R	736

factors

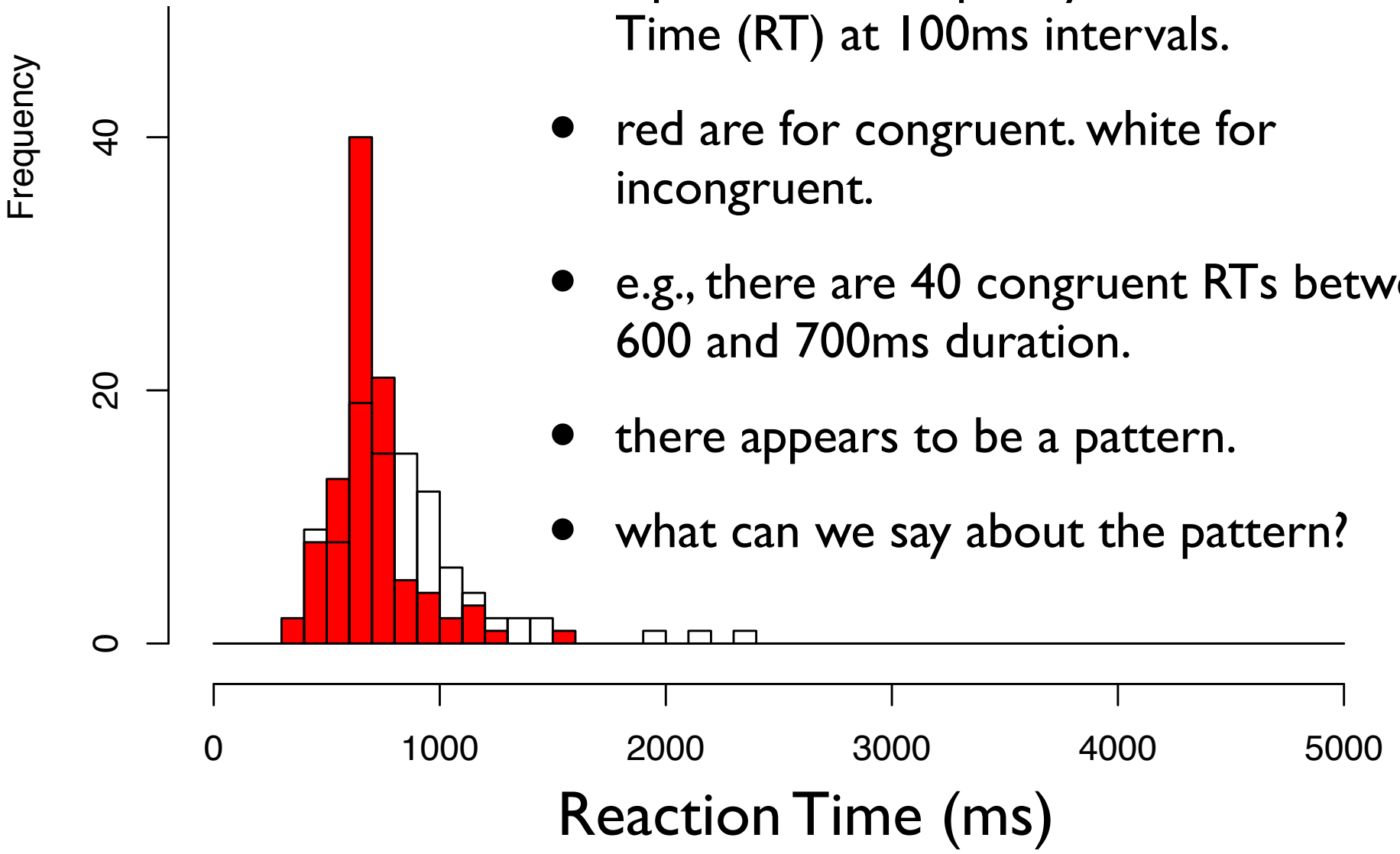
- note that unlike with the Social Capital data, the data here are **factored** into a "long-form".
- each response is represented on a separate row along with its **factor levels**.

how do we make sense of the data?

- this is a fraction of the raw data from just one participant!
- reaction times in 1000ths of a second (milliseconds).
- the participant has made some errors.
- multiple stimuli in multiple experimental conditions.
- is there evidence that people take longer to process incongruent words?



a frequency plot (a histogram)

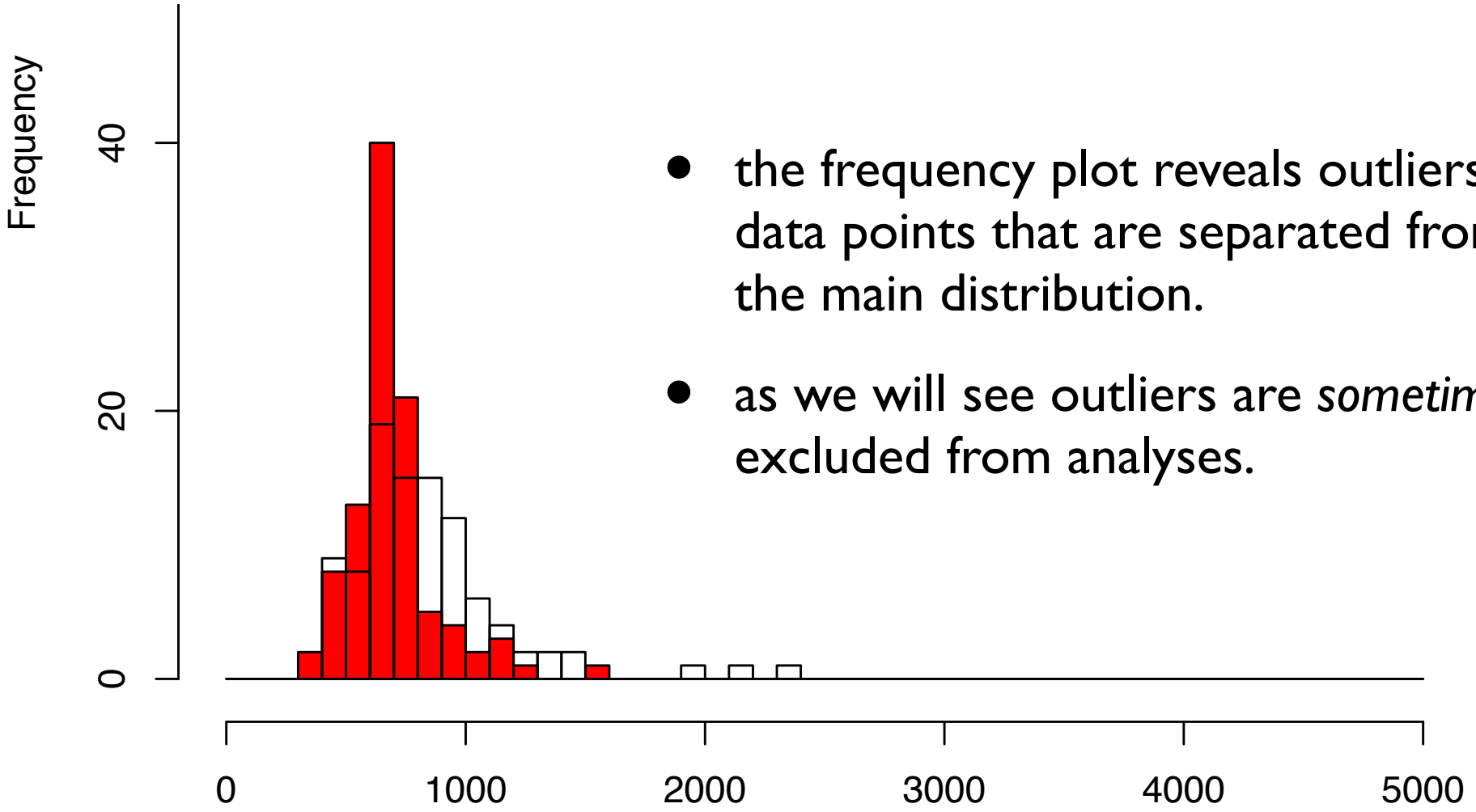


- a plot of the frequency of each Reaction Time (RT) at 100ms intervals.
- red are for congruent. white for incongruent.
- e.g., there are 40 congruent RTs between 600 and 700ms duration.
- there appears to be a pattern.
- what can we say about the pattern?

features of the frequency plot

- There are more values in the middle than at the extremes.
- it is **noisy**. While there is a pattern the curve is not perfectly smooth.
- it is **skewed**. The frequency distribution has a long-tail to the right. (there is a limit on how fast you can be (to the left) but no limit on how slow.)
- these are general properties of human reaction time curves though for tasks that take a longer duration the curves become progressively less skewed.

outliers



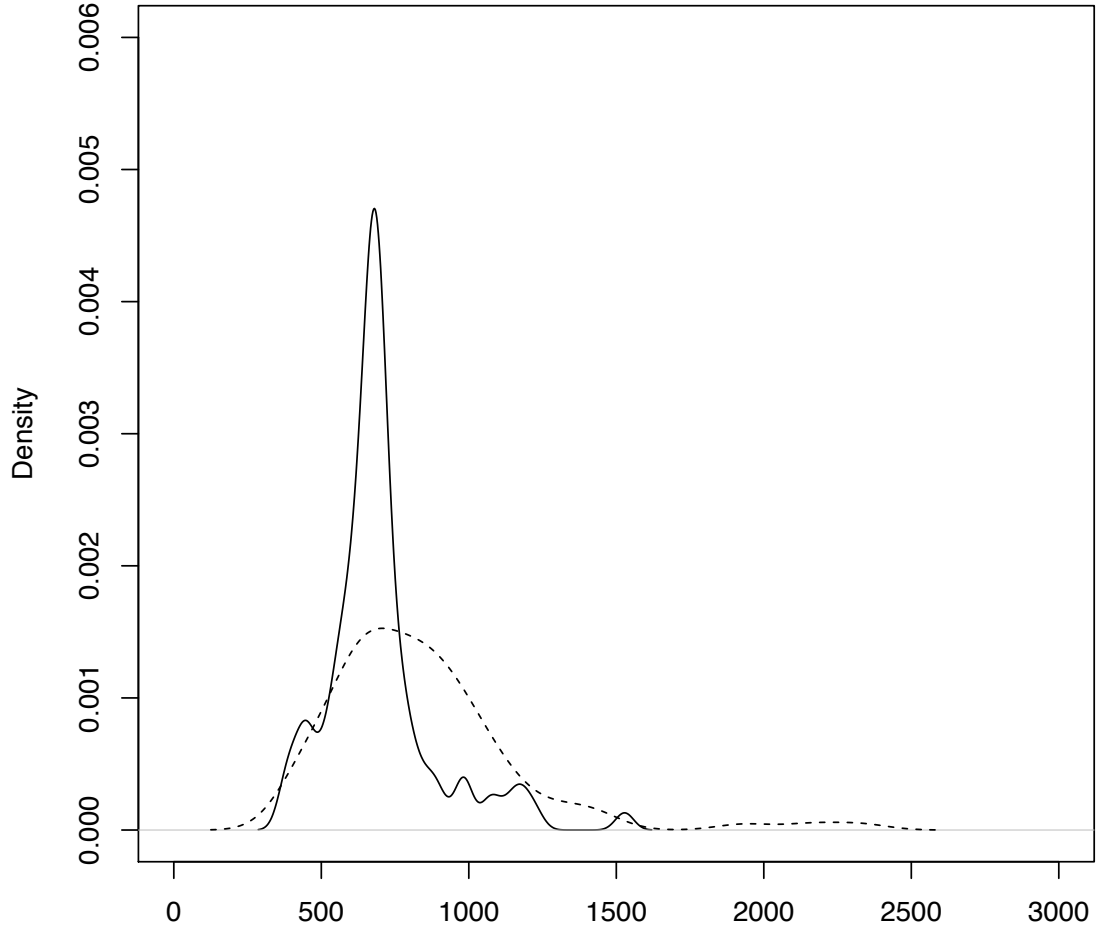
- the frequency plot reveals outliers -- data points that are separated from the main distribution.
- as we will see outliers are *sometimes* excluded from analyses.

why do response times vary?

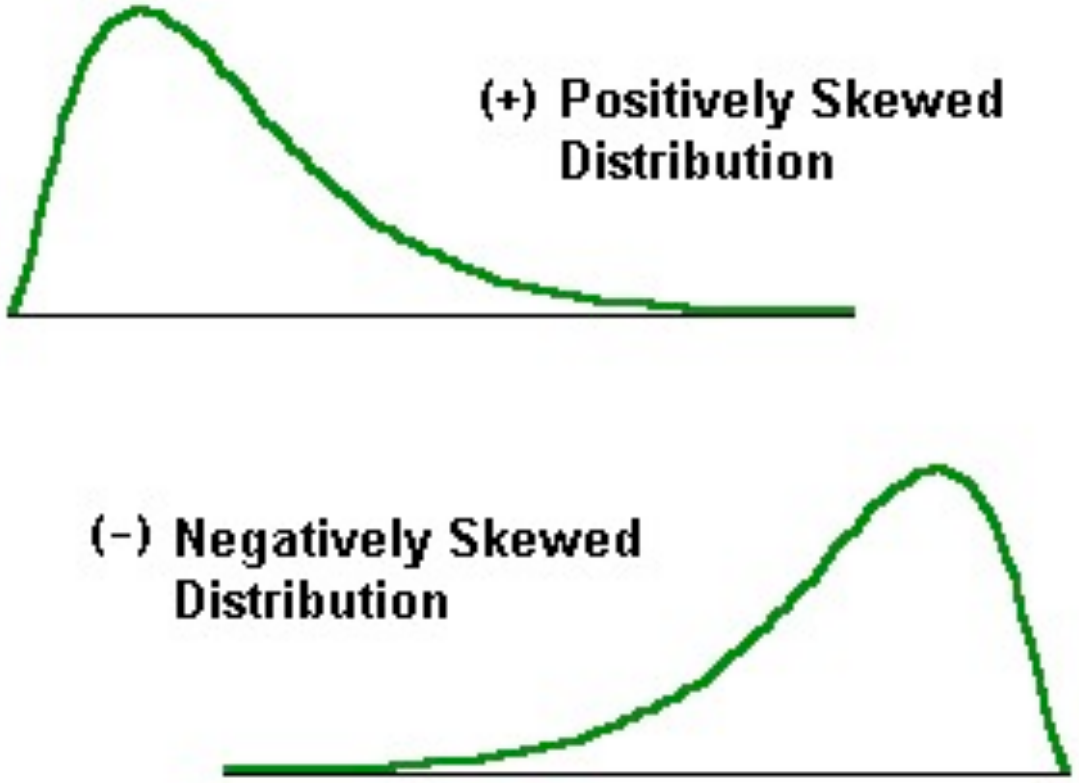
- the cognitive neural system is subject to noise - random disturbances of signal.
- smell, for example, is affected by thermodynamic noise because molecules arrive at the receptors at random rates. Similarly for vision and photons.
- perceptual amplification processes can add further noise.
- noise in neuron firing is also relevant.
- to generate movement neuronal signals are relayed and converted by mechanical forces in their muscle fibres. All of these processes are noisy.
- Together these various systems, and others, lead to trial-to-trial variation.

density - another way to plot Reaction Time

- density is proportional to the probability of drawing a value close to X from the population.

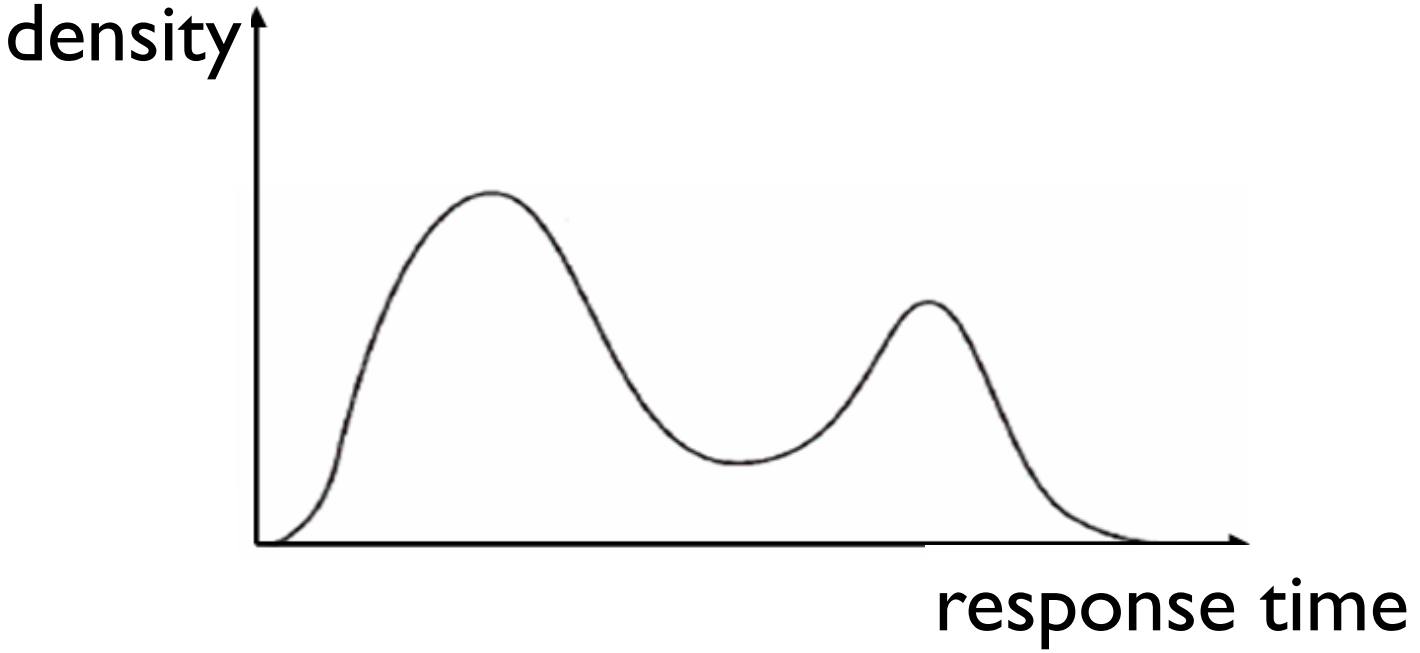


Reaction Times are "positively" skewed



Ratings of products on Amazon, e.g. tend to have a negative skew (high ratings are more likely than low ratings).

bimodal distribution



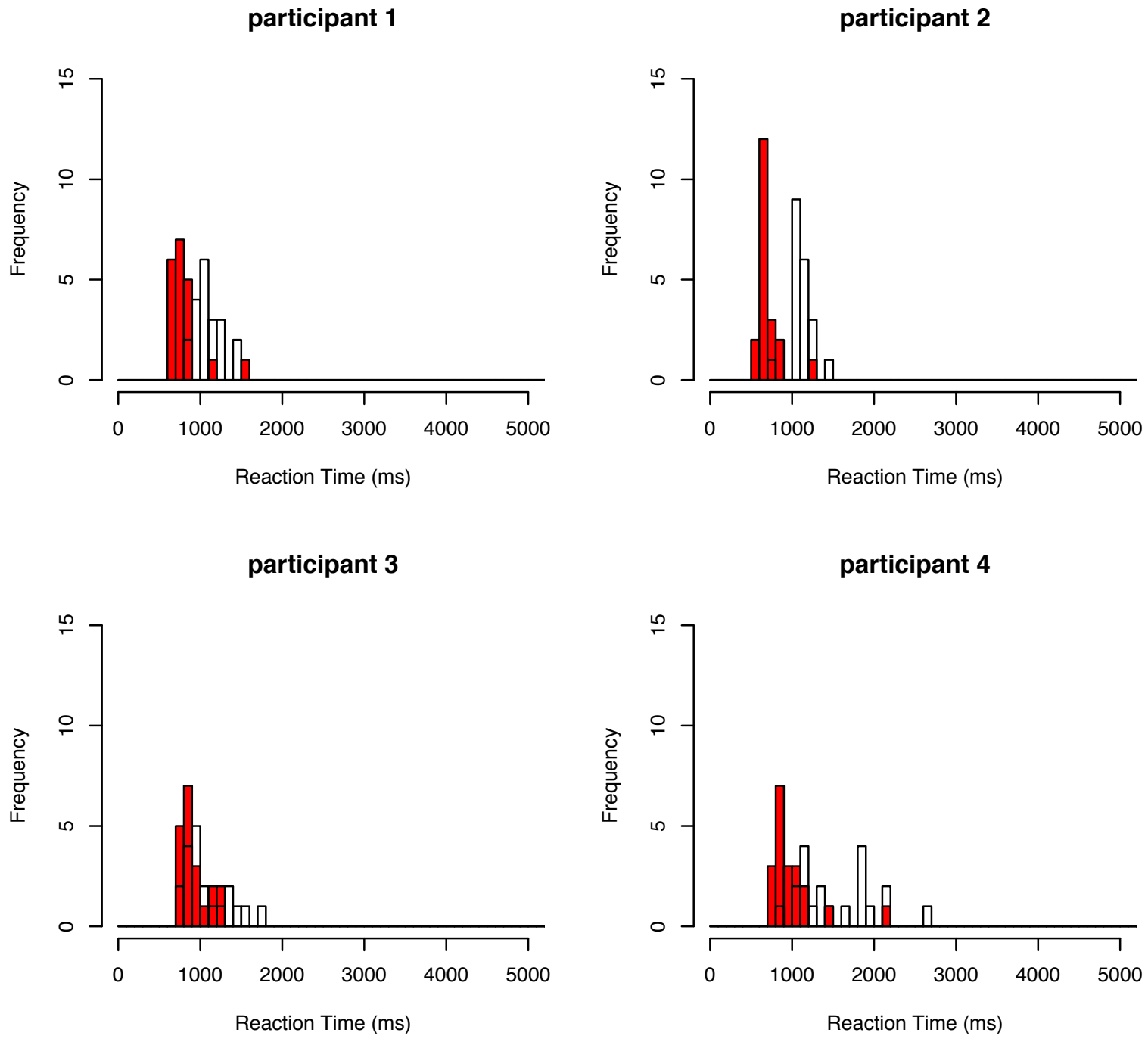
bimodal distributions might be found, for example, because of errors (e.g. failure to recall).

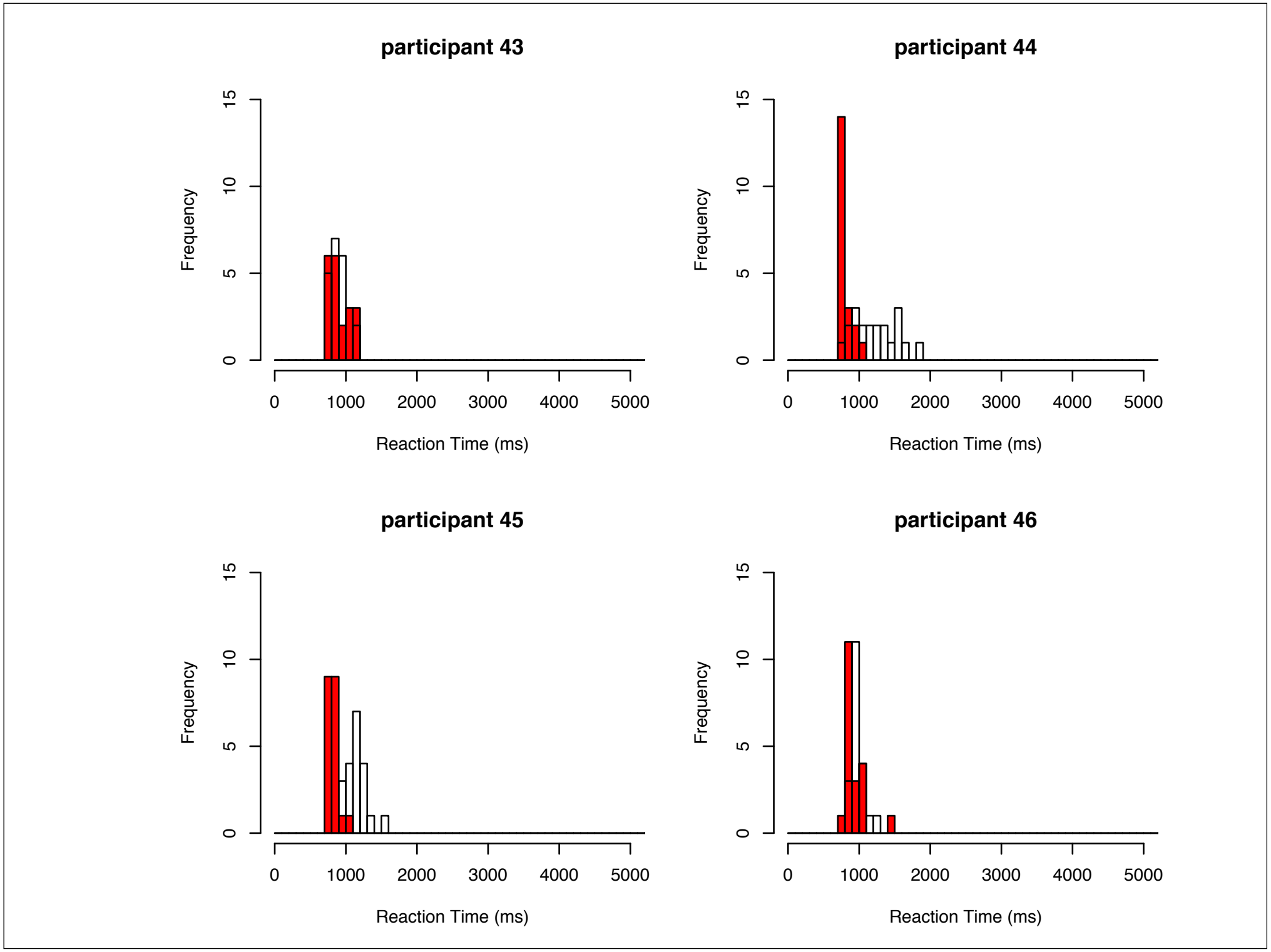
skew, bimodality.

