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?

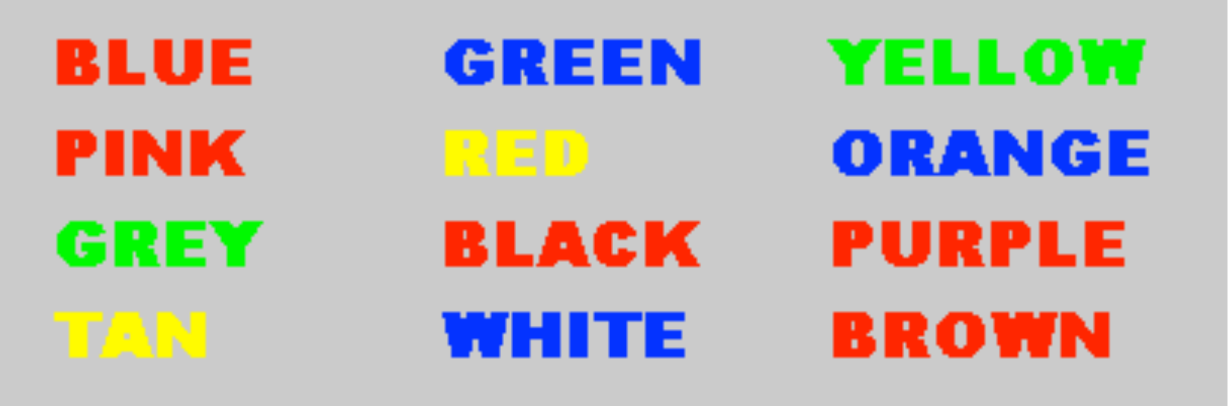
- Your mind has now primed "Google" (Sparrow et al., 2011).
- Sparrow, B., Liu, J. & Wegner, D.M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science, 333*, 776-778.

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.



BLUE	GREEN	YELLOW
PINK	RED	ORANGE
GREY	BLACK	PURPLE
TAN	WHITE	BROWN

- 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.

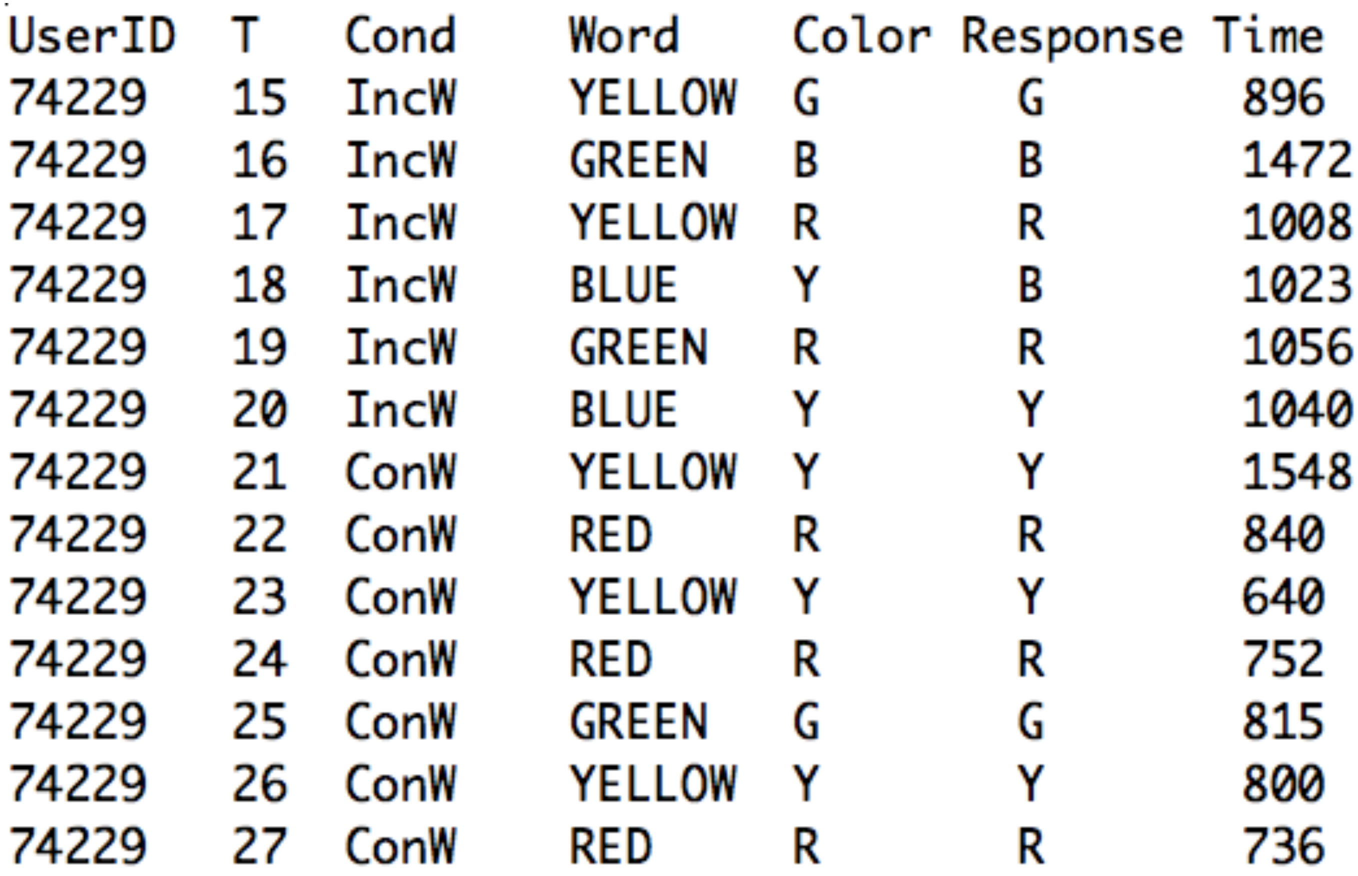
how is the experiment conducted?