- Variation, skew and bimodality are important properties because they tell us something about the underlying processes that generate these distributions.
- They are also important because many statistical tests should only be applied to raw data if it does **not** have these properties.

sampling

- Typically, in a study of human behaviour we do not measure just one participant, but rather a **sample** of a **population**.
- We do this because we are interested in making claims, and testing hypotheses, that concern populations.
- We do not want to conclude only that, "Andrew was slower ..."
- We do want to conclude that, "People are slower ..." ... but we want to do so without testing EVERYONE.
- We want to support claims that humans, in general, have particular qualities.
- For the Stroop experiment, data from multiple participants might look like...


participant 1
Frequency
0
5
10
15
0
1000
2000
3000
4000
5000
Reaction Time (ms)
participant 2
Frequency
0
5
10
15
0
1000
2000
3000
4000
5000
Reaction Time (ms)
participant 3
Frequency
0
5
10
15
0
1000
2000
3000
4000
5000
Reaction Time (ms)
participant 4
Frequency
0
5
10
15
0
1000
2000
3000
4000
5000
Reaction Time (ms)


participant 43
Frequency
0
5
10
15
0
1000
2000
3000
4000
5000
Reaction Time (ms)
participant 44
Frequency
0
5
10
15
0
1000
2000
3000
4000
5000
Reaction Time (ms)
participant 45
Frequency
0
5
10
15
0
1000
2000
3000
4000
5000
Reaction Time (ms)
participant 46
Frequency
0
5
10
15
0
1000
2000
3000
4000
5000
Reaction Time (ms)

central tendency

- These data are from a sample of 57 participants that was collected at the University of Michigan. (available at APA web site.)
- **N = 57**.
- When we have so much data from so many individuals, what do we do?
- How do we draw conclusions about the population given the sample?
- First we need a measure of central tendency then we need to contrast the resulting distributions.
- In other words we need to contrast distributions of means.

central tendency

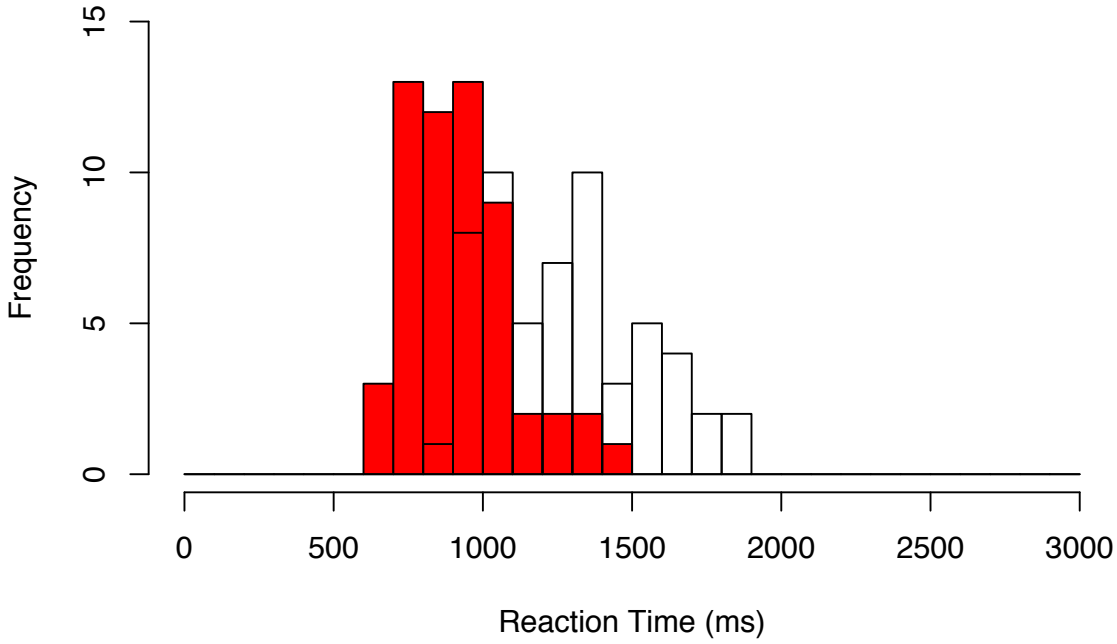
- Consider the following 3 data sets:
- {1, 1, 1, 2, 2, 3, 4, 5, 5, 6}: Mean = 3, Median = 2.5, Mode = 1
- {1, 27, 28, 29, 30}: Median = 28, Mode = XXX
- {1, 1, 1, 1, 1, 1, 2, 2, 3, 3, 3, 3, 50}:Mean =5.5, Median = 2, Mode = 1

Mean, mode, median advantages/ disadvantages

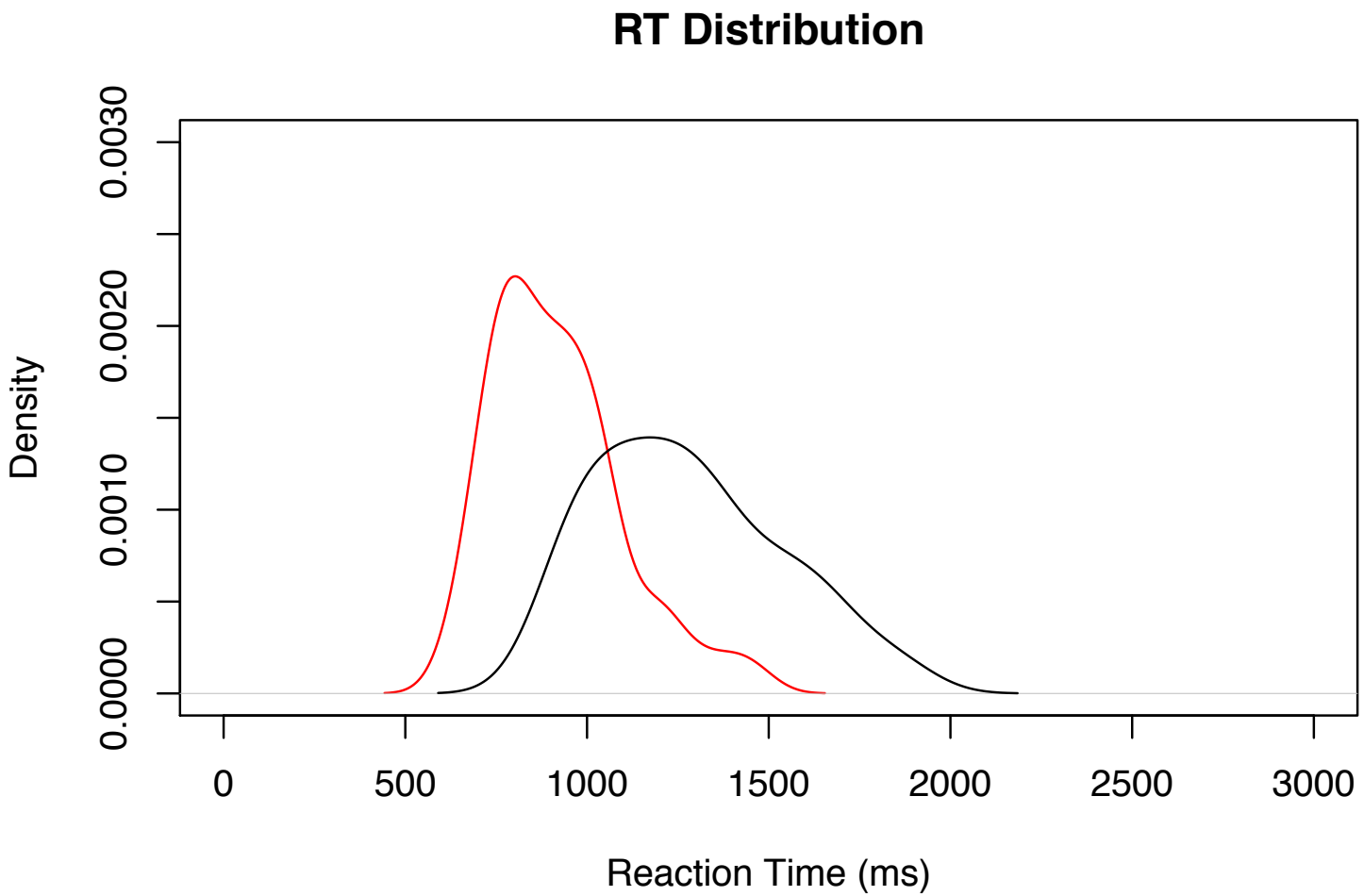
- The **mode** is a score that actually occurred,
- ... whereas the mean and sometimes the median may never have been present in the data.
- If 70% of your customers want "large" T-shirts and 30% want "small", then it does not make sense to stock the mean or "medium" size t-shirt.
- The major advantage of the **median** is that it is not affected by large extreme scores.
- The **mean** is by far the most commonly used.
- The sample mean is, in general, a better estimate of the central tendency of the population than either the median or the mode.

distributions of means

- the skew is considerably diminished.
- in fact, the **distribution of means** approaches a **normal** distribution.
- ... in this case, it is true for both the congruent and the incongruent Stroop data.

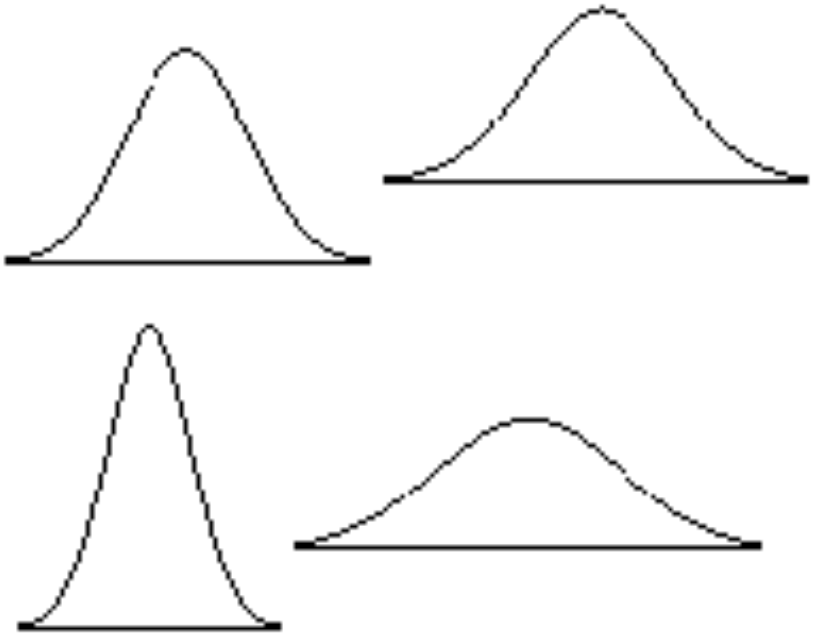


density plots



the Normal distribution

- Normal distributions have a distinctive bell shape density.

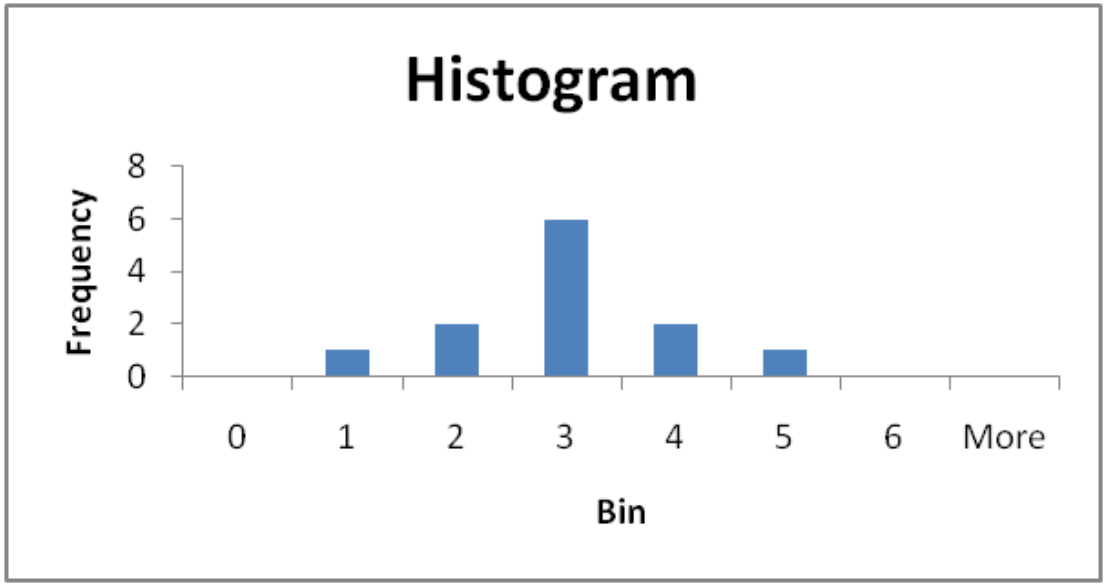


how can we analyse the distributions of means?

- there are many thousands of data points.
- the first step is to summarise the data from each participant.

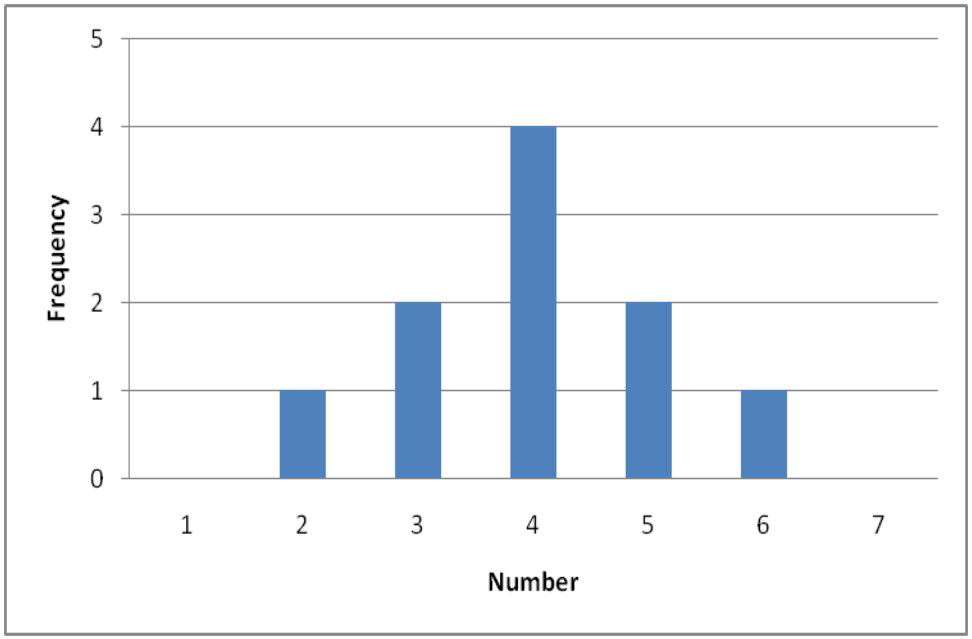
mean, mode, median for a Normal distribution

- Consider the data:
- {1, 2, 2, 3, 3, 3, 3, 3, 3, 4, 4, 5}
- Mean is (1+2+2+3+3+3+3+3+3+4+4+5)/12 = 36/12 = 3
- Median = 3
- Mode = 3



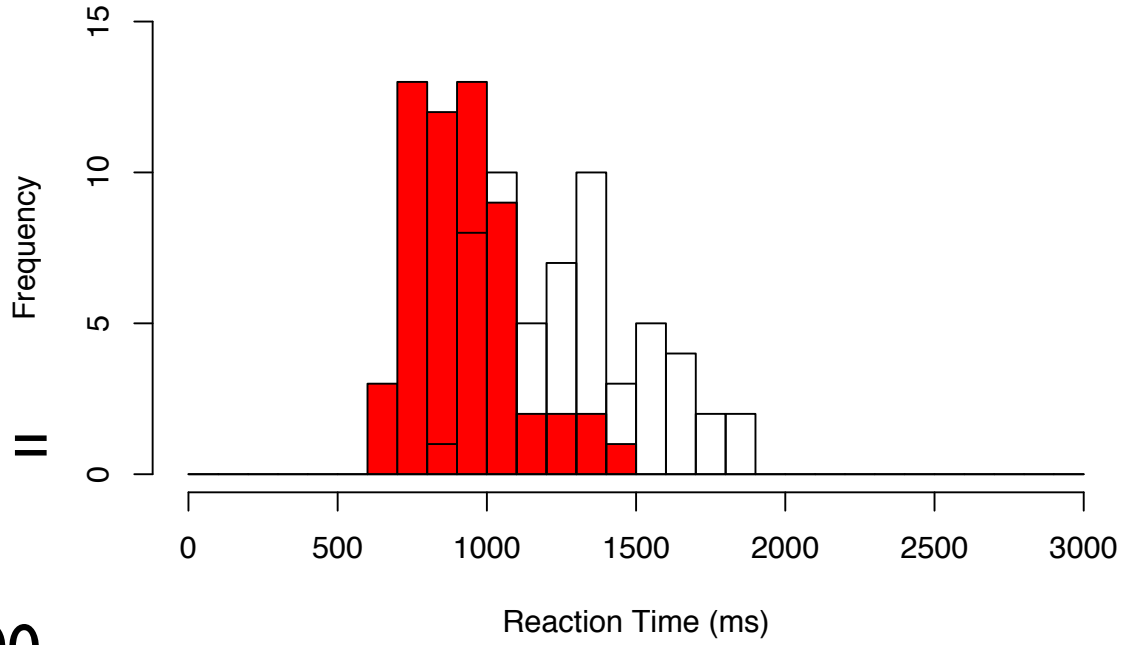
range

- Range is one measure of Spread/Variability
- DATA: {2, 3, 3, 4, 4, 4, 4, 5, 5, 6}
- (number of years since left school)
- Range is the difference between the maximum and minimum score
- Range = 6-2 = 4



distributions of means

- Mean
- congruent = 918ms.
- incongruent = 1279ms.
- Range
- congruent = 1500 - 600 = 900ms.
- incongruent = 1900 - 800 = 1100ms.
- The distributions overlap.

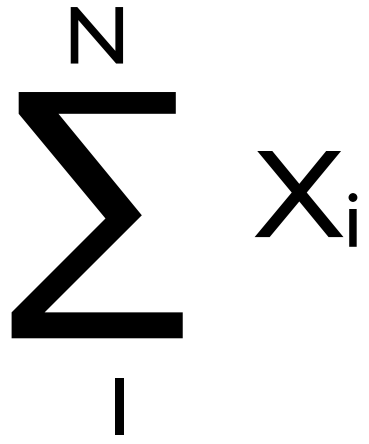


Notation

- DATA: {2, 3, 3, 4, 4, 4, 4, 5, 5, 6}
- We can refer to a set of scores, such as above as X.
- An individual number say 6 can be referred to with a subscript, say X_{10}.
- To refer to a single score without referring to which one then we can refer to X_i.
- We also need the summation symbol, sigma.

Summation

- Sum all of the X_is from i=1 to i=N.



Mean $\overline{X}$

$$\overline{X} = \frac{\sum_{I}^{N} X_i}{N}$$

central limit theorem

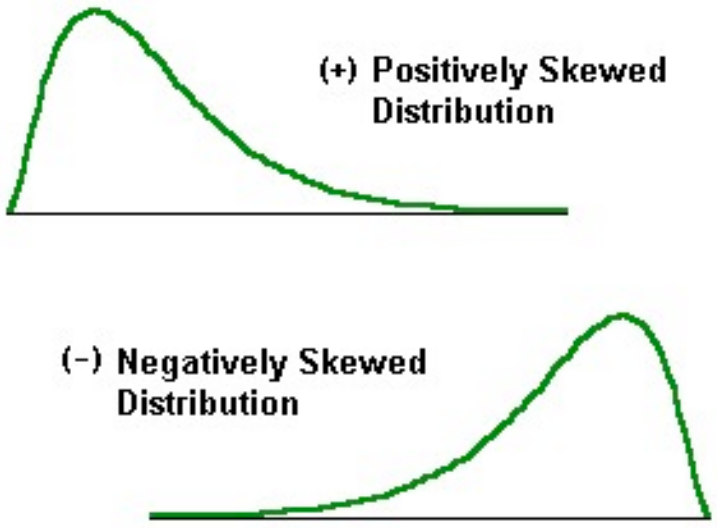
- We have seen that, despite skewed individual distributions, the distribution of mean Stroop RTs appears approximately Normal.
- In fact, there is a mathematical theorem, the Central Limit Theorem, that states that conditions under which the distribution of means will be approximately normal.
- In other words, the Central Limit Theorem tells us about the distribution of means that we would expect if we drew an infinite number of samples from the population and calculated the mean for each sample.

Central Limit Theorem

- ... is one of the most important theorems in statistics.
- It tells us that as N increases, the shape of the sampling distribution approaches normal, whatever the shape of the parent population.
- The rate at which the sampling distribution of the mean approaches normal is a function of the shape of the parent population.
- If the population itself is normal, then the sampling distribution of the mean will be exactly normal regardless of N.

implications of a skewed population for the sampling distribution of the mean

- If the population is markedly skewed, then larger sample sizes are required before the distribution of means approaches normal.



implications of a symmetric and unimodal but non-normal population for the sampling distribution of the mean.

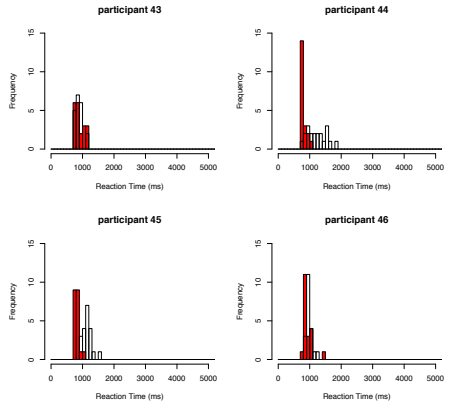
- If the population is symmetric and unimodal but non-normal, the sampling distribution of the mean will be nearly normal even for quite small sample sizes.

sampling distribution of the median

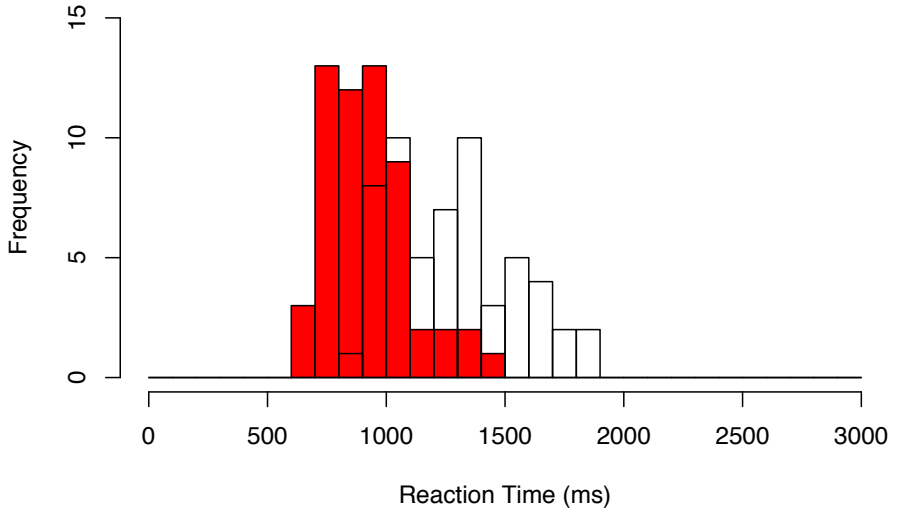
- Is not generally normal except, perhaps, for very large sample sizes.

Returning to the Stroop data

* individual RT distributions.
* one for each participant.



* sampling distributions of the mean.
* one for each condition (congruent / incongruent.)