- 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?



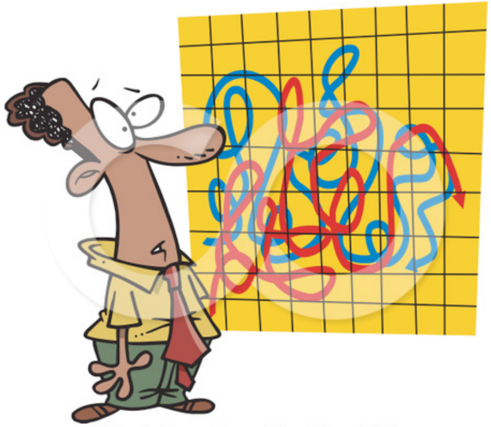
UserID	T	Cond	Word	Color	Response	Time
74229	15	IncW	YELLOW	G	G	896
74229	16	IncW	GREEN	B	B	1472
74229	17	IncW	YELLOW	R	R	1008
74229	18	IncW	BLUE	Y	B	1023
74229	19	IncW	GREEN	R	R	1056
74229	20	IncW	BLUE	Y	Y	1040
74229	21	ConW	YELLOW	Y	Y	1548
74229	22	ConW	RED	R	R	840
74229	23	ConW	YELLOW	Y	Y	640
74229	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**.