Hypothesis Testing

- We did not obtain a random sample of Reaction Times to congruent and incongruent words just so that we can draw frequency plots.
- Rather we want to test the hypothesis that the effect of incongruence is longer RTs.
- Having performed a **t-test** we are able to make the following statement:
- **There was a significant effect of incongruence $t(56) = 15.58$, $p < 0.001$ on reaction times.**
- 15.58 is the t statistic and it will be introduced in week 5.

R functions used in this lecture

- hist()
- density()
- plot()
- lines()
- t.test(incongruent, congruent, paired=TRUE, alternative="greater")
- legend()
- plot(density())
- pdf(height=11, width=8.5, file="means.pdf")
- par(omi=c(1,1,1,1))
- par(mfrow=c(2,2))
- for(i in 1:N) { }

Evaluation Methods and Statistics
Lecture 3

Professor Andrew Howes
Dr Ben Cowan

School of Computer Science
University of Birmingham

previously...

- Lecture 1: Why we need evidence-based argumentation. Example of Social Capital and facebook use.
- Lecture 2: Introduction to distributions.
  - Used an example of evidence that people have limited top-down control over sensorimotor processing (the Stroop effect).
  - Skewed distributions of Reaction Time (RT)
  - Normal distribution of means.
  - Showed how the same paradigm could be used to understand the effect of search engines on memory (transactive memory).

data types

- Data types:
- Nominal. **Categorical** data, e.g. names, participant identifiers.
- Ordinal. Is data that can be ordered, e.g. a person's favorite ice creams.
- Interval. Is like ordinal data but with addition that we know the size of the gaps between points in the scale.
- **Ratio**. Is like interval data but has a zero point. E.g. Reaction Times.

review the R practical class

- data can be stored in a data.frame
- the elements of a data.frame can be indexed.
- e.g. mydata[2,10] == 943

```
> mydata <- read.csv("StroopData3.csv" )
>
> mydata[1:5,]
   X ClassID UserID NumTrial Condition ColorOrWord WordDisplayed
1 41    4331  74229        1      IncW           C         YELLOW
2 42    4331  74229        2      IncW           C           BLUE
3 43    4331  74229        3      IncW           C         YELLOW
4 44    4331  74229        4      IncW           C            RED
5 45    4331  74229        5      IncW           C          GREEN
  ColorOfStimulus ColorOfResponse ReactionTime
1               G               G         1235
2               R               R          943
3               B               B         1008
```

format

- pdf() can be used to start a new pdf document. Subsequent plot() hist() etc. commands will be written to this document.
- par can be used to set various parameters.
- variables such as xx and yy can be defined.

```
> pdf( height=11, width=8.5,
file="individuals.pdf" )
> par( omi=c( 1,1,1,1 ) )
> par( mfrow=c(3,2) )
> xx = c(0,5000)
> yy = c(0,15)
```

factors

- factors provide compact ways to handle categorical data.
- A factor is a vector object used to specify a discrete classification (grouping) of the components of other vectors of the same length.

```
> IDs <- levels( factor( mydata$UserID ))
>
> IDs[1:5]
[1] "74229" "74626" "74652" "74653" "74654"
> ID[1]
[1] “74229”
```

for()

```
> for( i in 1:5 ) {
+   print(i)
+ }
1
2
3
4
5
```

subset()

```
> subject <- c(1,1,2,2,3,3)
> condition <-
c("con","inc","con","inc","con","inc")
>
> RT <- c(345,234,678,123,890,1024)
>
> toy <- data.frame( subject, condition, RT )
>
```

```
> toy
  subject condition   RT
1       1       con  345
2       1       inc  234
3       2       con  678
4       2       inc  123
5       3       con  890
6       3       inc 1024
>
> subset( toy, toy$subject == 1 )
  subject condition  RT
1       1       con 345
2       1       inc 234
>
```

putting for() and subset() together

```
> for( i in 1:n ) {
+   I <- subset( mydata, mydata$UserID == IDs[i] )
+   congruent <- subset( I, I$Condition == "ConW" )
+   hist( congruent$ReactionTime, ylim=yy, xlim=xx,
breaks=brks, col=2, main=paste("participant",i ),
xlab="Reaction Time (ms)" )
+ }
```

this lecture: basic statistics and experimental design

- significance
- falsification and the null hypothesis
- Independent and dependent variables.
- randomization
- variation
  - sum of squares
  - variance
  - standard deviation
- examples

Limitations on top-down control.

- People have a limited ability to control information processing. One study that supports this claim was reported by Stroop (1935). In a typical Stroop study participants are asked to name the ink colour of colour name words. Congruent colour words are printed in the same colour as the meaning of the word, e.g. the word green is printed in green ink. Noncongruent words are printed in a different colour, e.g. the word blue in red ink. Many studies have observed a **significant** effect of incongruence on reaction times. It takes longer for people to identify the ink colour of incongruent words. The Stroop effect provides some, though limited, evidence that the processing of words in the brain interferes with the colour naming task despite the explicit intention to do otherwise.

significance

- As we saw in the Stroop task there is often a large amount of data that requires summarisation before a statistical test can be conducted.
- Even then, someone wishing to make use of another person's study to support a claim as part of a broader argument further hides the details.
- In the previous paragraphs, for example, the summary consists of:
  - some details about the nature of the data
  - the statement that there was a **significant** finding.

falsification

- Significance tests allow us to test hypotheses.
- A good hypothesis is one that can be rejected (Popper).
- A good hypothesis is **falsifiable**. Consider:
- H0: There are no vultures on the UoB campus.
- H1: There are vultures on the UoB campus.
- Which of these is falsifiable?

evidence of absence

- absence of evidence is not evidence of absence.
- we can look for vultures all day and fail to find them but this does not allow us to reject H1 because all we have is absence of evidence.
- H0 is fundamentally different, we can reject H0 as soon as we see a vulture in the park.
- **H0 is falsifiable. It is the null hypothesis.**

the null hypothesis

- the significance of a statistical test tells us whether or not we can reject the null hypothesis.
- The null hypothesis says "nothing is happening"
- E.g. when we are comparing two sample means, as in the Stroop experiment, the null hypothesis is that the two samples are the same.
- E.g. when we are testing whether there is a correlation between two variables, as in the social capital and facebook experiment, the null hypothesis is that there is no correlation.
- Importantly the null hypothesis is **falsifiable**.
- We reject the null hypothesis when we can show that it is sufficiently unlikely.

significant findings allow us to reject the null hypothesis

- **The results indicated that intensity of facebook use and bonding social capital were correlated, $r(267) = .29$, $p < 0.001$.**
- **There was a significant effect of incongruence $t(56) = 15.58$, $p < 0.001$ on reaction times.**
- In both cases likelihood of the null hypothesis is less than 0.001 and we can therefore reject it.

variables

independent and dependent variables

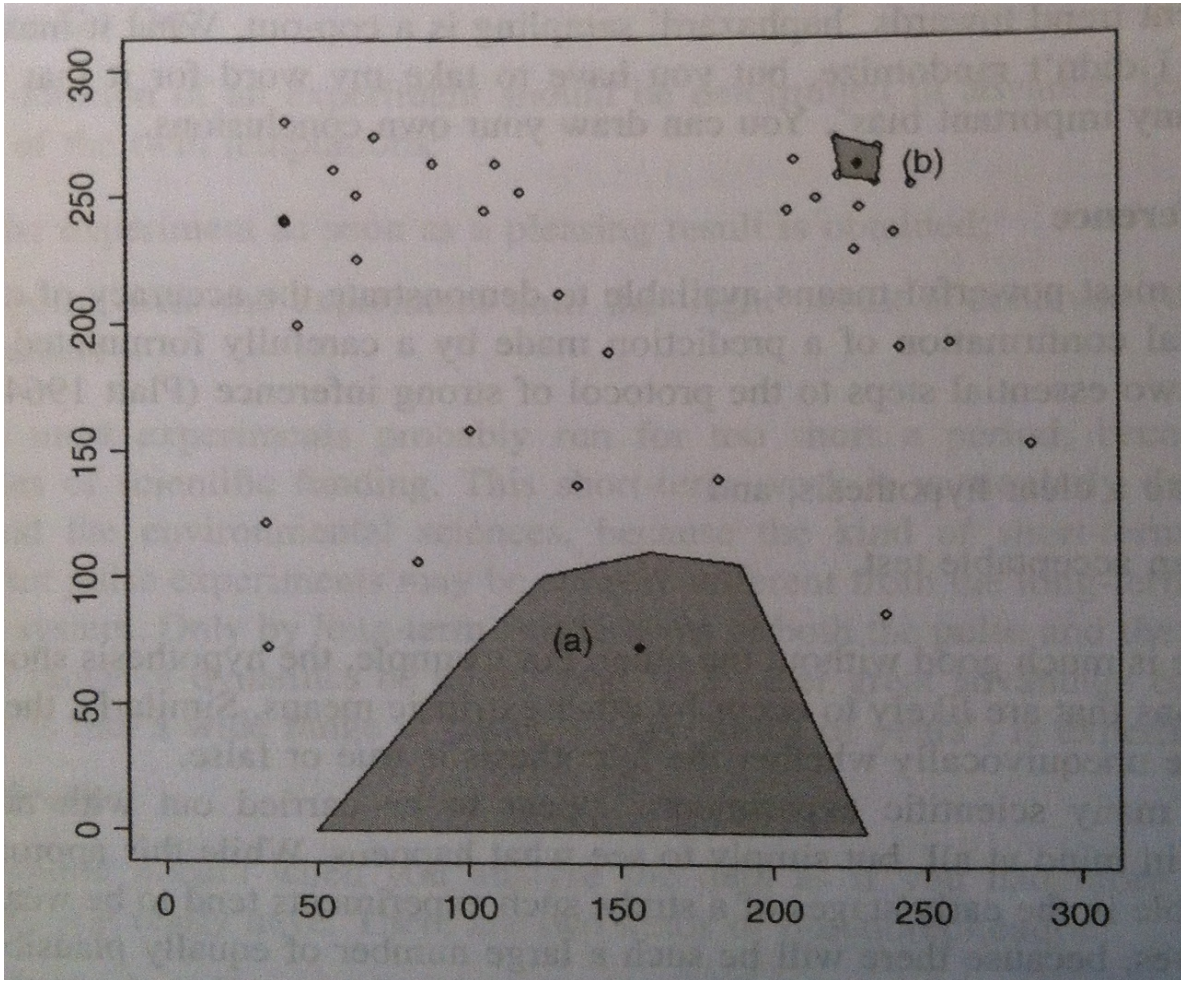
- The **independent** variables define the conditions of the experiment.
- In the Stroop experiment the independent variable was the relationship between the ink color and the word.
- There were two **levels** of the independent variable: congruent and incongruent.
- The dependent variable is what you measure.
- In Stroop the **dependent** variable was the reaction time (RT).

Randomization

- Without randomization we introduce bias into the sample.
- Bias reduces the validity with which the sample represents the population.

Randomization

- Consider the problem of selecting a tree from the forest. One approach would be to generate a series of random x,y coordinates and then take the nearest tree to each.



sampling

- However, this would over-sample trees that are relatively isolated from the others.
- The right approach would be to number the trees individually and then sample randomly.

Randomization

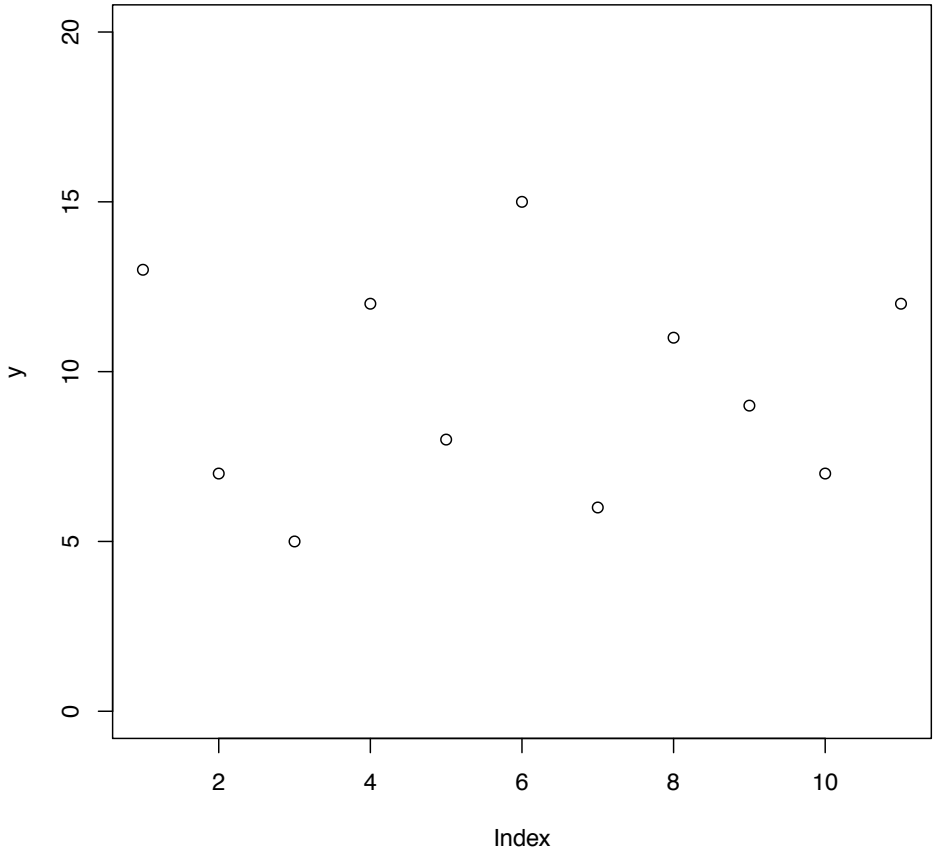
- When we are studying human behaviour we need to select participants randomly if we wish to draw inferences about the population.
- Full randomization from a global population is difficult to achieve.
- There are accepted limits on randomization.
- Student participants introduce age and IQ biases. They often introduce gender and socio-economic biases.
- There are unacceptable biases. E.g. An experiment such as the Ellison et al. facebook study might have measured bonding social capital at one university and bridging social capital another.

Variability

everything varies

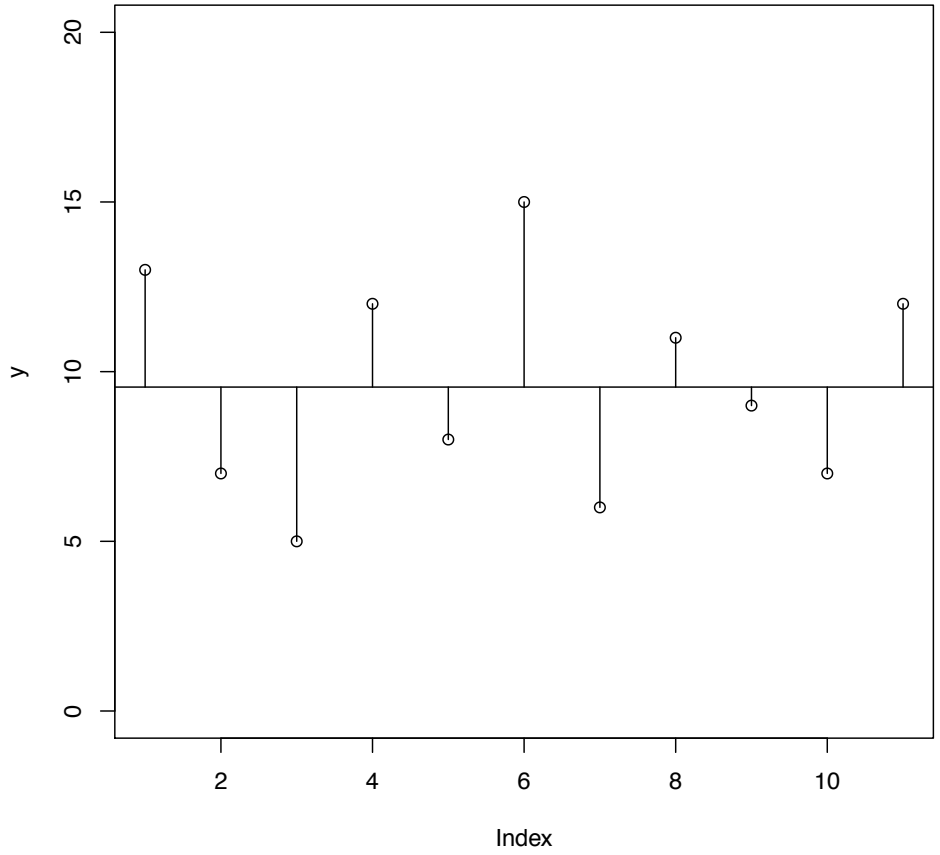
- Perhaps the most important concepts in statistics are those concerned with the representation of variability.
- Range is one simple measure of variation but it fails to capture the extent to which values are clustered, e.g. in a normal distribution.
- What would be a better measure of variability?

variation



- y <- c(13,7,5,12,8,15,6,11,9,7,12)
- plot(y, ylim=c(0,20))

differences between each value and the mean



mean(y)

- y-mean(y)
- [1] 3.4545455 -2.5454545 -4.5454545 2.4545455 -1.5454545 5.4545455
- [7] -3.5454545 1.4545455 -0.5454545 -2.5454545 2.4545455

sum of differences

- The longer the lines then the more variable the data. So, perhaps, use the length of the lines, the difference, as the
- y-mean(y)
- This looks like a promising measure of variability but there is a problem.
- When we add up the lengths of the lines we will get zero.
- [1] 3.4545455 -2.5454545 -4.5454545 2.4545455 -1.5454545 5.4545455
- [7] -3.5454545 1.4545455 -0.5454545 -2.5454545 2.4545455
- but sum(y-mean(y)) always equals zero!

sum of squares

- We could use the sum of the absolute differences (i.e. we ignore the minus signs). This is used in some analyses but some of the maths is difficult.
- Instead we use the **sum of squares**.
- The squared differences can be calculated with (y-mean(y))^2
- [1] 11.9338843 6.4793388 20.6611570 6.0247934 2.3884298 29.7520661
- [7] 12.5702479 2.1157025 0.2975207 6.4793388 6.0247934
- and the **sum of squares** is sum((y-mean(y))^2)

$$\sum (y - \overline{y})^2$$

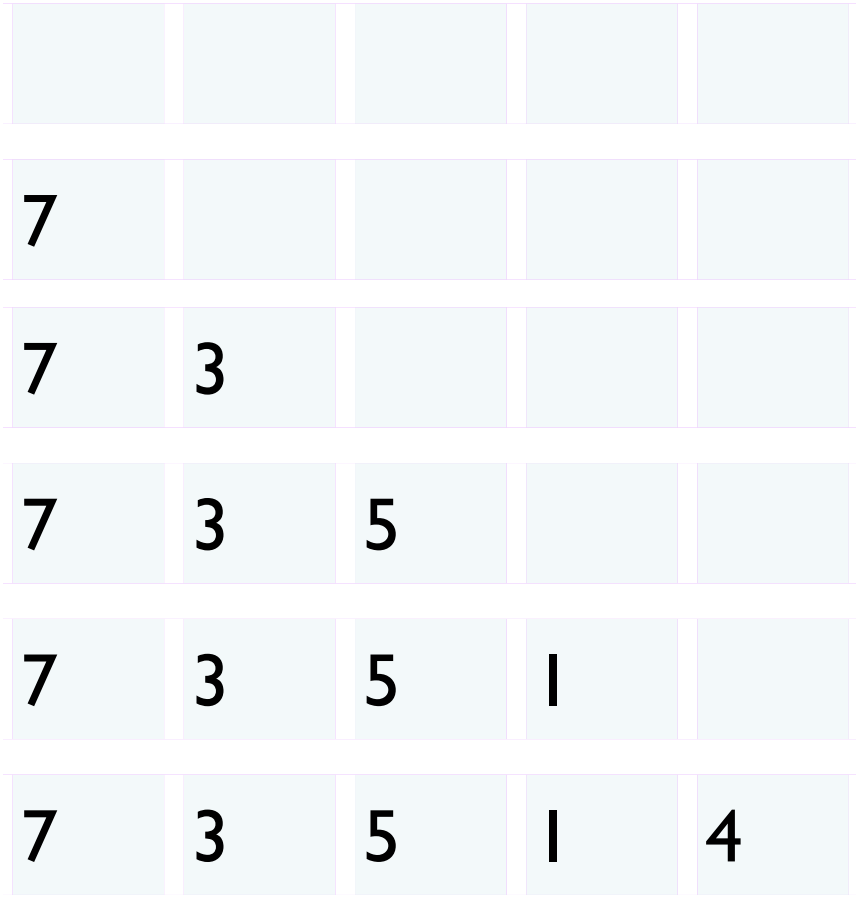
- [1] 104.7273
- The bigger the variability (the further points are from the mean) then the higher the sum of squares.

variance

- unfortunately, the sum of squares not only gets larger as variability increases (which is desirable) but also as we add more values to the vector of data.
- to correct for this, one thing that we could do is divide the sum of squares by the number of observations -- but this would underestimate the population variance!
- instead we must divide by the **degrees of freedom**.
- we need to take a brief diversion in order to understand degrees of freedom.

degrees of freedom

- Say that we have 5 observations.
- And the mean of these observations is 4.
- Then since there are 5 (N=5) observations, the sum must be 20.
- Now imagine that we have a cell for each observation (on the right).
- The last observation can only be one number. There is no "freedom" for the value of the last cell.
- So the degrees of freedom is 4.



degrees of freedom

- In general, the degrees of freedom of a statistical calculation is the number of data points minus the number of parameters that went into the calculation.
- For the calculation of variance the number of parameters is 1.
- the degrees of freedom for the calculation of variance is N-1.