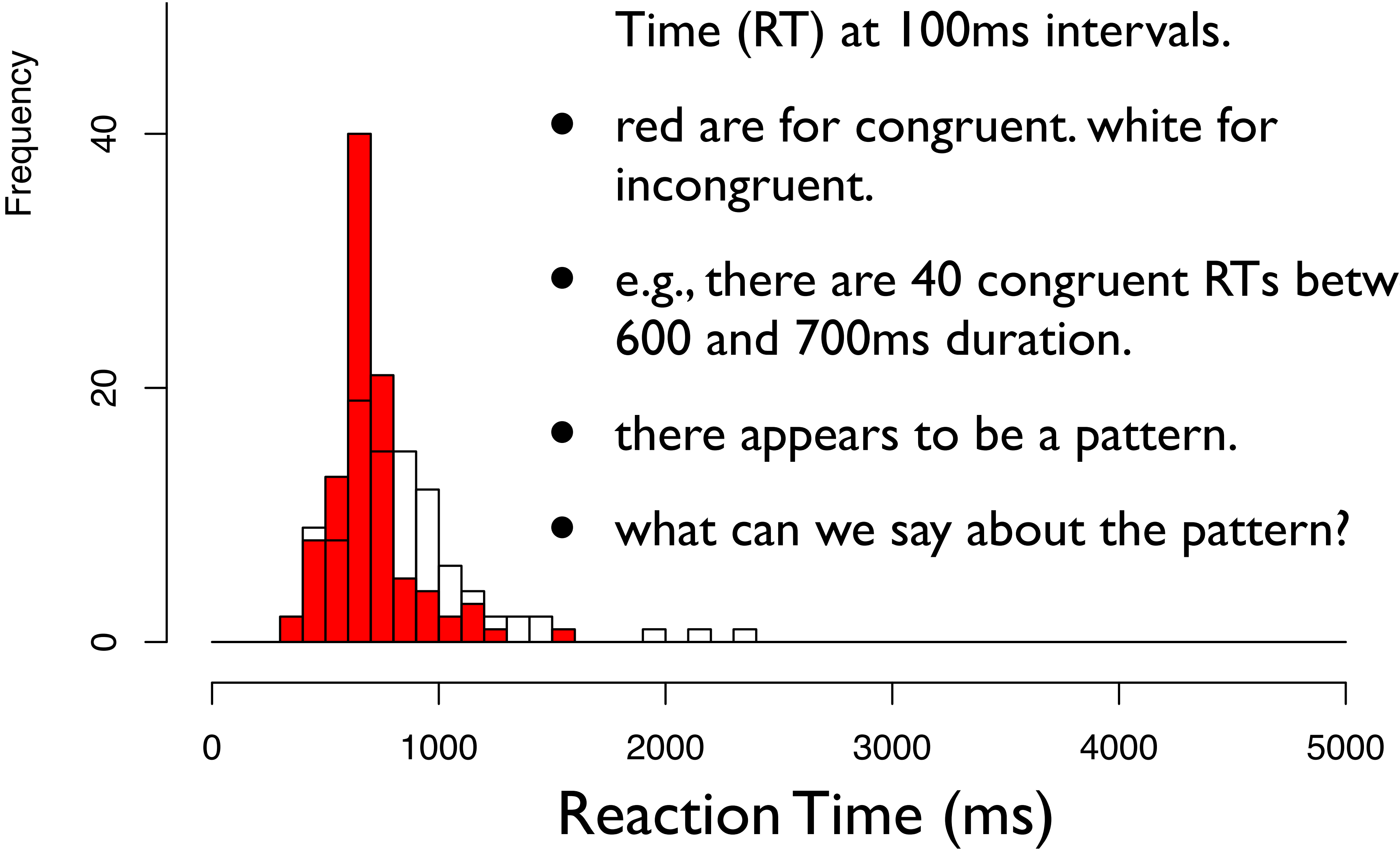
how do we make sense of the data?

- 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?



© Ron Leishman * www.ClipartOf.com/5686

a frequency plot (a histogram)