variance

$$\text{variance} = \frac{\text{sum of squares}}{\text{degrees of freedom}}$$

variance

$$s^2 = \frac{\text{sum of squares}}{\text{degrees of freedom}}$$

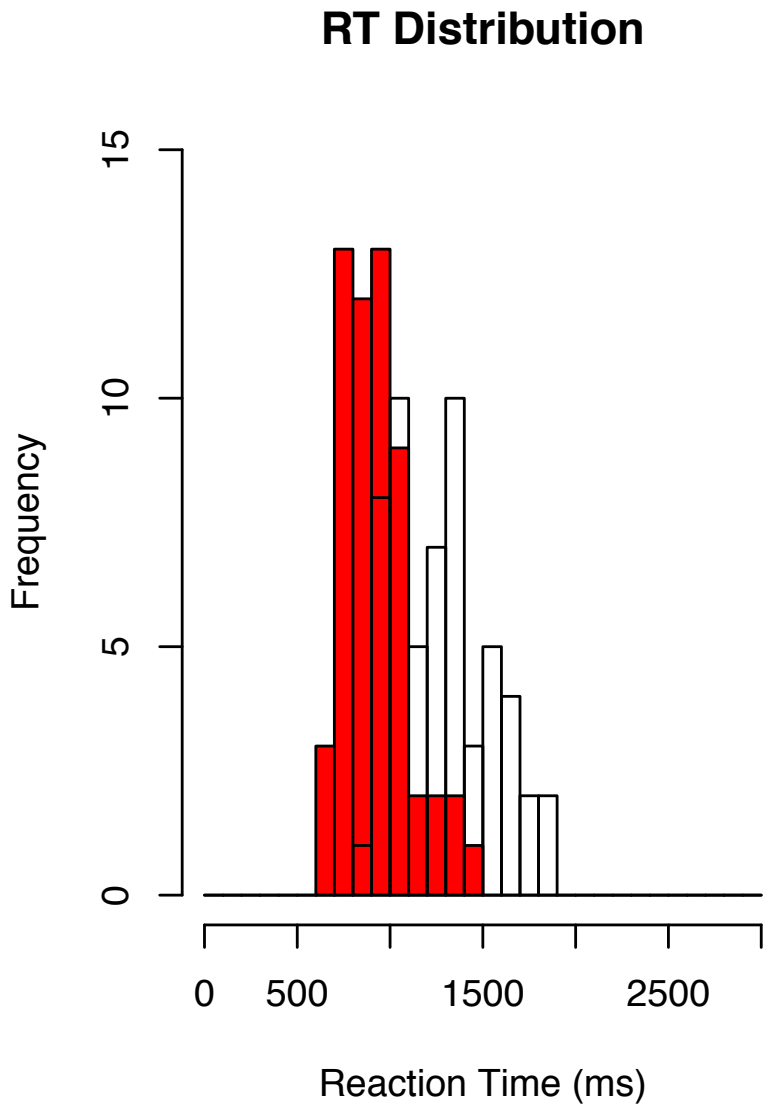
variance

$$s^2 = \frac{\sum (X - \overline{X})^2}{\text{degrees of freedom}}$$

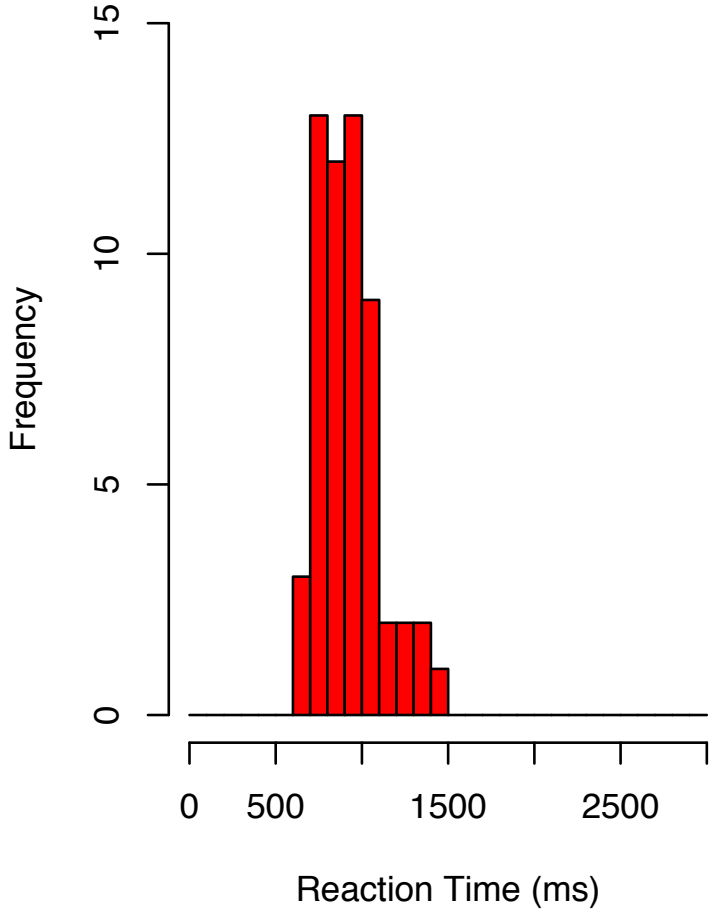
variance

$$s^2 = \frac{\sum (X - \overline{X})^2}{N - 1}$$

variance of Stroop data



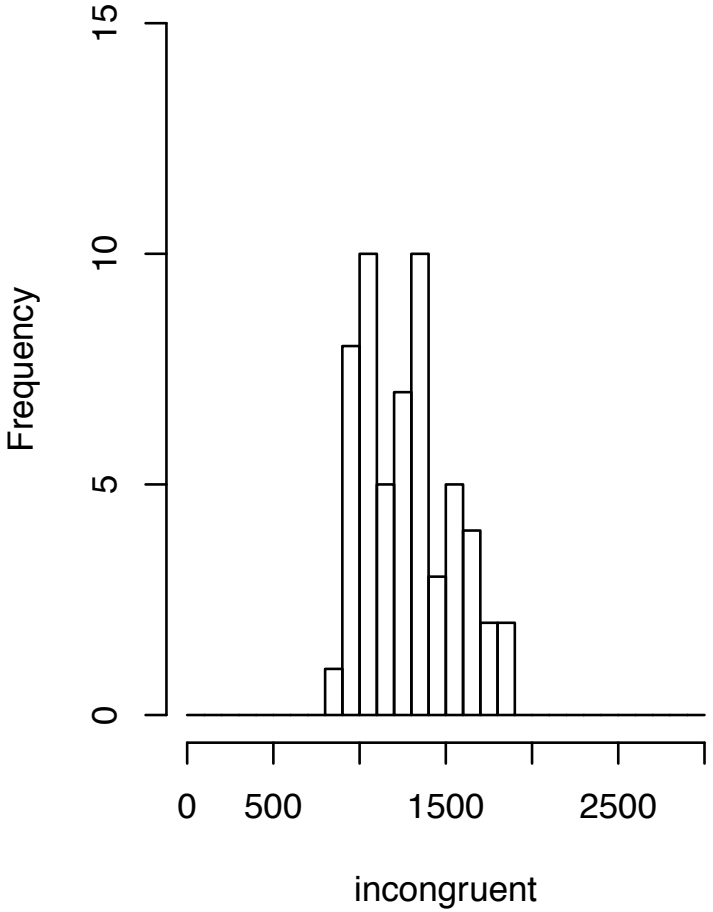
Congruent



mean = 918ms

variance = 32657ms^2

Incongruent



mean = 1278ms

variance = 66022ms^2

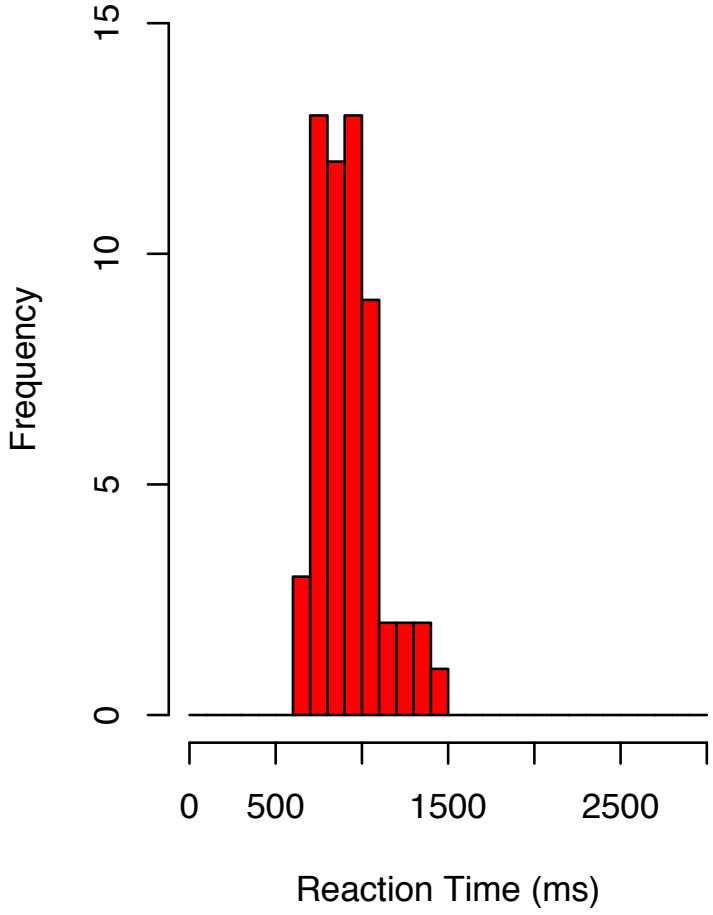
unequal variance

- It is important to know if there is unequal variance in two samples if the appropriate statistical test is to be selected.
- [More later.]

standard deviation

- ... we define the standard deviation *s* as the square root of the variance.

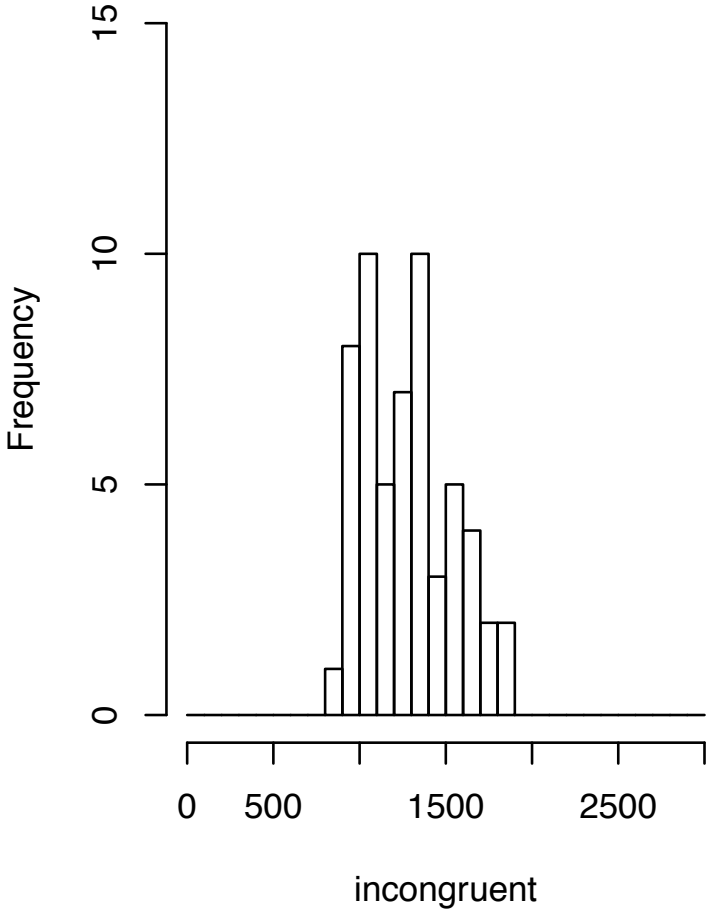
Congruent



mean = 918ms

standard deviation = 180.71ms

Incongruent



mean = 1278ms

standard deviation = 256.95ms

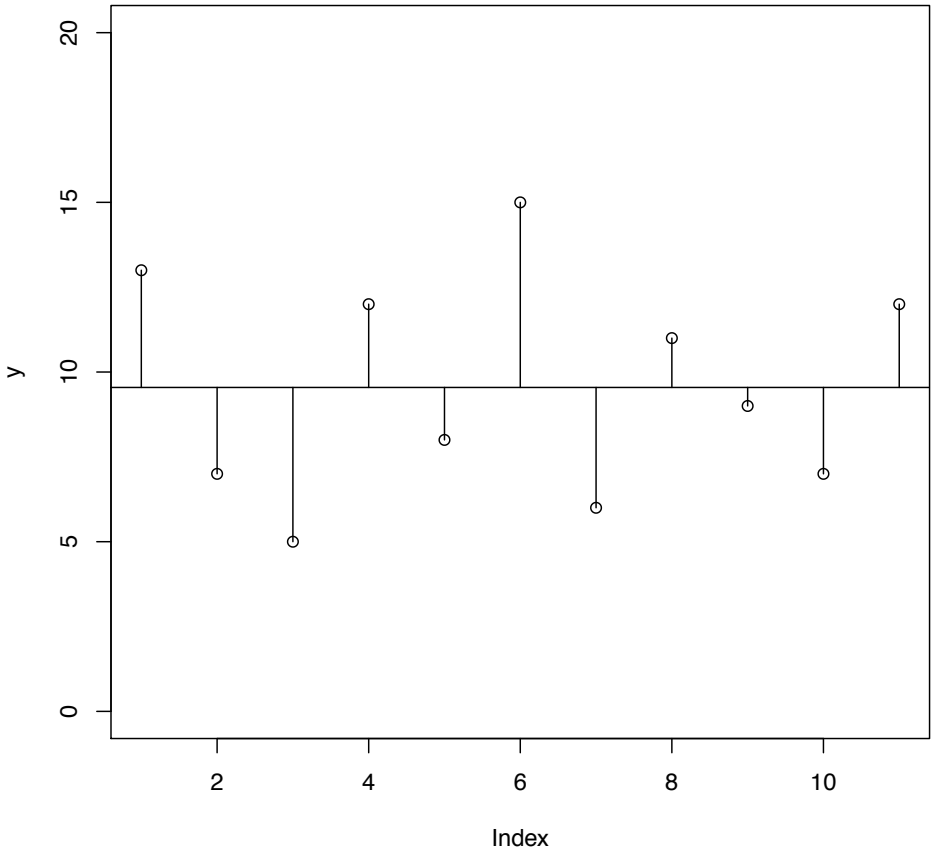
R

R CMD BATCH filename.r

- ... and use filename.r.Rout for debugging.
- also use the command print(paste()) to print out variable values.

differences between each value and the mean

- m <- mean(y)
- plot(y,ylim=c(0,20))
- lines(c(0,20), c(m,m))
- x <- seq(1,11)
- for(i in 1:11) {
- lines(c(x[i], x[i]), c(m, y[i]))
- }



m

variance

- variance <- function(b)
- {
- sum((b - mean(b))^2) / (length(b) - 1)
- }

- variance(y)
- [1] 10.47273
-

Ethics & Assessing papers

Ben Cowan & Andrew Howes
Lecture 5

Office Hours

- Room 134- Computer Science
- Tuesdays 1-3pm
- Email- b.r.cowan@cs.bham.ac.uk
- Email me to arrange a time

What will be covered

- Ethics
  - Major points to consider in experimentation
  - Deception
- Assessing a research paper
  - What each section should say
  - The importance of method and results sections
- The concept of social capital
  - What is it?
  - Questionnaire and data for assessment

Experiments with human participants

- Voluntary consent is absolutely essential.
- Experiments should yield fruitful results for the good of society...
- The experiment should be conducted as to avoid all unnecessary physical and mental suffering and injury

Experiments with human participants

- No experiment should be conducted where there is an a priori reason to believe that death or disabling injury will occur
- Degree of risk should never exceed that determined by humanitarian importance
- Proper preparations should be made and adequate facilities provided to protect the experimental subject

Experiments with human participants

- The experiment should be conducted only by scientifically qualified persons
- Participant should be at liberty to bring the experiment to an end
- The scientist in charge must be prepared to terminate the experiment at any stage...

Experiments with human participants

Participants are NOT crash test dummies.

There are many code of ethics

- Association of Social Anthropologists of the UK and Commonwealth.
- British Association of Social Workers - Code of Ethics.
- British Educational Research Association - Ethical Guidelines.
- British Psychological Society - Code of Ethics and Conduct.
- The Chatham House Rule.
- ESRC Research Ethics framework.
- National Children's Bureau Research guidelines.
- (see http://www.rcs.bham.ac.uk/ethics/links/index.shtml)

Safety guidelines

- Social Research Association - Code of practice for the safety of social researchers

- Universities & Colleges Employer Association - Safety in fieldwork and guidelines for working overseas.

Legislation

- Data Protection Act 1998
- Equality Act 2010
- Human Rights Act
- Mental Capacity Act 2005
- NHS Act 2006 (section 251)
- Police Act 1997
- Safeguarding Vulnerable Groups Act 2006
- Criminal Records Bureau (CRB) checks - Eligible positions requiring CRB

Ethics are not easy

- do not be complacent.
- do not make the mistake of believing that you know it all.
- academics employed by a university are obliged to consult an ethics committee.
- marketers and private individuals are not!

Deception

- Participants should not, *where possible,* be deceived.
- Why might this be an issue for the integrity of an experiment?

Deception

Let's take an example:

- bystander apathy (see Piliavin & Charng, 1990 for a review).
  - What influences a member of the public to help a bystander in need?
  - E.g. a drunk who had fallen over; or a well-dressed business person who had fainted?
- Should we tell the participants the true nature of the study? Why/why not?

Deception

- If a participant knows the desired outcome of the experiment then that is likely to affect their behavior.
- To study bystander apathy we cannot even inform participants that they are in a study prior to observing them without biasing the study!
- So we try and build an ethical reason for deceiving someone.

Debriefing

- Participants should not stay deceived.

- they should be debriefed fully as to:
  - The motivations of the study
  - The condition/s they took part in
  - Given contact details of the researcher for further questions

Over to you…..

You want to test whether fear of certain objects/ creatures is innate (present at birth) or whether it is learned

- What conditions could you have?
- Who are your participants?
- What are the ethical considerations?

Little Albert

- John Watson (a behaviorist), used a 9 month old child as a subject.
- Placed Albert in the middle of a room with a white lab rat. Albert was not scared.
- Over a period of two months Albert was then exposed to various things without any sort of conditioning; a white rabbit, a monkey, masks etc...

Little Albert

- Then Albert was again placed in a room with the rat. However, this time, when the rat was touched by Albert, Watson would make loud sounds behind him.
- When this occurred, Albert would get frightened and begin to cry. Watson continued to do this until eventually, Albert became distressed whenever exposed to the rat.

Little Albert

- Eventually, Albert associated anything fluffy or white with the loud noise.

- Albert was never desensitized to his fear.

Assessing a paper

Papers and Reports

- There are 4 core elements to a scientific paper
  - Introduction
  - Method
  - Results
  - Discussion
- Each serves an important purpose

Introduction

- Laying the ground for the work
  - Introducing the problem/question
  - Using existing literature
  - Argumentation and clear reporting of previous findings
  - Clear statement of hypotheses and aims of the research

Method

Fully describe the research methods

Participants

- How many participants did you use?
- Who were they?
- How were they sampled?

Materials

- What questionnaires did you use? Reference the original authors
- What materials did you use?

Method

Conditions (Independent variables)

- What were the conditions in your experiment?
- Was it within or between participants?
- How did you design them? What were the key manipulations?
- Counterbalancing?

Procedure

- A step by step guide of how the experiment was run
- Payment of participants
- Number of trials
- Debrief procedure

content of your methods section depend on the details of the study conducted.

Results

- Reporting the data analysis
  - What tests were used?
  - What program/packages?
  - Descriptive statistics (mean/standard deviation)
  - Graphs of the data
  - Reporting of statistics in APA style
    - E.g. r (88) = .78, p<.001
  - State whether null hypothesis can be rejected
  - Non technical explanation of the results

Discussion

- Placing the findings in the context of previous work
  - Reiterate findings in non technical way
  - Place findings in wider literature
  - Describe limitations
  - Expand with ideas for future work
  - Conclude with summary

Critiquing a paper

- Introduction
  - Does the "story" make sense? Is argumentation used effectively? Is there good evidence used for claims?

- Method
  - Is the sample selected appropriately? Are the conditions/measures effective? Is the experimental approach actually answering the question?

Critiquing a paper

- Results
  - Are the tests effective in assessing the hypothesis? Are they appropriate? Have assumptions been checked? Has the familywise error rate been controlled for?

- Discussion
  - Does the interpretation of the statistics make sense? Does the interpretation use relevant literature to bolster its claims? Is this literature of good quality itself?

Critiquing a paper

- Why is this important?
  - Lots gets published, not all of it is of good quality
  - Poor papers lead to wasted scientific effort
  - Spot problems and you will avoid wasting time
  - Spot problems and you will be helping the community
  - It is important that science is based on good quality, replicable evidence

Social Capital- Assessment

What is social capital?

understood roughly as "the good will that is engendered by the fabric of social relations and that can be mobilized to facilitate action."

- Resources which are available in one's network

Bridging Social Capital

- Also known as 'weak ties'
- Typically do not provide emotional support.
- But access to individual's outside one's close circle provides access to non-redundant information, resulting in benefits such as employment connections, novel information and perspectives

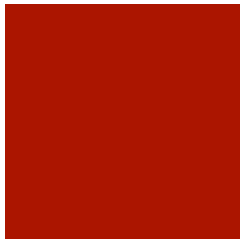
Bonding Social Capital

- Found between individuals in tightly-knit, emotionally close relationships.
  - e.g. family and close friends.
- Highly trusting relationships.
- With e.g., delayed reciprocation.
- Access to social and emotional support

Jung, Gray, Lampe & Ellison (2013)

- Social media (like Facebook) helps build, maintain and benefit from relationships
- Looking at dimensions of social capital as well as favour executions by friends.
- Task: people ask facebook friends for favour to complete a survey
- Found:
  - No relationship between favours and social capital
  - Sub scales of social capital "individual benefit" related to favour asking
  - People who have higher frequency of asking for help from Facebook friends had higher number of responses

Bridging social capital		
Outward-looking	1	Interacting with people in my Facebook network makes me interested in things that happen outside of my town.
	2	Interacting with people in my Facebook network makes me want to try new things.
	3	Interacting with people in my Facebook network makes me interested in what people unlike me are thinking.
	4	Talking with people in my Facebook network makes me curious about other places in the world.
Broader group	5	Interacting with people in my Facebook network makes me feel like part of a larger community.
	6	Interacting with people in my Facebook network makes me feel connected to the bigger picture.
	7	Interacting with people in my Facebook network reminds me that everyone in the world is connected.
	8	I am willing to spend time to support general Facebook community activities.
Meeting new people	9	Interacting with people in my Facebook network gives me new people to talk to.
	10	Through my Facebook network, I come in contact with new people all the time.
Bonding social capital		
Individual benefit	1	There are several people in my Facebook network I trust to help solve my problems.
	2	There is someone in my Facebook network I can turn to for advice about making very important decisions.
	3	There is no one in my Facebook network that I feel comfortable talking to about intimate personal problems.
	4	When I feel lonely, there are several people in my Facebook network I can talk to.
	5	If I needed an emergency loan of $500, I know someone in my Facebook network I can turn to.
	7	The people I interact with in my Facebook network would be good job references for me.
Collective action (More sacrifice)	6	The people I interact with in my Facebook network would put their reputation on the line for me.
	8	The people I interact with in my Facebook network would share their last dollar with me.
	9	I do not know people in my Facebook network well enough to get them to do anything important.
	10	The people I interact with in my Facebook network would help me fight an injustice.



Model 1						
Variables	1	2	3	4	5	6
1 Actual friends	-					
2 F of asking help	.375**	-				
3 SRI	.387**	.250**	-			
4 N of strategies	.004	-.101	.204*	-		
5 Bridging SC	.213*	.334**	.504**	.138	-	
6 Bonding SC	.238*	.195*	.259**	.128	.213*	-

Model 2									
Variables	1	2	3	4	5	6	7	8	9
1 Actual friends	-								
2 F of asking help	.375**	-							
3 SRI	.387**	.250**	-						
4 N of strategies	.004	-.101	.204*	-					
5 Outward looking	.129	.254**	.334**	.185	-				
6 Broader group	.214*	.278**	.443**	.168	.399**	-			
7 New people	.127	.218*	.351**	-.067	.382**	.284**	-		
8 Individual benefit	.164	.108	.184	.097	.147	.076	.046	-	
9 Collective action	.241*	.195*	.307**	.116	.196*	.350**	.052	.479**	-

Table 4. Correlation matrices

Ellison et al (2007)

- Relationship between the use of Facebook, and the formation and maintenance of social capital at Michigan State University

- Hypotheses
  - H1: Intensity of Facebook use will be positively associated with individual's perceived bridging social capital
  - H2: Intensity of Facebook use will be positively associated with individual's perceived bonding social capital

Measures

- Bridging social capital
- Bonding social capital
- Maintained social capital
- self-esteem
- Facebook intensity
- Life Satisfaction at MSU
- All questionnaire based- Scores calculated by taking the mean of the item scores on that scale

Table 2 Summary statistics for Facebook intensity

Individual Items and Scale	Mean	S.D.
Facebook Intensity[1] (Cronbach's alpha = 0.83)	**−0.08**	**0.79**
About how many total Facebook friends do you have at MSU or elsewhere? 0 = 10 or less, 1 = 11–50, 2 = 51–100, 3 = 101–150, 4 = 151–200, 5 = 201–250, 6 = 251–300, 7 = 301–400, 8 = more than 400	4.39	2.12
In the past week, on average, approximately how many minutes per day have you spent on Facebook? 0 = less than 10, 1 = 10–30, 2 = 31–60, 3 = 1–2 hours, 4 = 2–3 hours, 5 = more than 3 hours	1.07	1.16
Facebook is part of my everyday activity	3.12	1.26
I am proud to tell people I'm on Facebook	3.24	0.89
Facebook has become part of my daily routine	2.96	1.32
I feel out of touch when I haven't logged onto Facebook for a while	2.29	1.20
I feel I am part of the Facebook community	3.30	1.01
I would be sorry if Facebook shut down	3.45	1.14

Notes: [1]Individual items were first standardized before taking an average to create scale due to differing item scale ranges. [2]Unless provided, response categories ranged from 1 = strongly disagree to 5 = strongly agree.

Ellison et al., (2007)

- Findings
  - Facebook intensity strong predictor of three types of social capital measured
  - Strongest prediction was with bridging social capital

Limitation of Ellison et al. (2007)

- Focuses on social capital between MSU students
  - What about someone's wider social network?
- Conducted in 2006- Perhaps influenced by fashion?
- Causal direction impossible to establish
  - Does high social capital cause, or is caused by SNS use?

Our study

- Facebook intensity, Bonding & Bridging social capital
- Using measures from Jung instead of Ellison (2007)
  - Focus on Facebook network rather than local group
- Correlation study
- Based on what we know from the papers and social capital, what are our hypotheses?
- How would we analyse this with the data we have?

Today's practical

- Clean your dataset
- Assess the demographics- mean and SD of age, gender split of sample.
- Analyse questionnaire data
  - Create scale scores
  - Descriptives/graphs
  - Normality test of data
  - Analysis of the relationship between the variables (see last week if stuck for analysis ideas)

Hints for Practical

- Id10 does not use Facebook. Should you remove this row?

```
data[!data$id == "10", ]
```

- There are 6 questionnaires in the dataset. We are only interested in 3 of them.
- Reverse scoring items
  - bonding_3 and bonding_9 are not the same polarity as the other items. We need to reverse those scores e.g. make a score of 5 into 1; 4 into 2 etc..

    You can do this with a simple calculation
- Making scores for the questionnaire scales can be done by creating a mean of the scores in the relevant columns for each participant

```
rowMeans (data[row, column])
```

Regression

Evaluation Methods & Statistics Lecture 6

What will be covered

- Hypothesis Testing
- Types of data
- P value revisited
- Regression
  - What it does
  - How it does it

Our Research Question

- People may be anxious about posting tweets due to judgments from others
- This anxiety might explain why some twitter users do not post frequently
- Does posting anxiety predict the number of twitter posts?