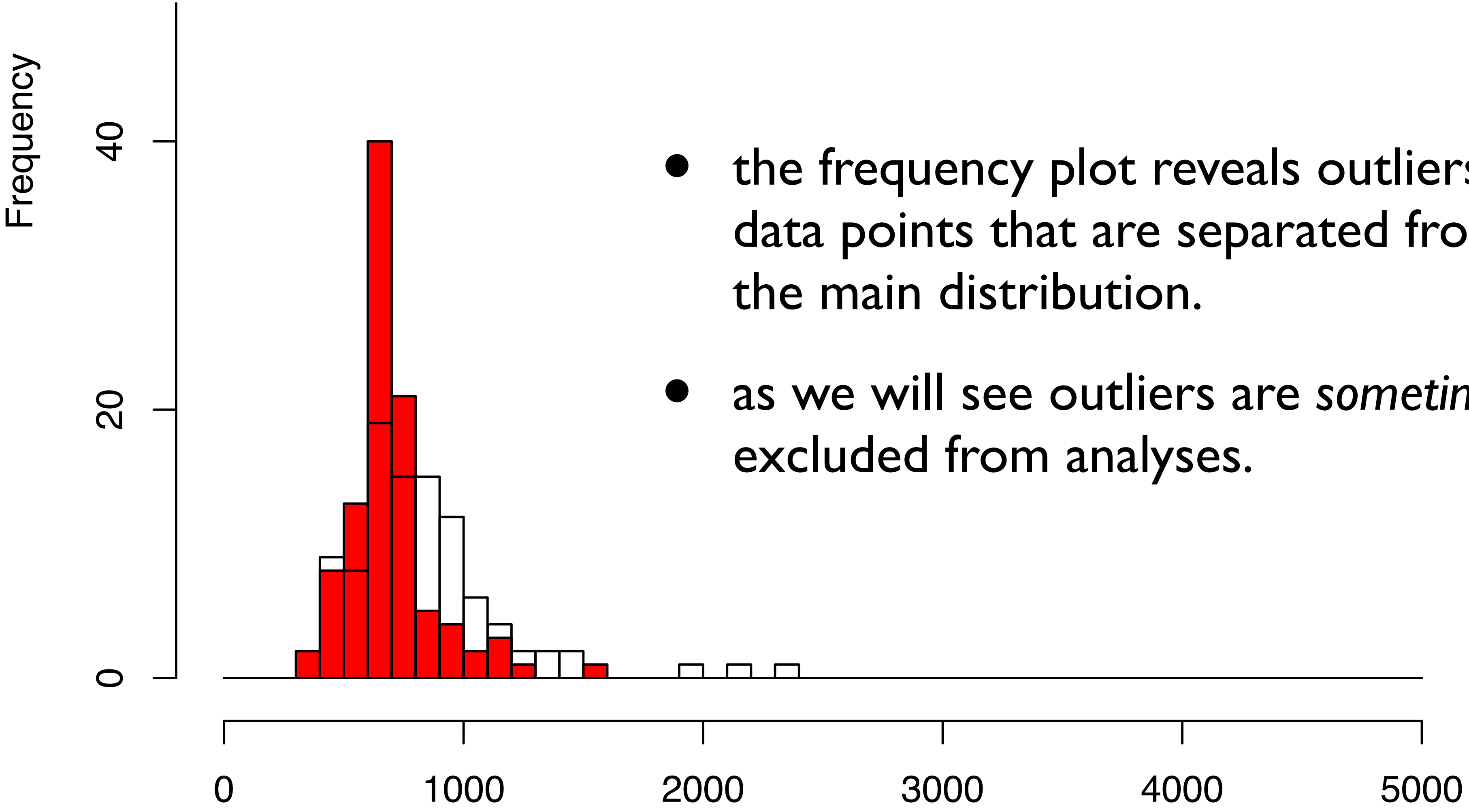


- a plot of the frequency of each Reaction Time (RT) at 100ms intervals.
- red are for congruent. white for incongruent.
- e.g., there are 40 congruent RTs between 600 and 700ms duration.
- there appears to be a pattern.
- what can we say about the pattern?

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



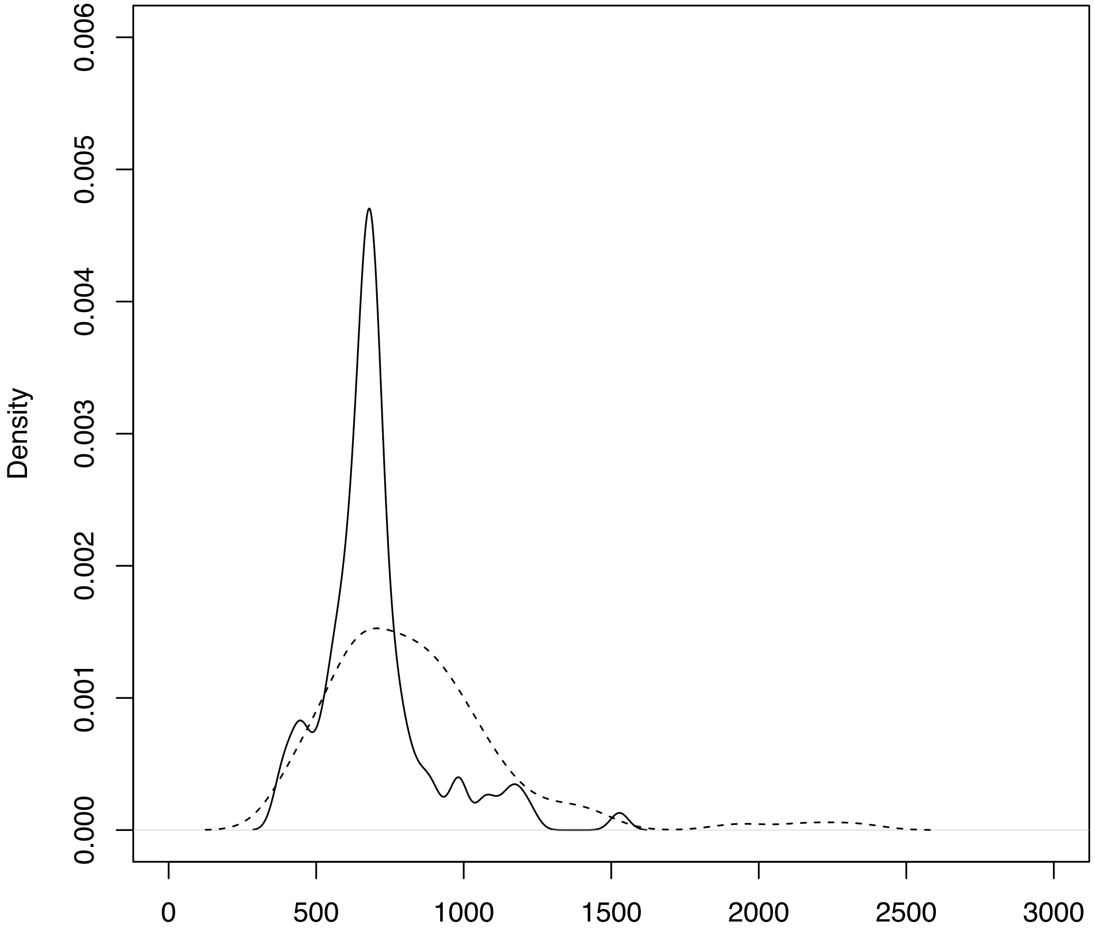
- the frequency plot reveals outliers -- data points that are separated from the main distribution.
- as we will see outliers are *sometimes* excluded from analyses.