The Scientific Process

1. Generate a hypothesis
2. Design experiment/study to test the hypothesis
3. Collect data from sample
4. Fit statistical model to the data
5. Assess how well this model represents the data (the model fit)

Hypothesis testing

The Null Hypothesis (H0)

Posting anxiety **does not** predict twitter posting

The Research Hypothesis(H1)

Posting anxiety **significantly predicts** twitter posting

Hypothesis testing

- H0 given more weight
  - Experiment run to disprove H0 => will not reject it unless evidence is sufficiently strong
  - Discount the simple before adopting something more complex (Occam's razor)

Types of Data- Categorical

- Nominal
  - Data ascribing objects or values to distinct categories
    - Eg. High, Low
- Ordinal
  - Nominal data with explicit order/ranks
    - 1st, 2nd, 3rd
- Although we know category and order we don't know the quantity of difference between values

Types of Data- Continuous

- Interval
  - Equal intervals in the scale
  - E.g. 5 point Likert Scale

- Ratio
  - Equal intervals with a true 0 point
    - E.g. Reaction Time
    - Distance along scale should be divisible
    - X on scale should equal 2x/2

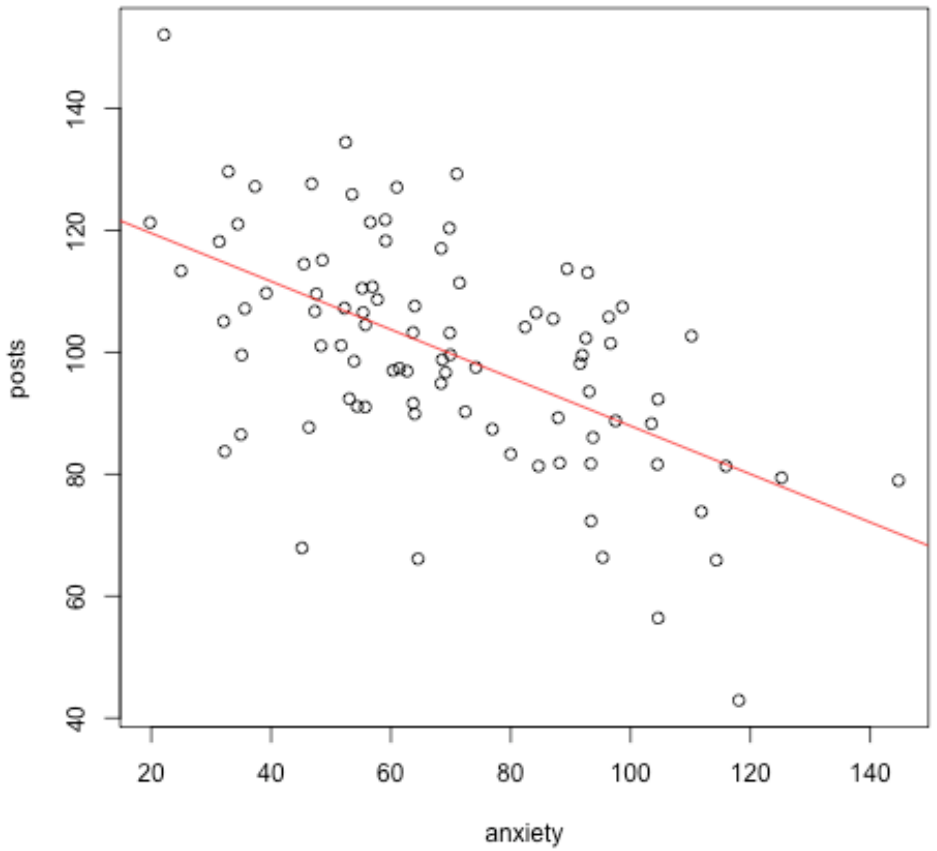
What is Regression?

- Significance of a predictor variable **(posting anxiety)** on outcome variable **(twitter posts)**
- Predictor can be continuous or categorical
- Allows us to predict future posts based on knowledge of predictor value
- **Here we are looking at linear regression (straight line model)**

Equation of Straight Line

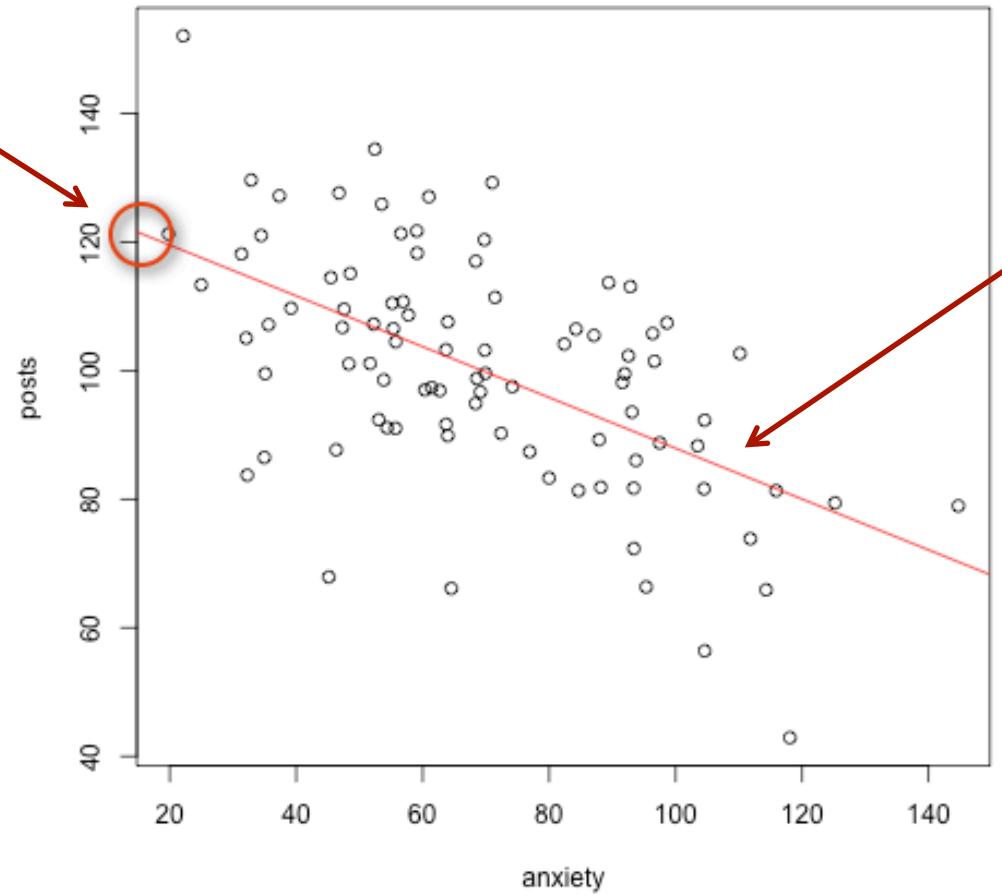
- y= mx +c --- Familiar?
- $Y_i = \beta_1 X_i + \beta_0 + \varepsilon_i$
  - Y_i= Outcome Variable value
  - β_0= intercept
  - β_1= gradient of regression line
  - X_i= Participant $_i$ score on our predictor variable
  - ε_i =residual difference between score predicted for Y_i and the one actually attained in the data

Our Data



Our Data

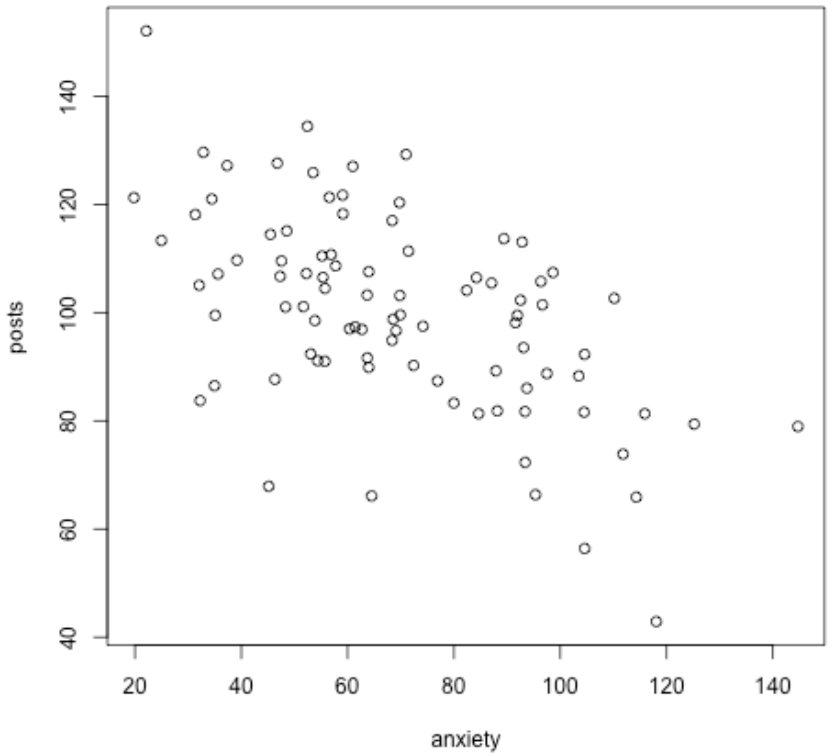
Intercept
β_0



Gradient of the line- β_1

Change in *y* with a unit change in x

How do we fit this model?

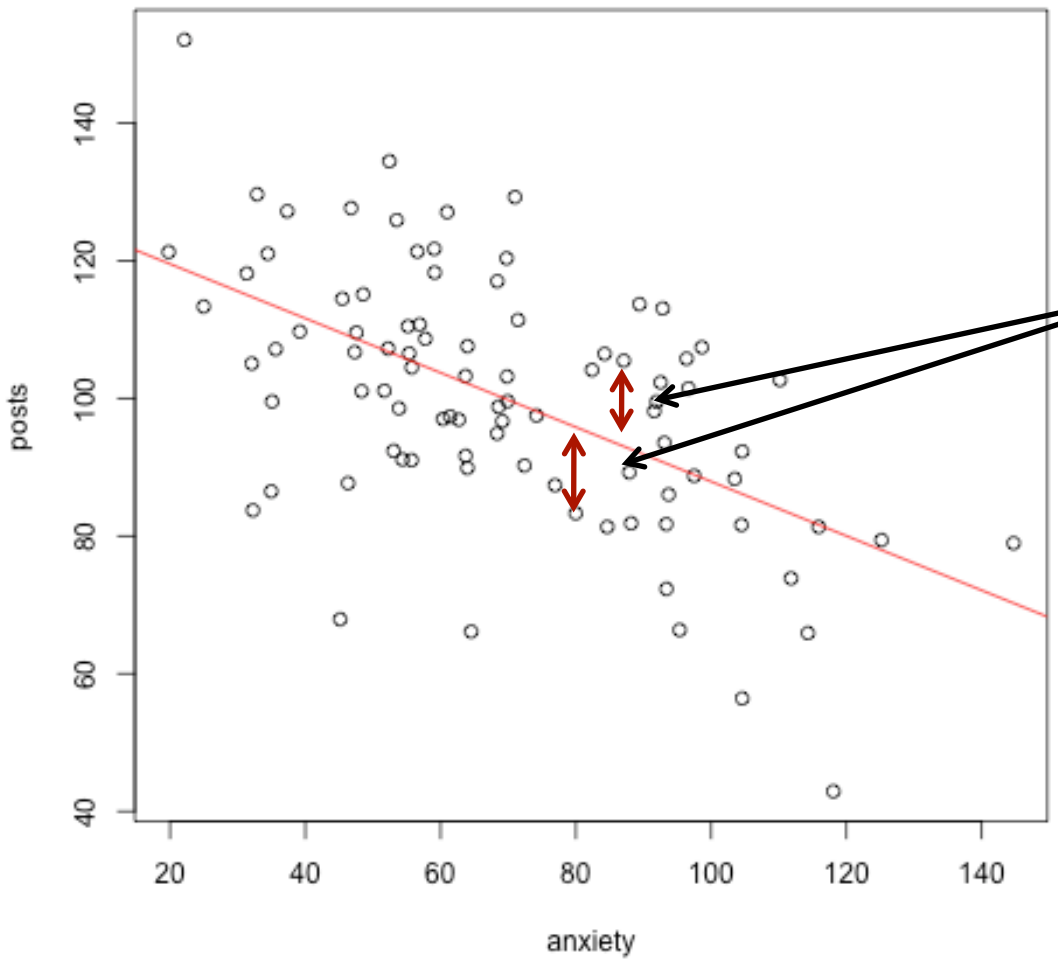


- Data looks linear
- We could just draw a line we feel is the *best fit*
- Very subjective
- Wouldn't be sure it was the best fit
- Use technique called **method of least squares**

Method of Least Squares

- Residual difference between the line and the actual data point
- The line of best fit (regression line)= line that leads to minimum residual
- The residuals are squared and summed (SS) as some will be negative some positive.

Our Data

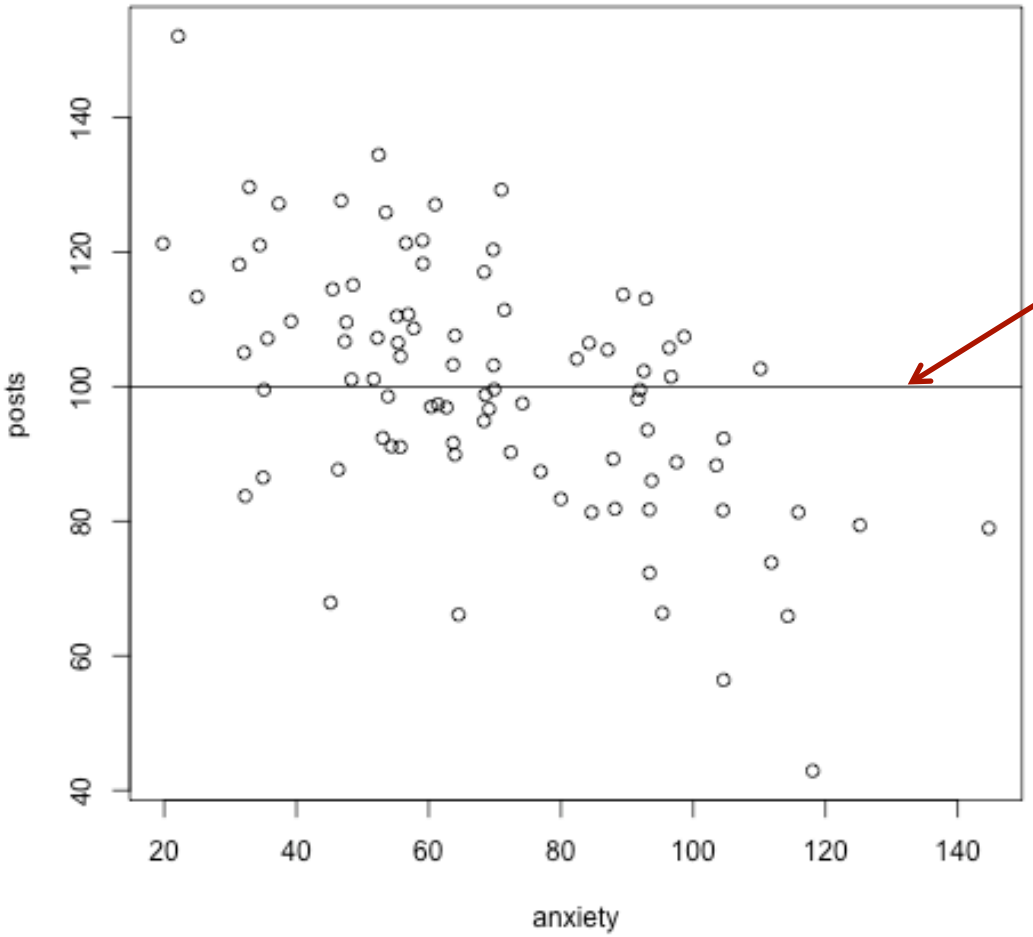


Residuals

Is it the best model?

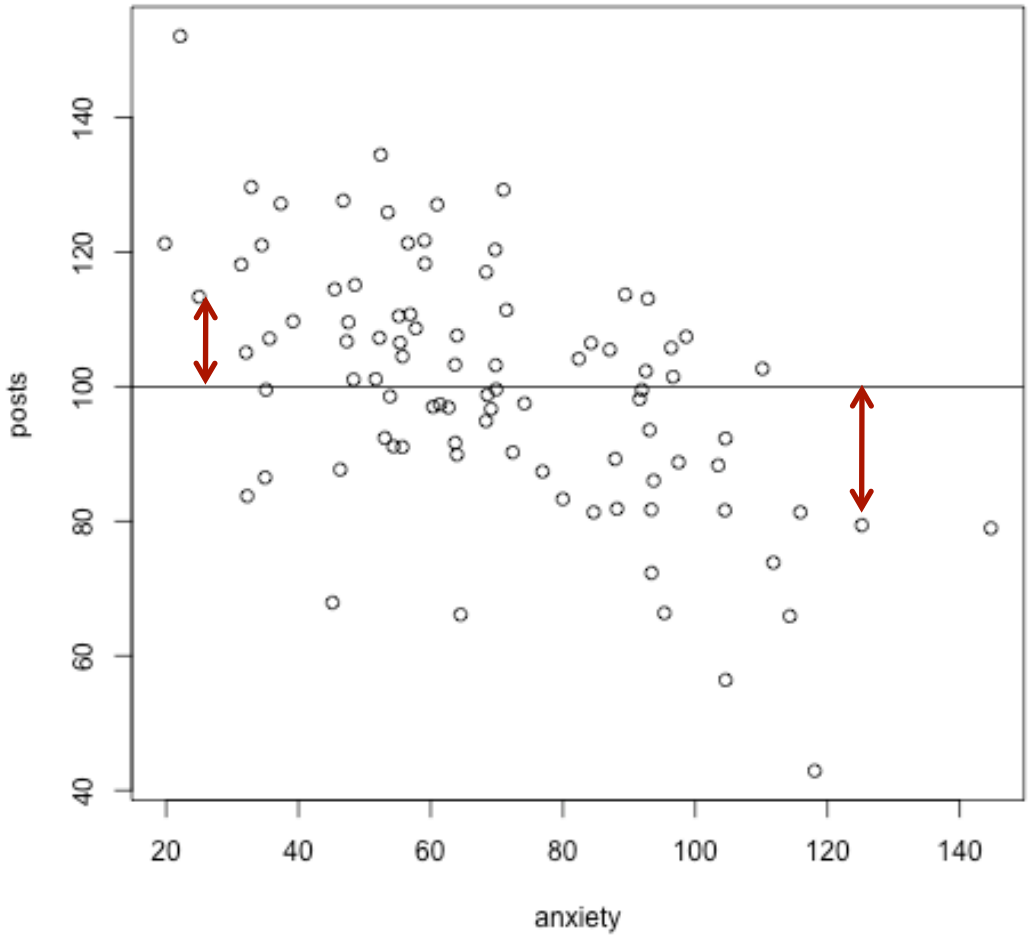
- Although it may be the line of best fit available, it might still be a poor description of the data
- We need to find out how adequate the best fit model is
- We therefore compare the best fit model to the most basic model (the mean) using Sum of Squares

Our Data



Mean of y-
Simplest Model

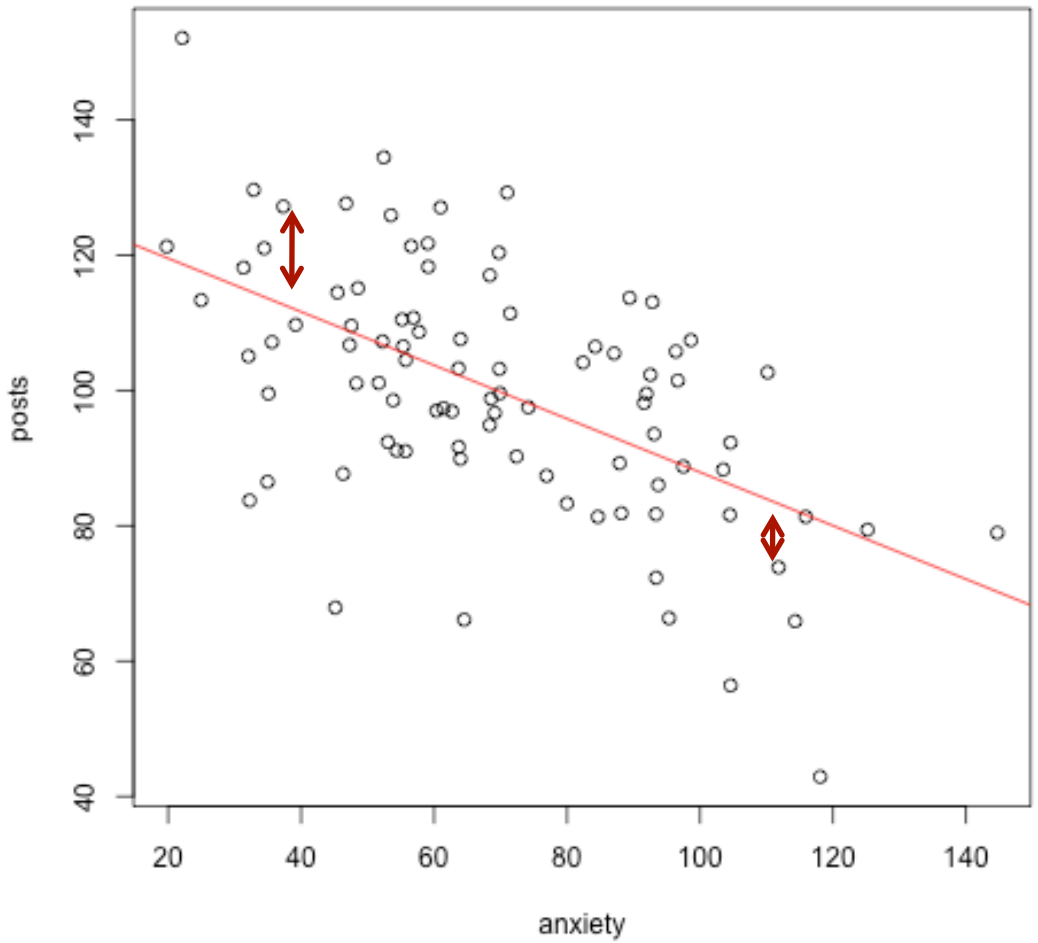
Total Sum of Squares (SST)



SST= Σ (Observed – Mean Y)^2

Total amount of differences present when simplest model applied

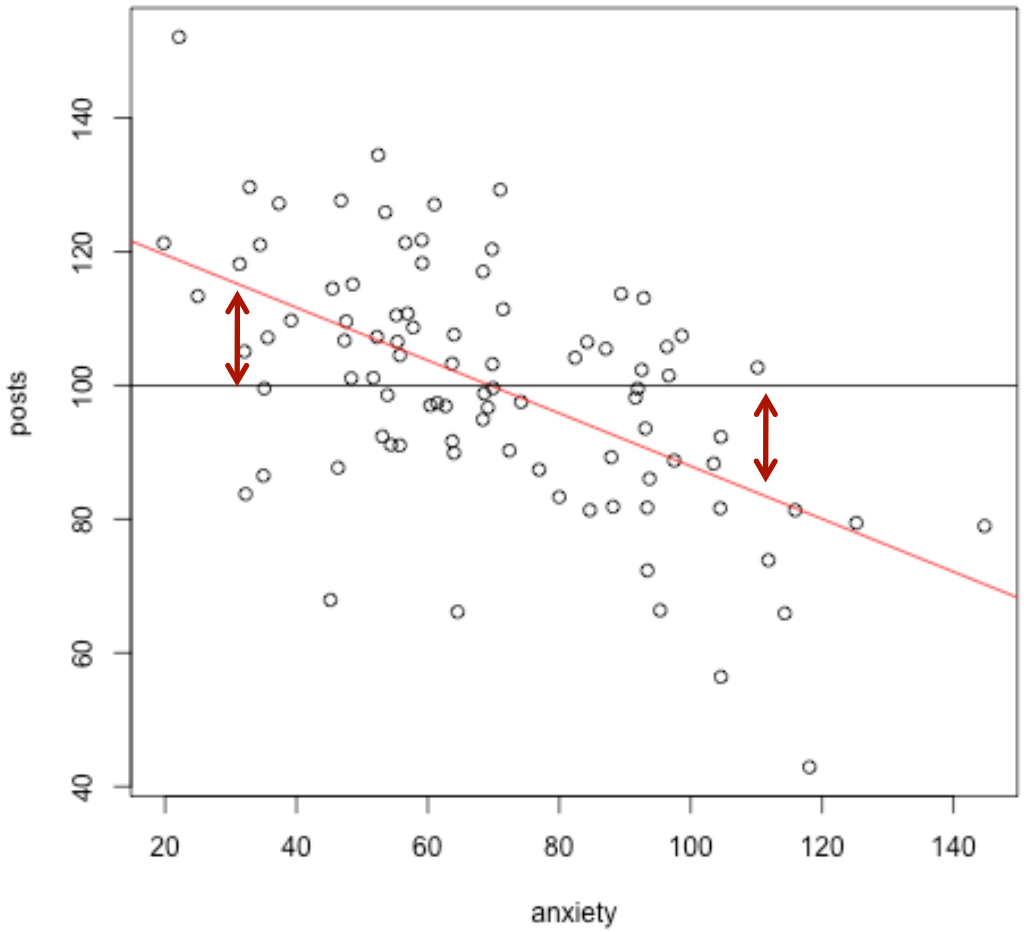
Residual Sum of Squares (SSR)



SSR= Σ (Observed – Model)^2

Total amount of differences present with best fit model applied

Model Sum of Squares (SSM)



SSM= Σ (Model – Mean Y)^2

Total amount of differences between the predicted Y from best fit and the mean of Y

Alternatively: SSM= SST- SSR

How good is the model?