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.

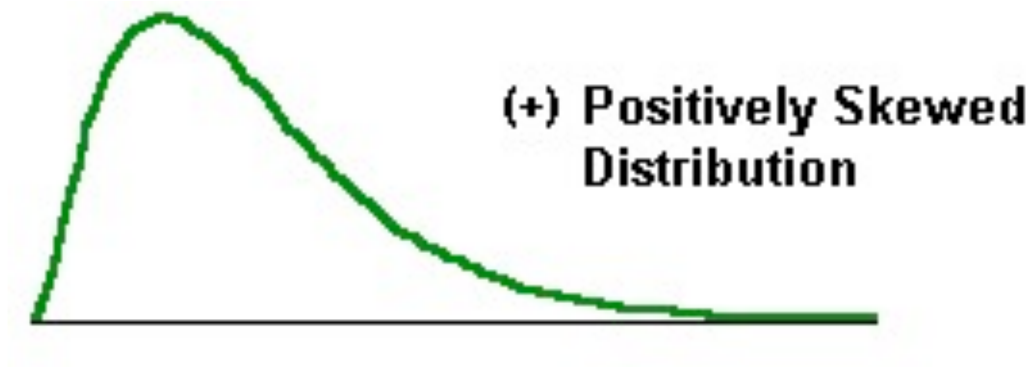
density - another way to plot Reaction Time

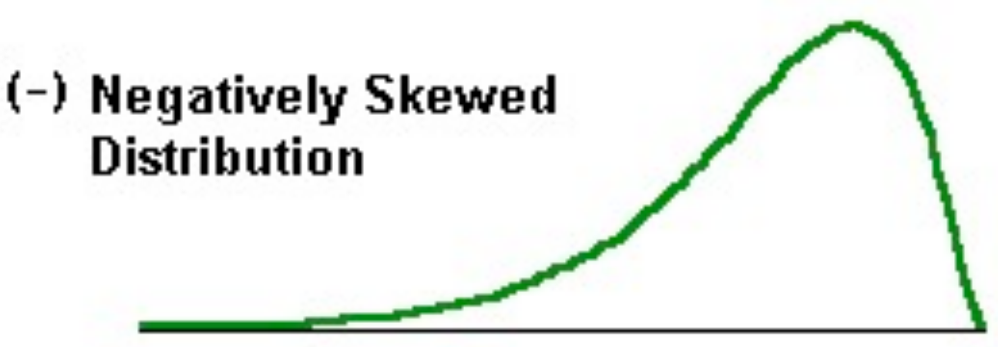
- density is proportional to the probability of drawing a value close to X from the population.



Reaction Time (ms)

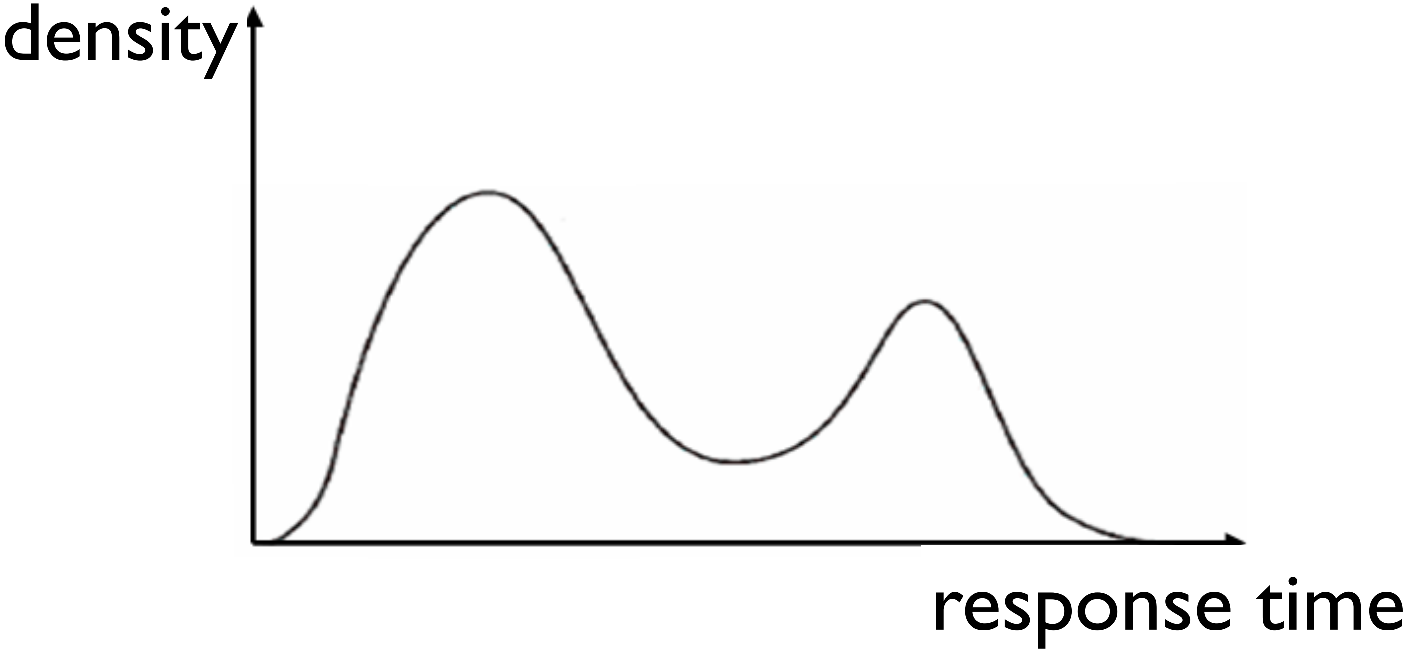
Reaction Times are “positively” skewed





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

bimodal distribution