- If SSM is large then suggests model has made big improvement over just using the mean
- If SSM is small then model has made little improvement over the mean
- We can get a number to show us how much improvement (R^2)
- R^2= SSM/SST
- How much variation explained by the model as a proportion of how much there was in the first place
- R^2x100 = % variation explained by model

How good is this model?

- Can also assess this through F Ratio Test
- F = Improvement due to model (MS_M)/difference between model and observed data (MS_R)
- We want F to be large (MS_M to be large and MS_R to be small)

Is the predictor significant?

- Does anxiety significantly predict posts?
- Mean model assumes β_1 to be 0
  - No increase in y with unit increase in x
- We want to see whether the β_1 for regression model is significantly larger than 0
- This is done by running a t-test on these values (more next week)
- P value is the probability that the obtained t value would occur if β_1 was actually 0

The Philosophy of Statistics

test statistic= s^2 explained by model / s^2 unexplained by model

we know how frequently certain test statistic values occur

We can therefore calculate the probability of obtaining that value by chance (the p value)

The Philosophy of Statistics

- As the test stat gets larger there is less likelihood that the test stat is due to chance

- When this likelihood falls below 0.05 (Fisher's Criterion) we say that our findings are statistically significant

  - In other words we accept the H1 and reject H0

What is a P Value?

- The probability that the results obtained occurred by chance assuming there is no effect at all
- As p value gets lower then more certain we can reject our null hypothesis
- p<= 0.05, p <= 0.01, p<= 0.001

How do we get a p value?

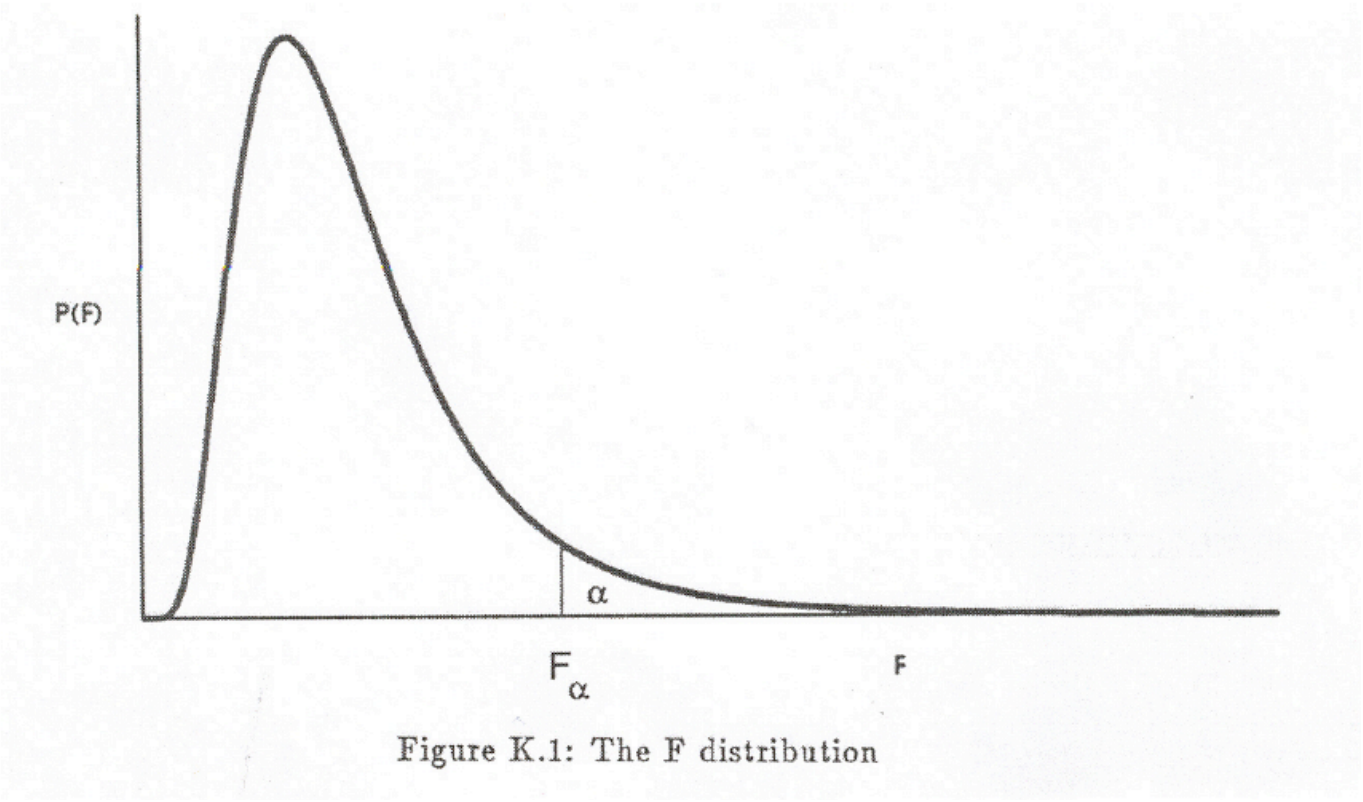


Figure K.1: The F distribution

Reporting Statistics

- American Psychological Association Style Guide
- E.g. Correlation reporting:
  - **r(82)= -.79, p<0.001**

Degrees of freedom (n-2)

Pearson's correlation coefficient

P value

- Has instructions for each test
- Guide is available online

Reporting Statistics

- Reporting regression
  - Report regression table

  - In text reporting of predictor findings and regression analysis results:
    - $b = -.34$, $t(225) = 6.53$, $p < .001$
    - $R^2 = .12$, $F(1, 225) = 42.64$, $p < .001$

Readings

- Field, A., Miles, J., & Field, Z. (2012). Discovering Statistics Using R. Chapter 7

T-test

Evaluation Methods & Statistics – Lecture 7

Benjamin Cowan & Andrew Howes

What we will cover

- Experiment Design
- Independent, Dependent Variables
- 2 samples t-test
  - Independent
  - Dependent
- Reporting the findings

A Study…..

Research Example

- Consequences of a secondary task on driving
- Does using a mobile phone to text cause driving quality to deteriorate?

Independent and dependent variables

- We systematically manipulate the IV

- We want to see how this systematic manipulation impacts an outcome (or DV)

Research Set-Up

Independent Variable (IV)

- **Secondary Driving Task**

This independent variable has **2 levels** or **factors**

1) Control Group

- Just driving (no secondary task)

2) Texting Group

- Participants asked to text whilst driving

Dependent Variable (DV)

- **Driving Quality Score**

Main task for the participants is to drive for an hour in a simulator

Hypotheses

Null Hypothesis

- There will be no significant difference between **secondary driving tasks** on **driving quality score**

Research Hypothesis

- There will be a significant difference between **secondary driving tasks** on **driving quality score**

Experiment Design

Methods of data collection

Between-subjects

Different groups of people take part in each driving task

Within-Subjects

Same group of people take part in each driving task

IV Level 1 Participants

IV Level 2 Participants

IV Level 1 Participants

IV Level 2 Participants

Methods of data collection

- This influences what statistical test you use

- It also influences how you design your experiment
  - Counterbalancing (In Within Subjects)
  - Sample matching (in Between Subjects)
  - To minimise confound effects on the variance

Two types of DV Variation

Unsystematic

Differences in DV due to unknown (or unmeasured) factors

e.g. personality, change in room environment between conditions 1 and 2

Systematic

Differences in DV due to the change in IV

e.g. Changes in driving score due to systematic manipulation of IV

Between-Subjects

Unsystematic variation due to sample differences

- naturally vary in IQ, Personality, attention span

We can try to control for this

- Include high risk variables as part of design
- E.g. high and low attention span as another IV

Randomly allocate participants to condition

- Spreads the variation across conditions

Within Subjects

- Unsystematic variation controlled
  - More power to identify effects
- Systematic confounds
  - Practice effects
  - Boredom effects
- Form a risk to our conclusions if we assume that systematic variation is all due to IV
- Counterbalance to reduce impact of this confound
  - 50% of participants- Condition 1 before 2
  - 50% of participants- Condition 2 before 1

T-Test

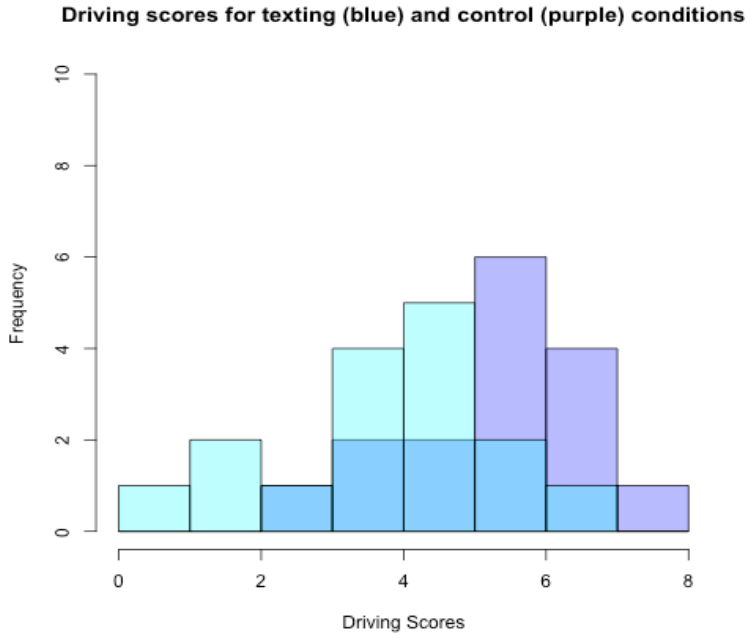
Guinness & t-test

- William Sealy Gosset (1876-1937)
- Worked for Guinness as a statistician
- They needed stats to examine which types of barley had the best yield
- In 1908 published a paper which introduced the t-distribution under the pseudonym "Student"

T-Test

- T-test is looking at **differences**
- It is testing to see if the two sample means gathered are from different populations
- Why might they be from different populations?
  - Due to the levels of our IV

T-test



T-Test: Rationale

If the sample mean difference is larger than we expect

- We have collected two sample by chance that are atypical of the population

OR

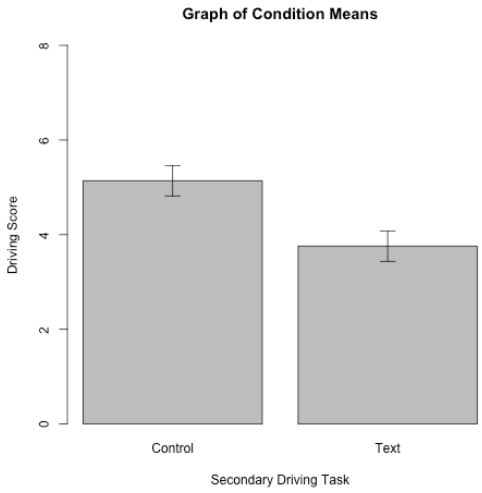
- The two samples are from different populations

T-Test: Rationale

If the sample comes from the same population (i.e. as assumed in H_0)

- we'd expect these means to be roughly equal
- **Large differences** between sample means should occur rarely

If they do, it is because sample means are from different populations



Types of t-test

- **Independent means t-test**
  - Different participants are allocated to each of the secondary driving task conditions
  - AKA independent measures & independent samples

- **Dependent means t-test**
  - Each participant completes both of the secondary driving task conditions
  - AKA matched pairs & paired samples

T-Test: The Formula

- Keep in mind with t-test we are interested in **differences**

The Dependent t-test: Calculation

Mean difference between samples (Experiment effect)

Difference expected between population means (because of H0 then this is 0)

test statistic for the t-test

$$t = \frac{\overline{D} - \mu_D}{S_D / \sqrt{N}}$$

Standard error of the differences

Dependent t-test calculation

$$\overline{D}$$

score_control	score_text	D
3.85	5	-1.15
5.29	4.96	0.33
5.52	4.09	1.43
5.07	3.77	1.3
5.11	5.24	-0.13
4.04	4.6	-0.56
5.34	5.36	-0.02
6.66	6.41	0.25

- The mean of the D values

Standard Error of Differences

$$S_D / \sqrt{N}$$

Used as variability gauge between sample means

- If this is small -- most samples should have similar means, thus we would expect a small D
- If this is large – large D is likely

Independent t-test

The same premise as the dependent but some important differences

Instead of differences between pairs of scores it looks at differences between **overall means**

Because we don't have pairs of scores for each participant

Independent t-test

This means we cannot calculate the SE of differences by using the **differences in the sample**

- There are 2 independent samples

Calculated using the **variance sum law**

- The **variance of a difference** between two independent variables is equal to the sum of their variances
- Then square root this to get SE

Effectively this is doing the same thing as the denominator in previous equation

Independent t-test equation- with equal sample sizes

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)}}$$

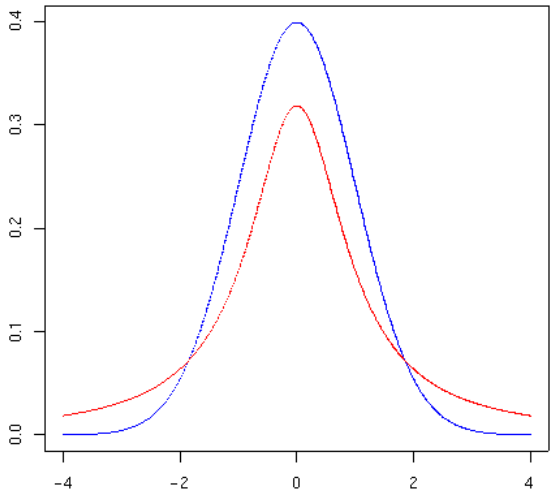
T-test: R output

```
> t.test (control, text, type=Student)

        Welch Two Sample t-test

data:  control and text
t = 2.673, df = 28.564, p-value = 0.01229
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.3236936 2.4388064
sample estimates:
mean of x mean of y
 5.138125  3.756875
```

T Distribution



Reporting the t-test

- Need to report
  - The t statistic
  - The degrees of freedom (N-2)
  - The significance value (or full p statistic)

e.g. t (dof) = t statistic, pvalue

The Concept of Degrees of Freedom

- The number of observations that are free to vary
- Imagine a rugby team:
  - 15 people needed
  - If you arrive last then you have no choice in where to play
  - But the other 14 did
  - There are therefore 14 degrees of freedom (15-1)
  - In our t-test our degrees of freedom are the number of differences that are free to vary

The Concept of Degrees of Freedom

Degrees of freedom

- Dependent t-test
  - N-1
  - Because we are using 1 variable (mean of differences)
- Independent t-test
  - N-2
  - Because we are using 2 variables in the calculation

Parametric assumptions of t-test

Both tests

- Data is normally distributed (Shapiro Wilk test)
- Data is interval or ratio scale

Independent t-test only

- Variance in populations are roughly equal (Equality of variance)- Levene's test
- Scores are independent (i.e. They come from different people).

Shapiro-Wilk (Normality)

- Compares sample distribution to normal distribution
- If our sample varies significantly from this distribution, what would we expect?

```
> shapiro.test (control)

        Shapiro-Wilk normality test

data:  control
W = 0.9584, p-value = 0.6332
```

Shapiro-Wilk (Normality)

- Compares sample distribution to normal distribution
- If our sample varies significantly from this distribution, what would we expect?
- Test will be statistically significant
- We want it to be **non significant** (p>.05)

```
> shapiro.test (control)

        Shapiro-Wilk normality test

data:  control
W = 0.9584, p-value = 0.6332
```

Shapiro-Wilk

Levene's Test (Homogeinity of Variance)

- Compares variance in each sample to see if they are roughly equal
- We want both variances to be similar
- Do we want the test to be statistically significant or not?

```
> leveneTest(data$score, data$condition)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  1  0.9755 0.3312
      30
>
```

Levene's Test (Homogeinity of Variance)

- Compares variance in each sample to see if they are roughly equal
- We want both variances to be similar
- Do we want the test to be statistically significant or not?
- We want it to be **non significant** (p>.05)

```
> leveneTest(data$score, data$condition)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  1  0.9755 0.3312
      30
>
```

Levene's Test

Commands used in practical

Data Manipulation

- subset()

Graphing Data:

- barplot()
- boxplot()
- se.bar()
  - This is our own "homemade" function

Descriptives & Assumptions

- mysummary()
  - Again, a homemade function
- shapiro.test()
- leveneTest()

Analysis

- t.test()

Task for next week

- Complete the dependent and independent t-test manual calculation on Canvas
- Do this using a calculator (no R!)
- Solution will be posted online

Readings for this lecture

- Field, Miles & Field (2012) Chapter 9
- Howell (2010) Chapter 7 p.194-213

Comparing 3 Conditions- ANOVA

Evaluation Methods & Statistics- Lecture 9

Benjamin Cowan & Andrew Howes

Research Example (Lecture 7)

- Consequences of a secondary task on driving

- Does using a mobile phone to text cause driving quality to deteriorate?

Research Example (This Week)

- Consequences of a secondary task on driving
  - Texting
  - Talking on phone
- Compared to just driving (control)

How would we design this experiment?

How would we design this experiment?

- IV- Secondary Driving Task
  - Level 1- Control Group (No secondary task)
  - Level 2- Texting
  - Level 3- Talking

- DV-Driving score

How would we analyse the data?

- We could do 3 t-tests
  - Control to Texting
  - Control to Talking
  - Talking to Texting
- This would inflate our *Type I error rate*

Type I error

- When we believe there is genuine effect in our population…..but actually there isn't (a false positive)

Type II error

- When we believe there is **no** effect in the population…….but there is

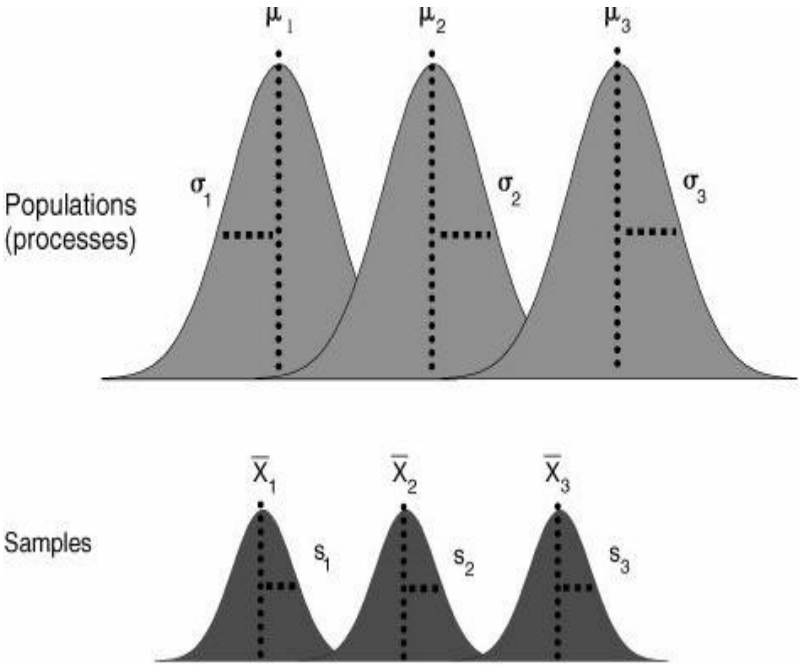
- Lot of natural variation between samples, too stringent controls for Type I error, low power of stats to find effects

Familywise error rate

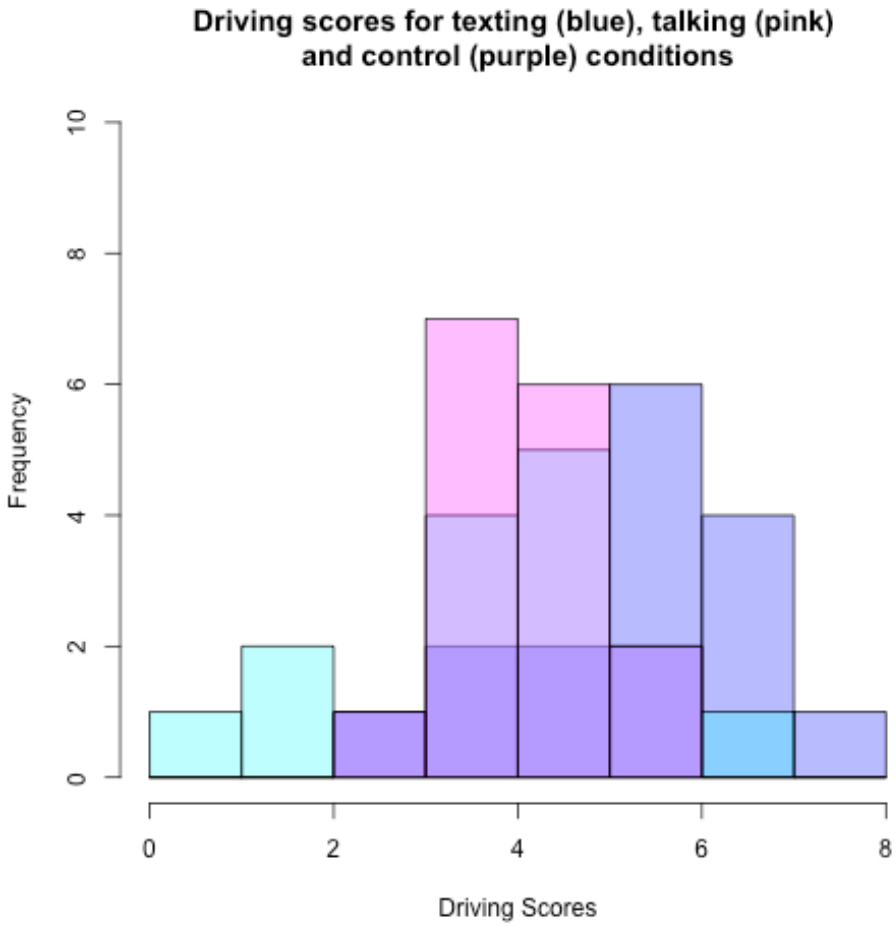
- If we have 3 tests in a *family of tests* and assume each is independent
- If we use Fishers level of 0.05 as our level of significance…
- The probability of a false positive (Type 1 error) in all of these tests
  - 0.95 x 0.95 X 0.95 = 0.857
  - => Probability of Type 1 error is 1- 0.857=0.143
  - That is far greater than the Type I error for each test separately (0.05)
  - We therefore use ANOVA rather than lots of t-tests

ANOVA- The Idea

- Compare 3 (or more) means to identify whether they are significantly different
  - i.e. whether they come from different populations
- Or more accurately…..we are testing the null hypothesis that the samples come from the same population.
- It is what we call an *omnibus test*
  - It tells us there is a significant difference, not where it is.