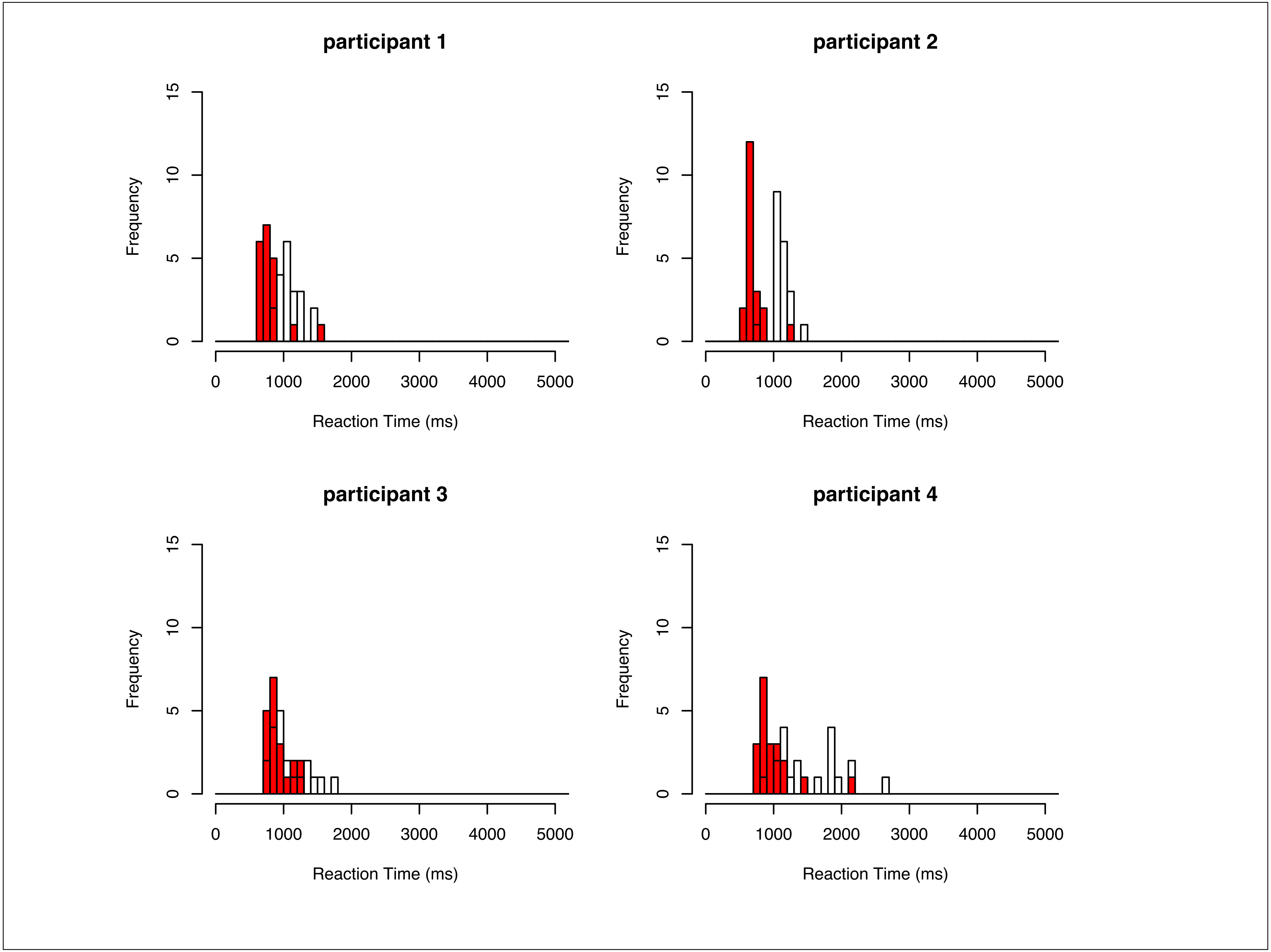
bimodal distributions might be found, for example, because of errors (e.g. failure to recall).

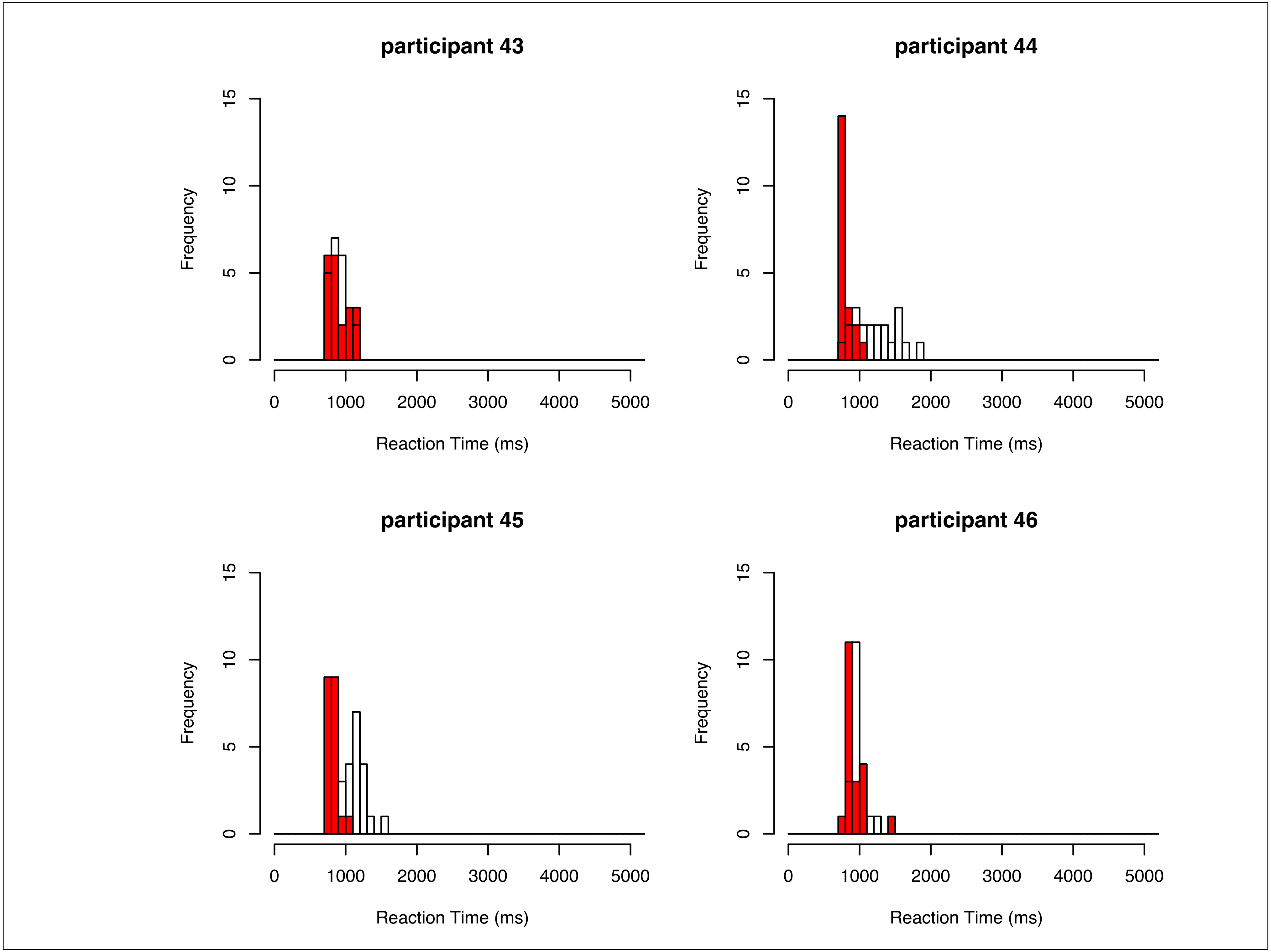
skew, bimodality.