Our Data

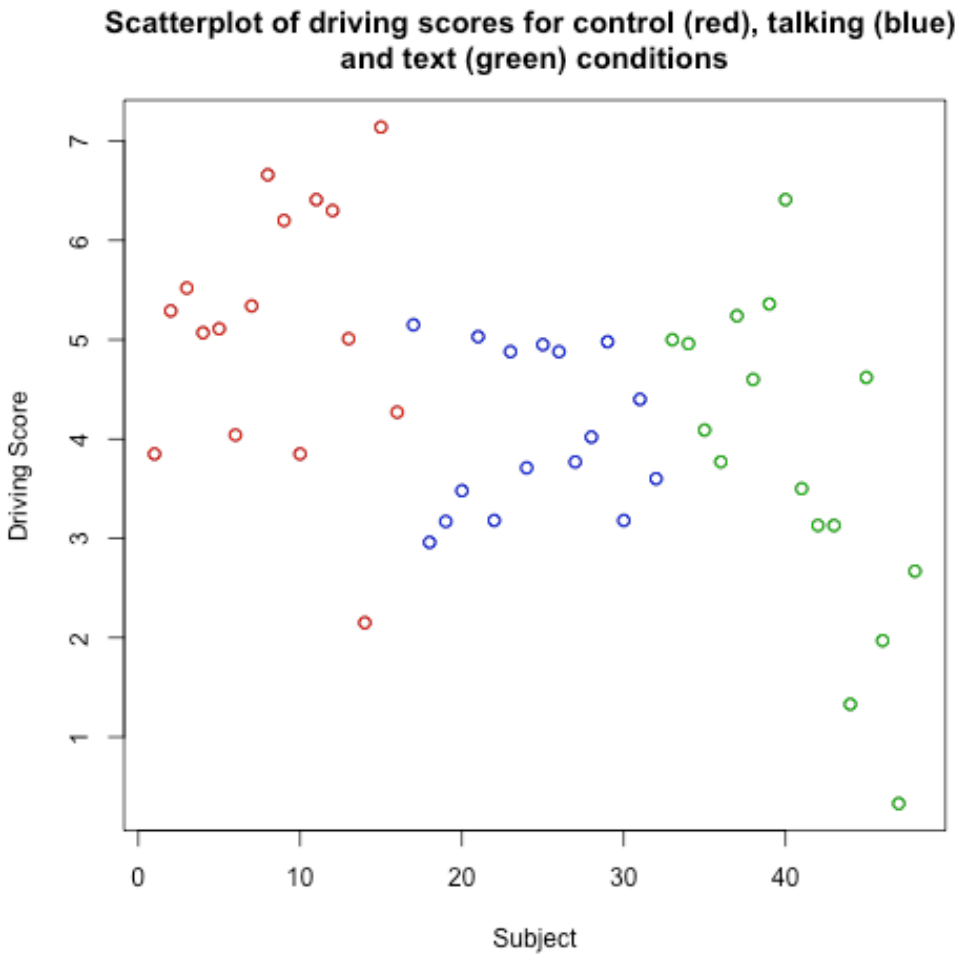

Driving scores for texting (blue), talking (pink)
and control (purple) conditions
Frequency
0
2
4
6
8
10
0
2
4
6
8
Driving Scores

The Key: ANOVA & F Ratio

- F ratio is the ratio of **explained (that accounted for by the model we are proposing)** to **unexplained** variation

- This is calculated using the **Mean Squares**

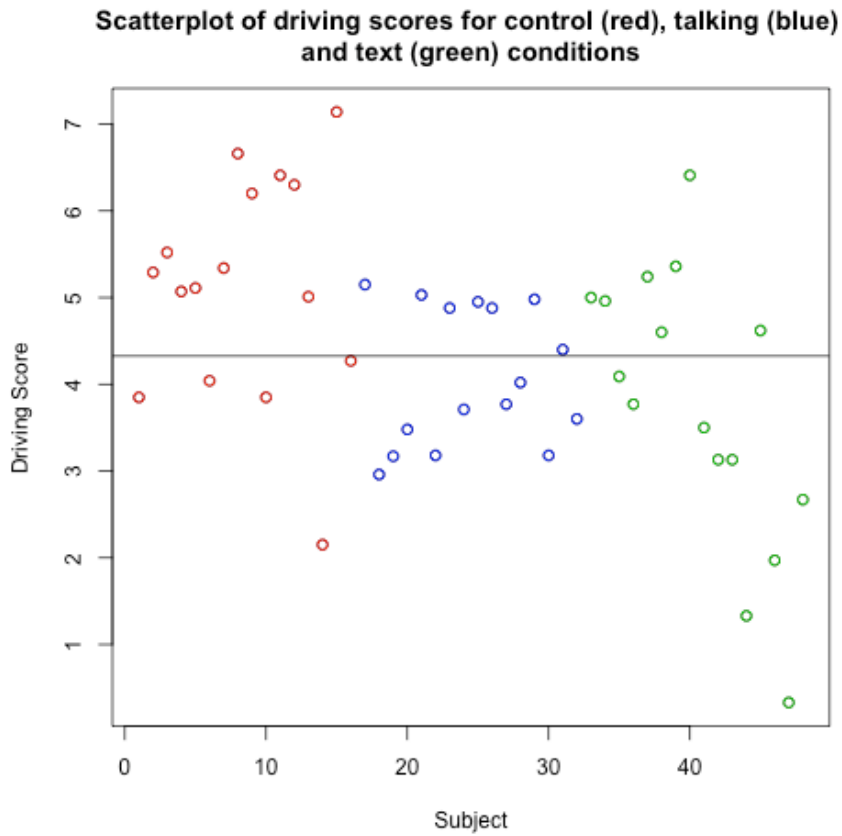
Our Data


Scatterplot of driving scores for control (red), talking (blue) and text (green) conditions
Driving Score
1
2
3
4
5
6
7
0
10
20
30
40
Subject

The mindset of "models"

the mean is a statistical model, just sometimes not a very good one…….

Does the statistical model we have proposed explains the variation better?



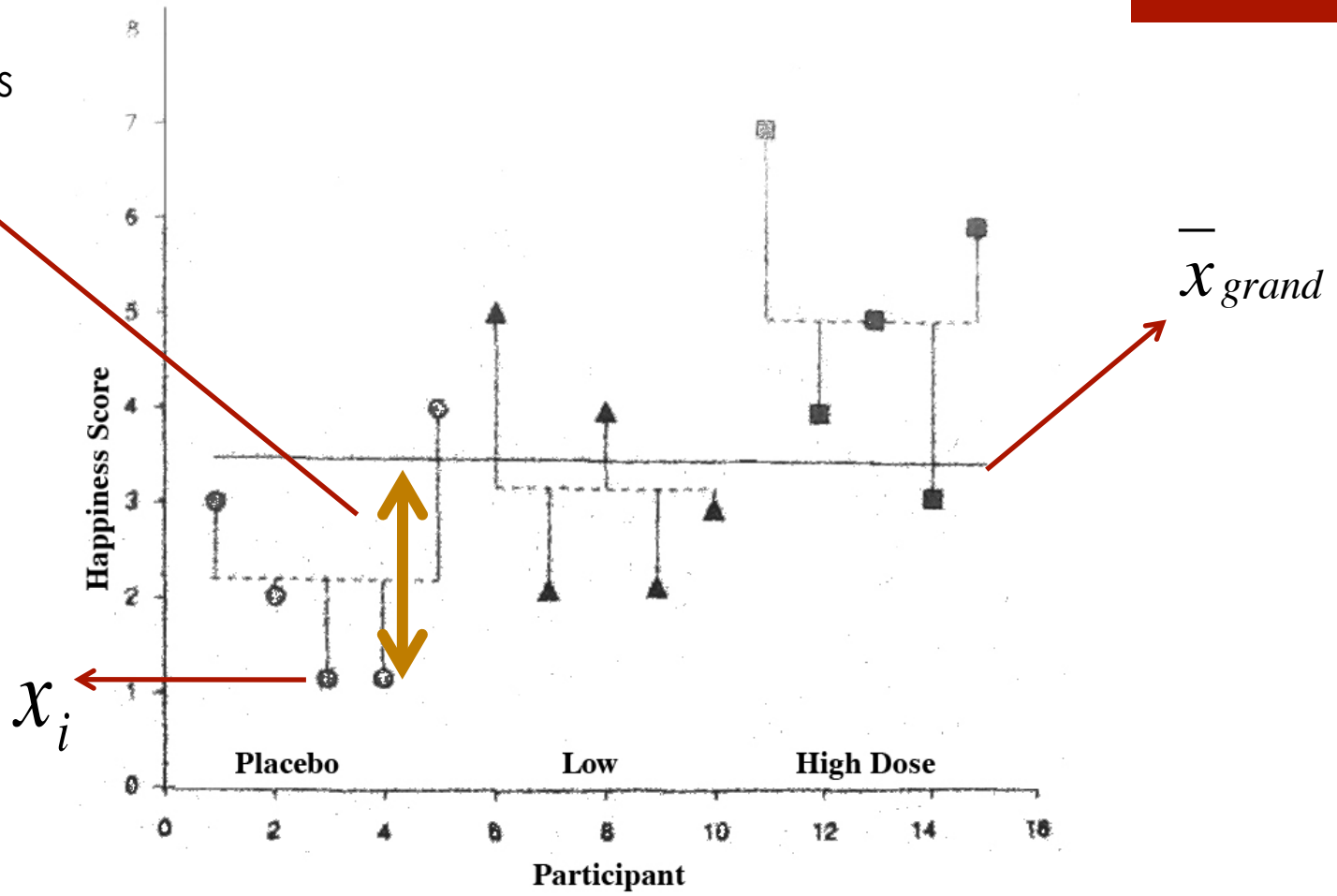
Step 1- Total Sum of Squares

- The total amount of variation in our data
- This should look familiar

$$SS_T = \sum \left(x_i - \overline{x}_{grand} \right)^2$$

Step 1- Graphically

What the equation is doing



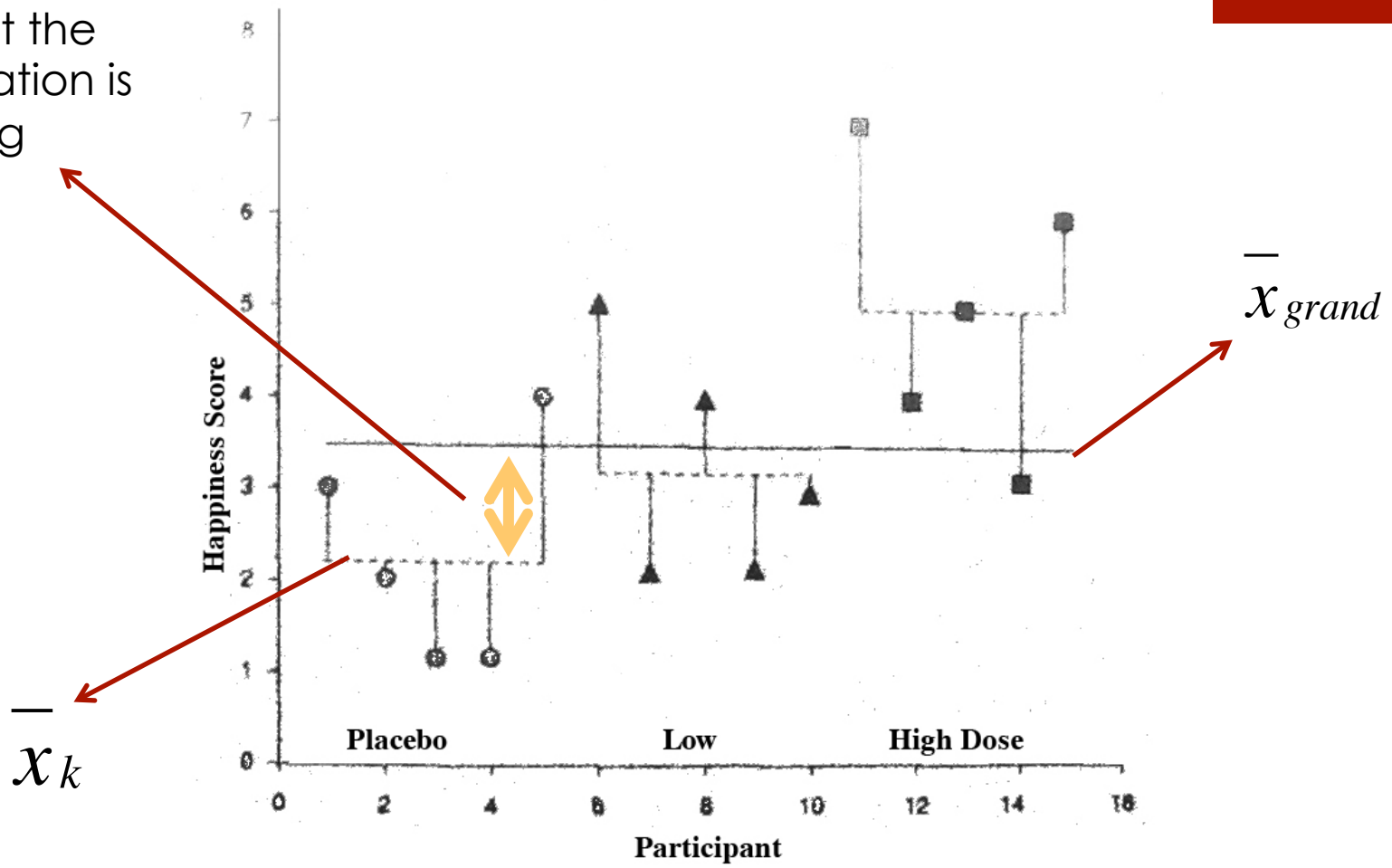
Step 2- Model Sum of Squares

- We now need to know how much variation our model can explain
- How much the total variation can be explained due to data points coming from different groups in "the perfect model"
- n_k is the amount of people in that condition

$$SS_M = \sum n_k (\overline{x}_k - \overline{x}_{grand})^2$$

Step 2- Graphically

What the equation is doing



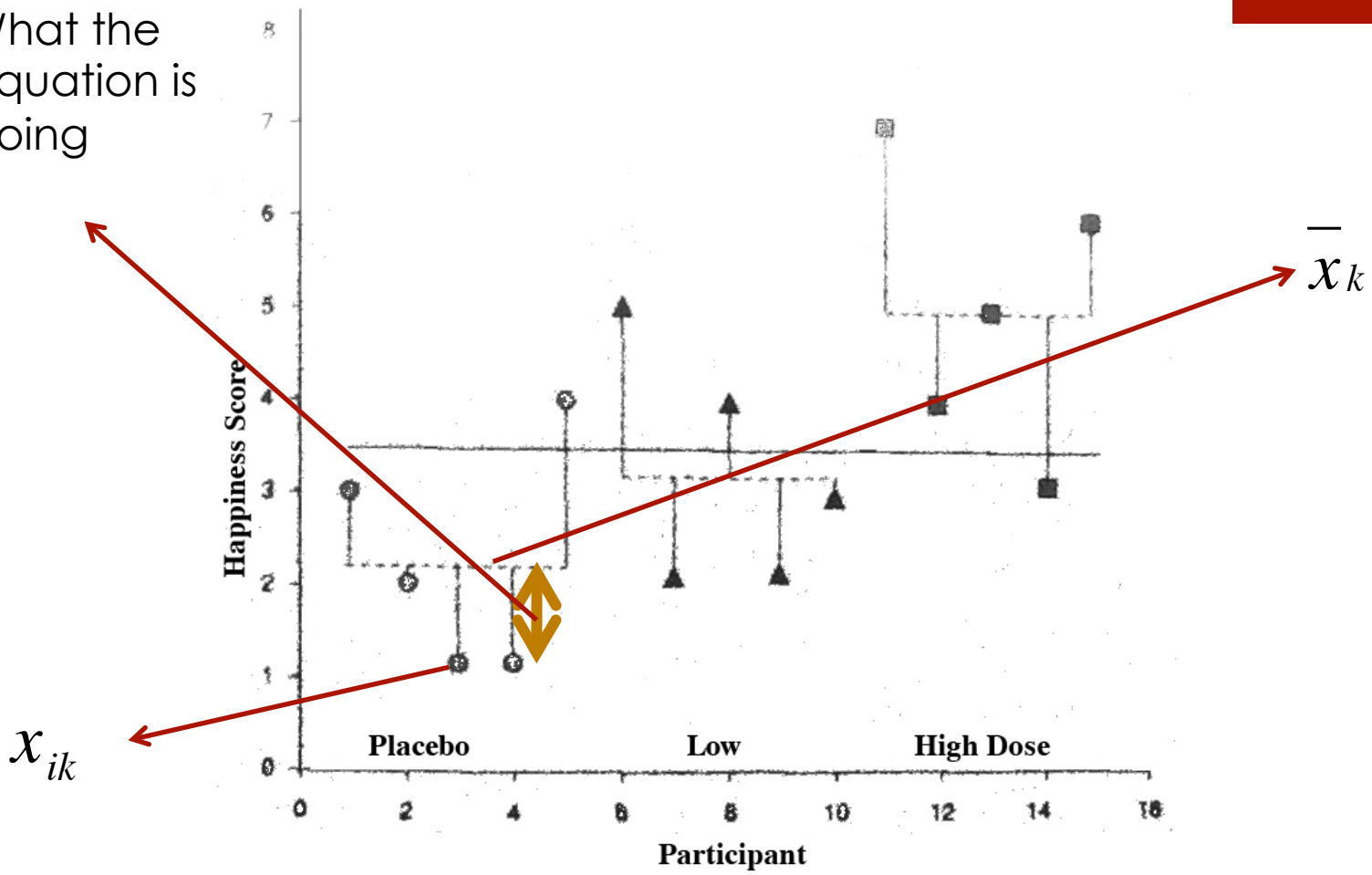
Step 3- Residual Sum of Squares

- How much of the variation cannot be explained by the model i.e. what error is there in the model prediction?
- Easy way to calculate: $SS_R = SS_T - SS_M$
- But here is the real formula

$$SS_R = \sum \left(x_{ik} - \overline{x}_k \right)^2$$

Step 3- Graphically

What the equation is doing



Sum of Squares & Mean Squares

- These are summed values
  - Therefore impacted by the number of scores in the sum (remember variance in Lecture 3)

- We can get around this by dividing by the respective degrees of freedom for each SS

Degrees of Freedom for each SS

- Degrees of Freedom for SS_T (dfT):
  - N-1
- Degrees of Freedom for SS_M(dfM):
  - Number of Conditions (k) -1
- Degrees of Freedom for SS_R(dfR):
  - N-k

F Ratio

- The F Ratio is calculated using the:

- Mean Squares model (MS_M):
  - SS_M/df_M

- Mean Squares residual (error) (MS_R):
  - SS_R/df_R

F Ratio

$$\frac{\text{Variation explained by our model}}{\text{Variation unexplained by our model}}$$

F Ratio

$$\frac{\text{Mean Square Model } (MS_M)}{\text{Mean Square Residual } (MS_R)}$$

F Distribution

- F Distribution for specific pair of degrees of freedom

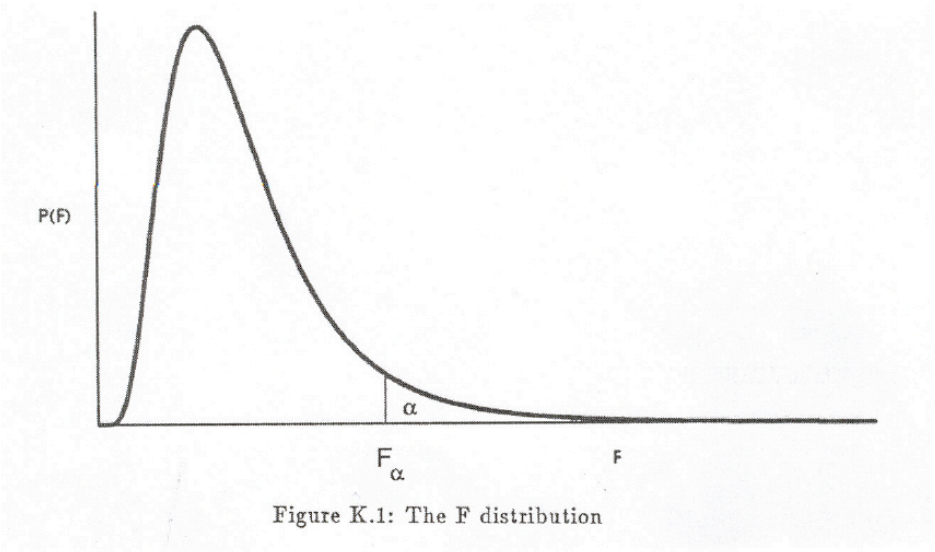


Figure K.1: The F distribution

- Table of Critical Values

Critical values of F for the 0.05 significance level:

	1	2	3	4	5	6	7
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29
10	4.97	4.10	3.71	3.48	3.33	3.22	3.14
11	4.84	3.98	3.59	3.36	3.20	3.10	3.01
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91

One Way Independent ANOVA- Assumptions

- Normally distributed data (what test?)
- Equality of Variance (what test?)
- Interval or ratio data
- Independent data

One Way Independent ANOVA- Assumptions

- Normally distributed data (Shapiro-Wilk)
- Equality of Variance (Levene's)
- Interval or ratio data
- Independent data

Repeated Measures ANOVA- Sphericity

- independence of data doesn't hold
  - data is from the same participants
- Instead we look for sphericity
  - variance of the differences between scores in each treatment are equal
  - Calculate difference between pairs of scores in all possible combination of treatment levels ,then calculate the variance of these differences

Mauchly's test of sphericity

- It tests the hypothesis that the variances of the differences are equal (H0)

- If I got $p<0.05$ for this test would it be good or bad?

Mauchly's test of sphericity

- It tests the hypothesis that the variances of the differences are equal (H0)

- If I got $p<0.05$ for this test would it be good or bad?
- It would be bad as it states there is a significant difference between variance of differences

- Corrections exist if this is the case, usually Greenhouse-Geisser correction is used

One Way Repeated Measures ANOVA- Assumptions

- Normally distributed data for each condition (Shapiro-Wilkes test)
- Sphericity
- Interval or ratio data

Running Independent ANOVA in R

Code:

```
model <-aov(score~condition, data=data)

summary(model)
```

Output:

```
> summary(model)
            Df Sum Sq Mean Sq F value Pr(>F)
condition    2  16.67   8.337   5.097 0.0101 *
Residuals   45  73.61   1.636
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Running Repeated Measures ANOVA in R

Code:

Install.packages ('ez'); library ('ez')

analysis <-ezANOVA(data=wai.Tot.Anova, dv=.(wai.score), wid=.(userid), within=.(wai.time), between=.(condition), type=3)

analysis

Output:

```
> analysis
$ANOVA
              Effect DFn DFd         F           p p<.05          ges
2          condition   1  42 0.1435817 0.70665485       0.0029651789
3           wai.time   2  84 3.7235569 0.02822214     * 0.0113989211
4 condition:wai.time   2  84 0.2371799 0.78937583       0.0007339116

$`Mauchly's Test for Sphericity`
              Effect         W           p p<.05
3           wai.time 0.7000419 0.0006684095     *
4 condition:wai.time 0.7000419 0.0006684095     *

$`Sphericity Corrections`
              Effect       GGe      p[GG] p[GG]<.05       HFe      p[HF]
3           wai.time 0.7692556 0.04021494         * 0.792803 0.03878871
4 condition:wai.time 0.7692556 0.73058235           0.792803 0.73751657
  p[HF]<.05
3         *
4
```

Reporting ANOVA

- F ratio
- Degrees of Freedom (dof_M, dof_R)
- P value
- Back to the example we saw earlier:
  - **F(2, 45)= 5.097, p<0.05**
- We can therefore state that there is a significant effect of secondary task on driving score

Omnibus test & Post Hoc

- Main effect of secondary driving task: $F(2, 45) = 5.097$, $p < 0.05$
- Significant effect of our experiment conditions on driving score
- But how does this break down?
  - Control >Text?
  - Control > Text?
  - Text > Talk?
- We need ***post hoc tests***

Post Hoc Tests

- Used when no specific a priori predictions about the data we have
- They are used for exploratory data analysis
- Pairwise comparisons
  - Like performing t-tests on all the pairs of mean in our data
  - There are many to choose from……..

A selection of common post hoc tests

- LSD (Least Significant Difference)
  - Analogous to multiple t-tests
- Bonferroni
  - Uses Bonferroni correction to control for Type I
  - With multiple comparisons this may be too conservative (increase chance of Type II error)
- Tukey's test
  - Control Type I and better when testing large number of means

Which one to choose?

- Trade off between:
  - Type I error rate likelihood
  - Statistical power (ability to find an effect if there is one)
  - Whether assumptions of ANOVA have been violated, although most are robust to minor variations

Running Post Hoc tests in R

Code:

```
pairwise.t.test (data$score, data$condition, paired=FALSE,
p.adjust.method="bonferroni")
```

Output:

```
    Pairwise comparisons using t tests with pooled SD

data:  data$score and data$condition

     control talk
talk 0.073   -
text 0.011   1.000

P value adjustment method: bonferroni
```

Lecture Readings and Further concepts to consider

- Core:- Field (2009) Chapters 8 & 11 (pages 427-454)

- Other concepts to consider:
  - **Statistical Power:** Cohen (1992). A power primer. Psychological Bulletin
  - **Planned Contrasts**: Field (2009), Chapter 8, p.325-339

Two Way ANOVA

Evaluation Methods & Statistics- Lecture 11

Benjamin Cowan & Andrew Howes

Research Example (Last Time)

- Consequences of a secondary task on driving
  - Texting
  - Talking on phone
  - Control (no secondary task)

Research Example (This week)

Consequences of a secondary task on driving

- Texting
- Talking on phone
- Control (no secondary task)

Gender effect

- Male
- Female

What are our DV and IVs?

What are our DV and IVs?

IV 1- Secondary Driving Task

- Level 1- Control Group (No secondary task)
- Level 2- Texting
- Level 3- Talking

IV 2- Participant Gender

- Male
- Female

DV-Driving score

How would we analyse the data?

- We could do lots of t-tests
  - Control to Texting for Males
  - Control to Talking for Males
  - Talking to Texting for Males
  - Control to Texting for Females
  - Control to Talking for Females
  - Talking to Texting for Females
  - Control (Males) to Control (Females) etc…..

- This would inflate our *Type I error rate*

Factorial ANOVA- The Idea

- One Way ANOVA cannot deal with two factors (i.e. two Independent variables)

- We need to use a different type of ANOVA if we have two independent variables

Types of Factorial ANOVA

- Independent factorial design
  - All independent variables are between subjects
- Repeated Measures
  - All independent variables are within subjects
  - E.g. Measuring satisfaction pre and post (IV 1) 3 different interfaces (IV 2) all experienced one after another
- Mixed Design
  - Some independent variables are within subjects and some are between subjects
  - Gender (IV1) on pre and post scores (IV2)

Types of Factorial ANOVA

- What type of ANOVA are we going to use in our example?

Types of Factorial ANOVA

What type of ANOVA are we going to use in our example?

- Independent factorial design
  - All independent variables are between subjects

ANOVA Names

- Number of IV's
- Number of levels in IV's
- Experiment design used to gather data
  - Independent (Between subjects)
  - Repeated Measures (Within subjects)

Example

If we had:

- IV 1- Gender (2 levels- Between)
- IV2- Secondary Task (2 levels- Between)

2x2 Independent ANOVA

– or-

Two Way Independent ANOVA

Example

If we had:

- IV 1- Gender (2 levels- Between Subjects)
- IV2- Secondary Task (3 levels- Between Subjects)

2x3 Independent ANOVA

– or-

Two Way Independent ANOVA

Example

If we had:

- IV 1- Gender (2 levels) (Between Subject)
- IV2- Secondary Task (3 levels) (Within Subjects)

2x3 Mixed Design ANOVA

-or-

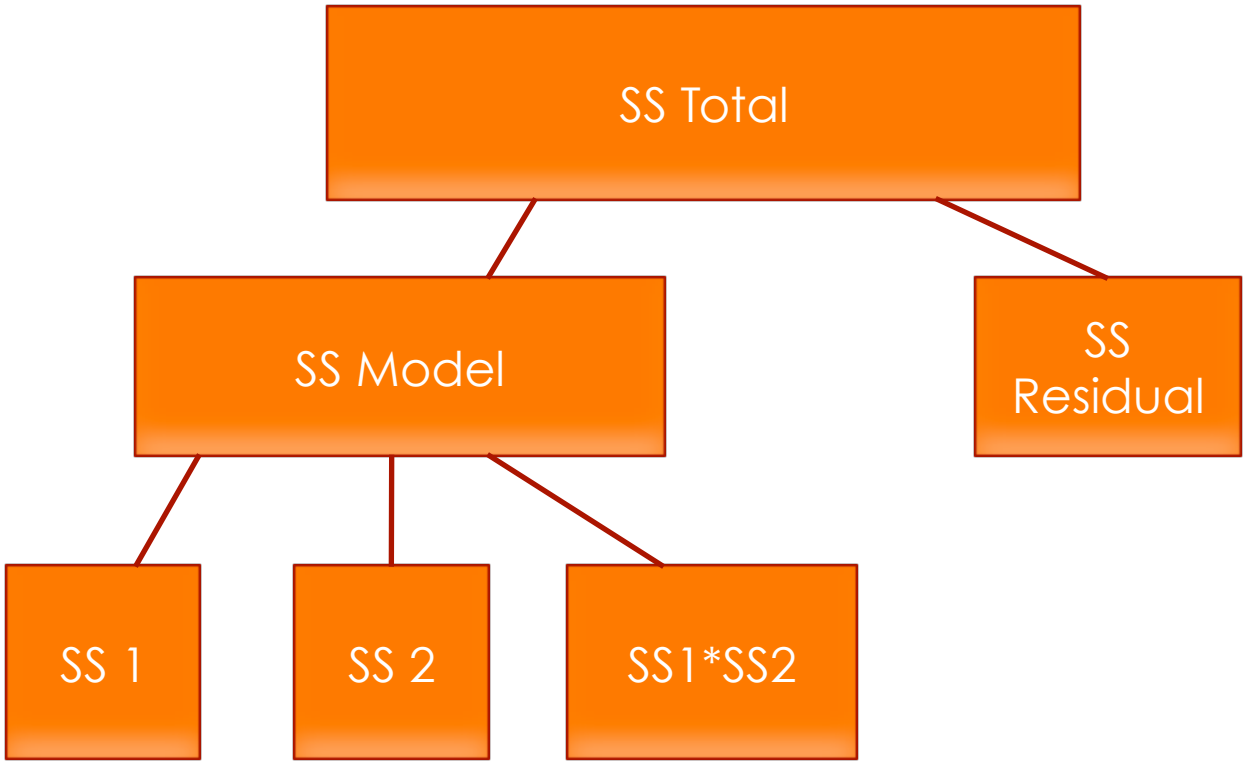
Two Way Mixed Design ANOVA

The Key (again): ANOVA & F Ratio

- F ratio is the ratio of **explained (that accounted for by the model we are proposing)** to **unexplained** variation

- This is calculated using the **Mean Squares**

What ANOVA is doing



Step 1- Total Sum of Squares

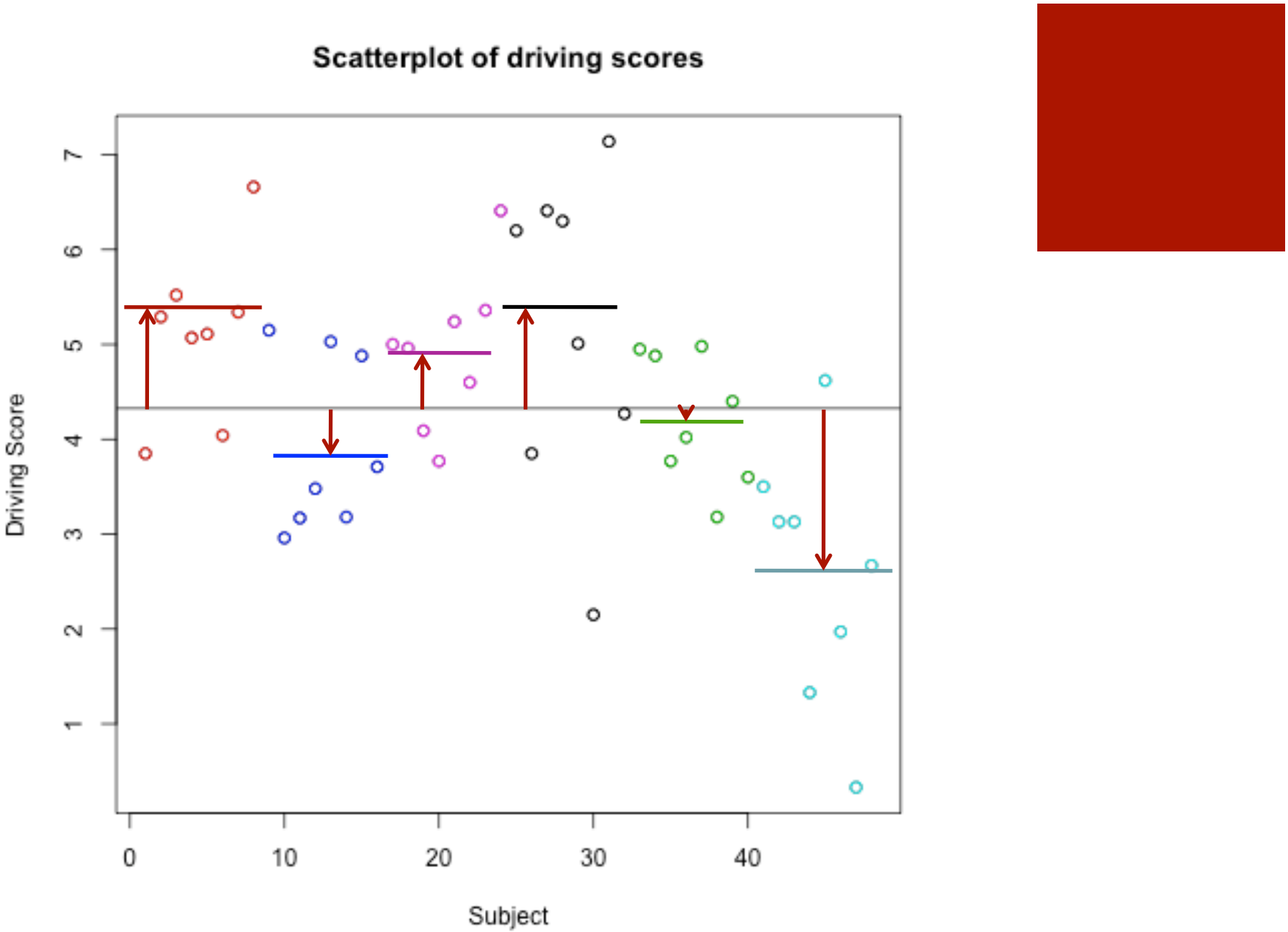
- The total amount of variation in our data
- This should look familiar (see Lecture 6)

$$SS_T = \sum \left(x_i - \overline{x}_{grand} \right)^2$$

Step 2- Model Sum of Squares

- We now need to know how much variation our model can explain
- How much the total variation can be explained due to data points coming from different groups in "the perfect model"
- n_k is the number of people in that condition
- Sum all levels of IVs together

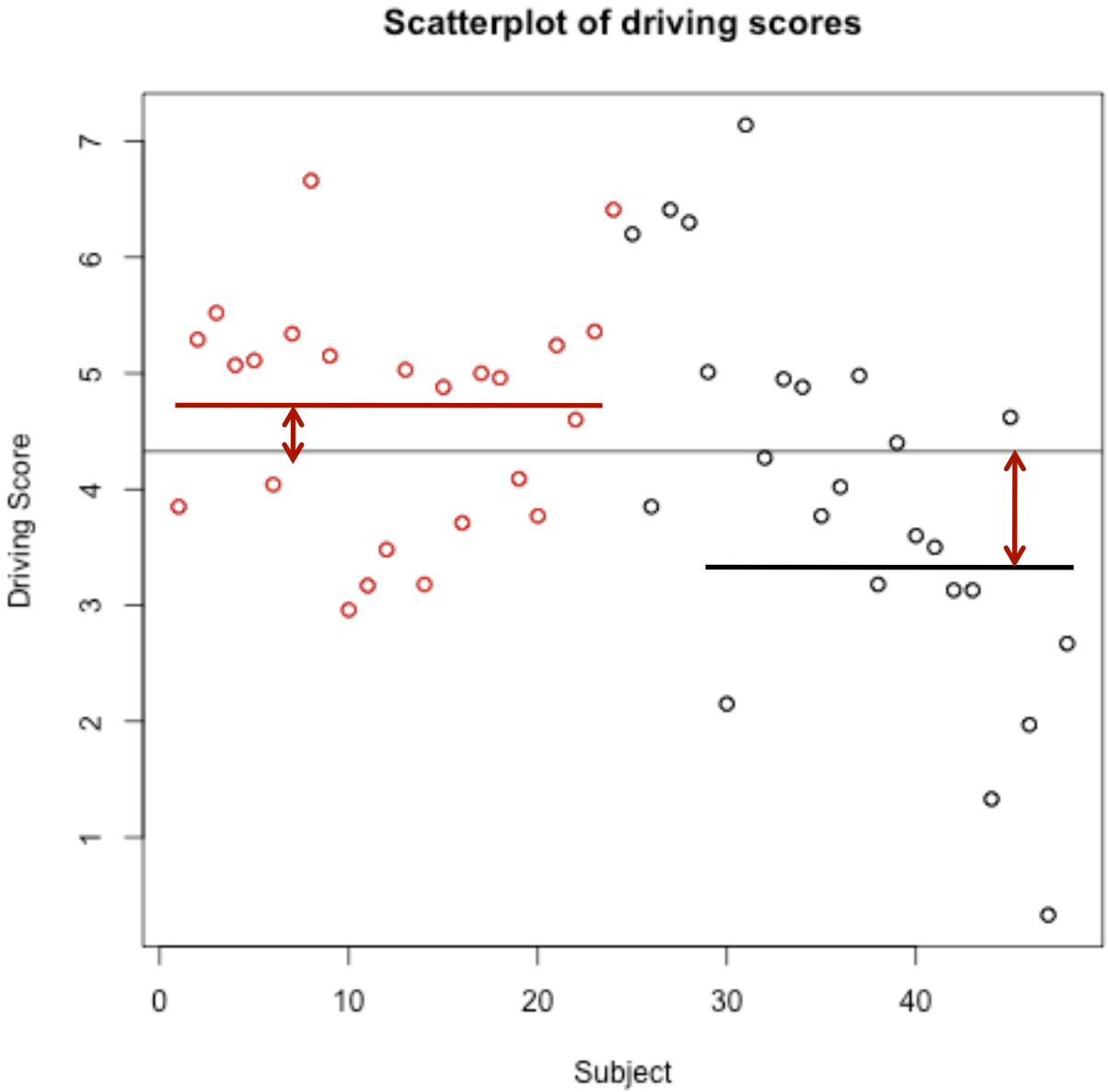
$$SS_M = \sum n_k (\bar{x}_k - \bar{x}_{grand})^2$$


Scatterplot of driving scores
Driving Score
1
2
3
4
5
6
7
0
10
20
30
40
Subject

Step 3- Main Effect of Gender

- We group data by levels of Gender alone
- How much the total variation can be explained due to data points coming from different gender groups only in "the perfect model"
- n_k is the number of people in that condition
- Sum both levels of IV together

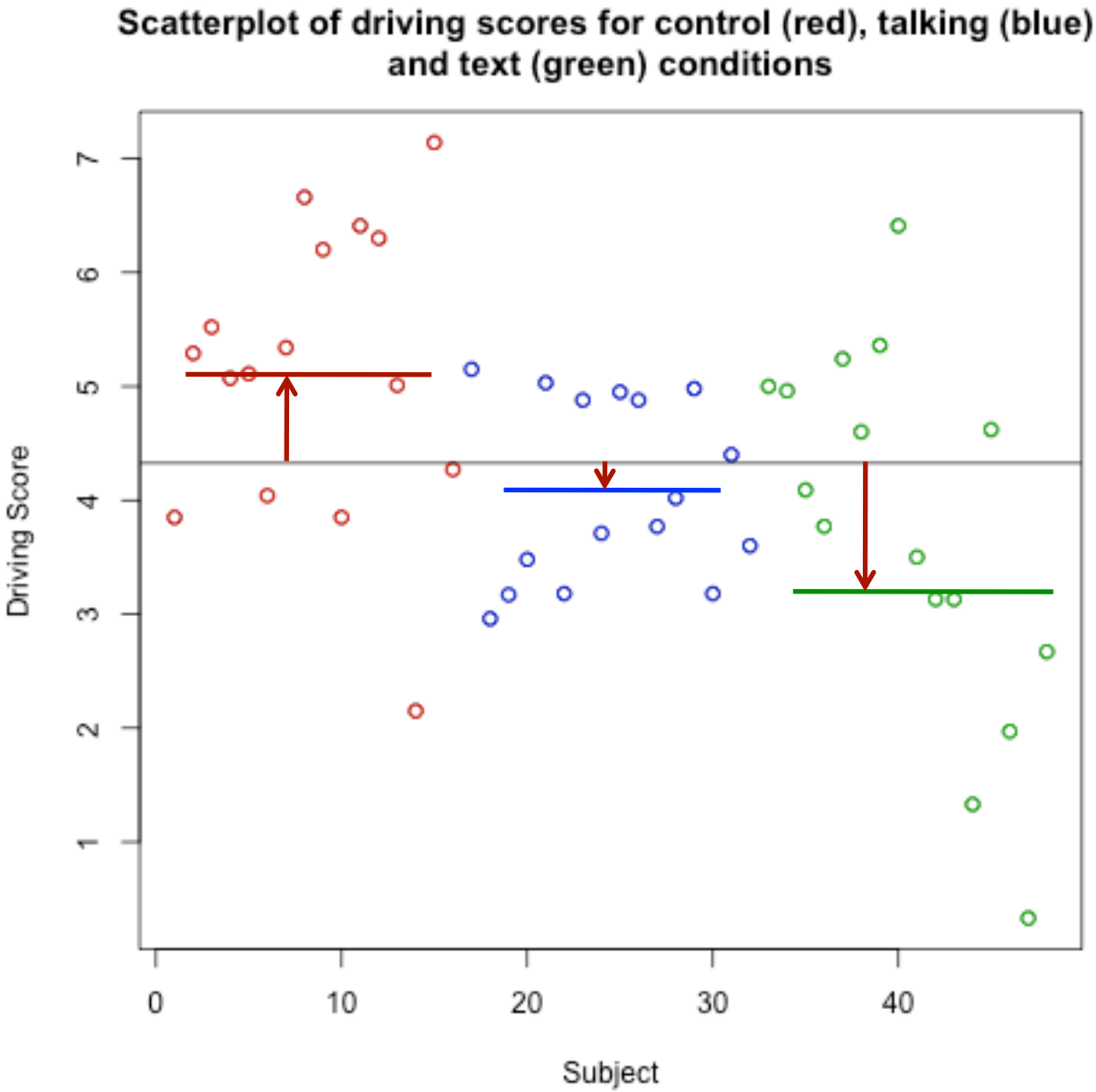
$$SS_1 = \sum n_k(\overline{x}_k - \overline{x}_{grand})^2$$


Scatterplot of driving scores
Driving Score
Subject
0
10
20
30
40
1
2
3
4
5
6
7

Step 4- Main Effect of Task

- We group data by levels of Task alone
- How much the total variation can be explained due to data points coming from different Task conditions alone in "the perfect model"
- n_k is the number of people in that condition
- Sum levels of IV together

$$SS_2 = \sum n_k (\overline{x}_k - \overline{x}_{grand})^2$$


Scatterplot of driving scores for control (red), talking (blue) and text (green) conditions
Driving Score
1
2
3
4
5
6
7
0
10
20
30
40
Subject

Step 5- Interaction (Gender *Task)

- SSm made up of 3 components:
  - SS1, SS2, SS1*2
  - Therefore easiest way is to take away SS1 and SS2 from SSM

$$SS_{1*2} = SS_M - SS_1 - SS_2$$

Step 6- Residual Sum of Squares

- How much of the variation cannot be explained by the model i.e. what error is there in the model prediction?

- Easy way to calculate: $SS_R = SS_T - SS_M$

Degrees of Freedom for each SS

- Degrees of Freedom for SS_T (dfT):
  - N-1 = 48-1 = 47
- Degrees of Freedom for SS_M(dfM):
  - Gender: Number of Conditions (k) -1 = 1
  - Task: Number of Conditions (k) -1 = 2
  - Gender*Task= df Gender*df Task = 2
- Degrees of Freedom for SS_R(dfR):
  - (N-1)*Number of groups = (8-1)*6= 42

F Ratio

- F Ratio gained for each effect (IV) and the interaction

- Mean Squares model (MS_M):
  - SS_M/df_M
  - This is done for main effects SS and interaction SS

- Mean Squares residual (error) (MS_R):
  - SS_R/df_R

F Ratio

$$\frac{\text{Mean Square Model } (MS_M)}{\text{Mean Square Residual } (MS_R)}$$

F Ratio

$$\frac{\text{Variation explained by our model}}{\text{Variation unexplained by our model}}$$

F Distribution

- F Distribution for specific pair of degrees of freedom

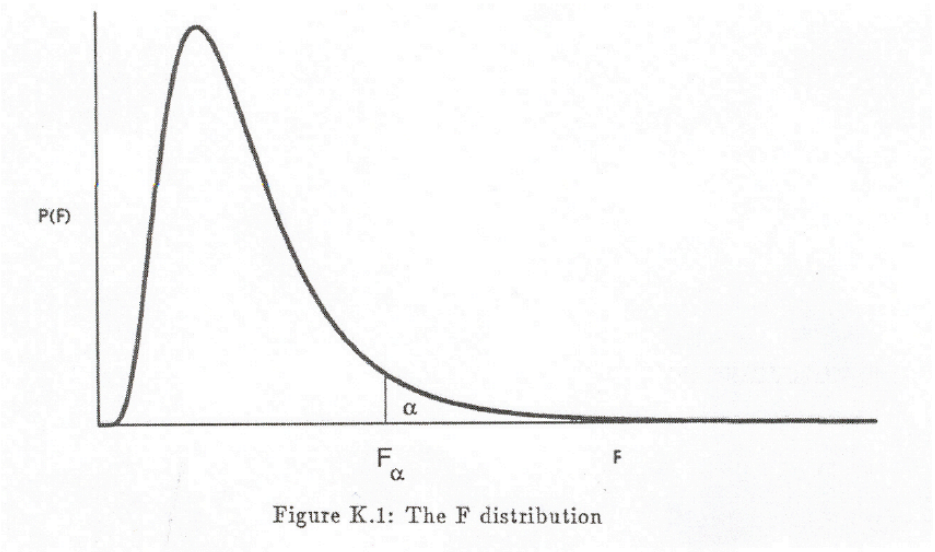


Figure K.1: The F distribution

- Table of Critical Values

Critical values of F for the 0.05 significance level:

	1	2	3	4	5	6	7
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29
10	4.97	4.10	3.71	3.48	3.33	3.22	3.14
11	4.84	3.98	3.59	3.36	3.20	3.10	3.01
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91

Output in R

```
> summary(analysis)
                 Df Sum Sq Mean Sq F value  Pr(>F)
gender            1   5.39   5.387   4.409 0.04180 *
condition         2  16.67   8.337   6.823 0.00272 **
gender:condition  2  16.91   8.453   6.918 0.00253 **
Residuals        42  51.32   1.222
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Main Effect

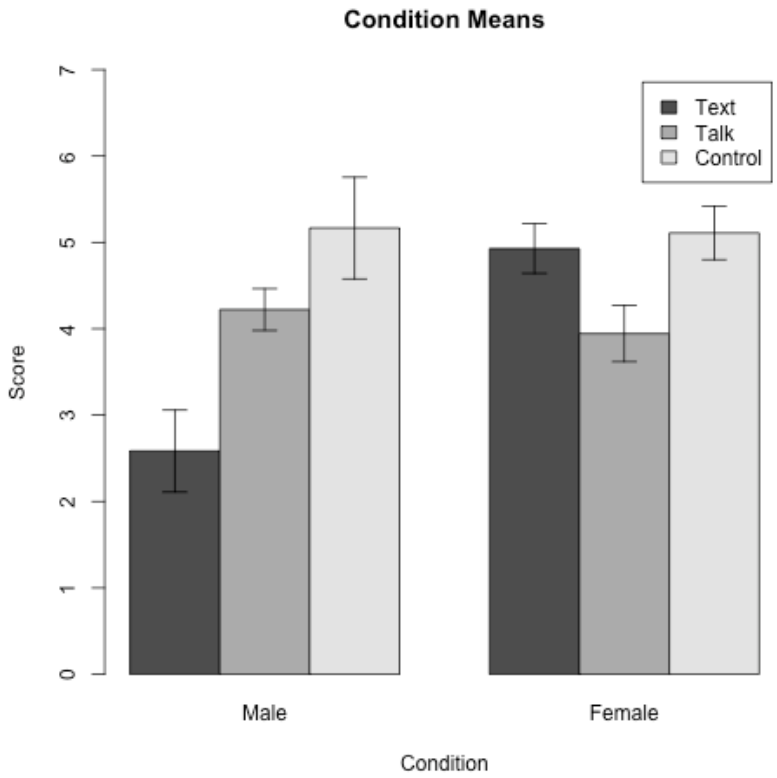
- Significant effect of an IV irrespective of the other IV

- Significant effect of gender on driving score irrespective of the task completed

Output in R

```
> summary(analysis)
                 Df Sum Sq Mean Sq F value  Pr(>F)
gender            1   5.39   5.387   4.409 0.04180 *
condition         2  16.67   8.337   6.823 0.00272 **
gender:condition  2  16.91   8.453   6.918 0.00253 **
Residuals        42  51.32   1.222
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Interactions



- The effect of secondary task on score is not that same for male and females
- interaction effect
- Females have a higher score in the texting condition compared to males
- Interactions supersedes main effects

Omnibus test & Post Hoc

- ANOVA is omnibus test
- How do they break down?
  - Male vs Female ? (Main Effect)
  - Control vs Text, Control vs Talk, Talk vs Text ? (Main effect of Task)
  - Male:Control vs Female Control etc (interaction)
- Again we need ***post hoc tests (See last lecture)***

Reporting ANOVA

- F ratio
- Degrees of Freedom (dof_M, dof_R)
- P value
  - **F(2, 42)= 5.097, p<0.05**

Output in R

```
> summary(analysis)
                 Df Sum Sq Mean Sq F value  Pr(>F)
gender            1   5.39   5.387   4.409 0.04180 *
condition         2  16.67   8.337   6.823 0.00272 **
gender:condition  2  16.91   8.453   6.918 0.00253 **
Residuals        42  51.32   1.222
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Reporting ANOVA

e.g. Main effect of Gender

- **F(1, 42)= 4.409, p<0.05**

Reading

- Field (2012) Discovering Statistics Using R, Chapters 12 -14