- 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
participant 44
participant 45
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

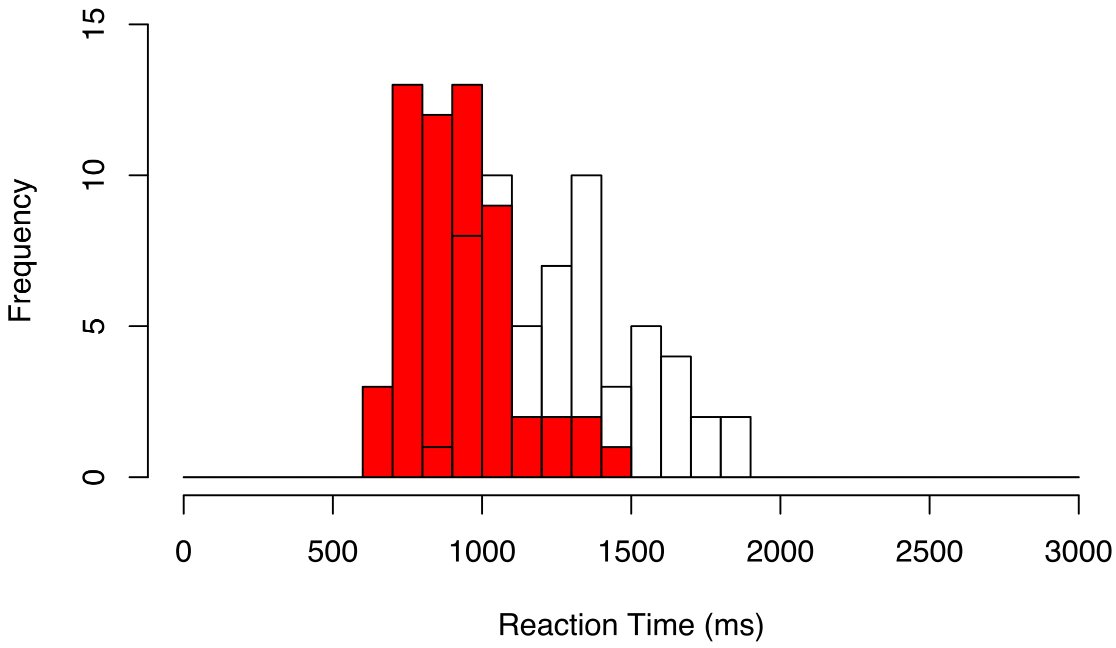
- Consider the following 3 data sets:
- {1, 1, 1, 2, 2, 3, 4, 5, 5, 6}: Mean = 3, Median = 2.5, Mode = 1
- {1, 27, 28, 29, 30}: Median = 28, Mode = XXX
- {1, 1, 1, 1, 1, 1, 2, 2, 3, 3, 3, 3, 50}:Mean =5.5, Median = 2, Mode = 1

Mean, mode, median advantages/ disadvantages

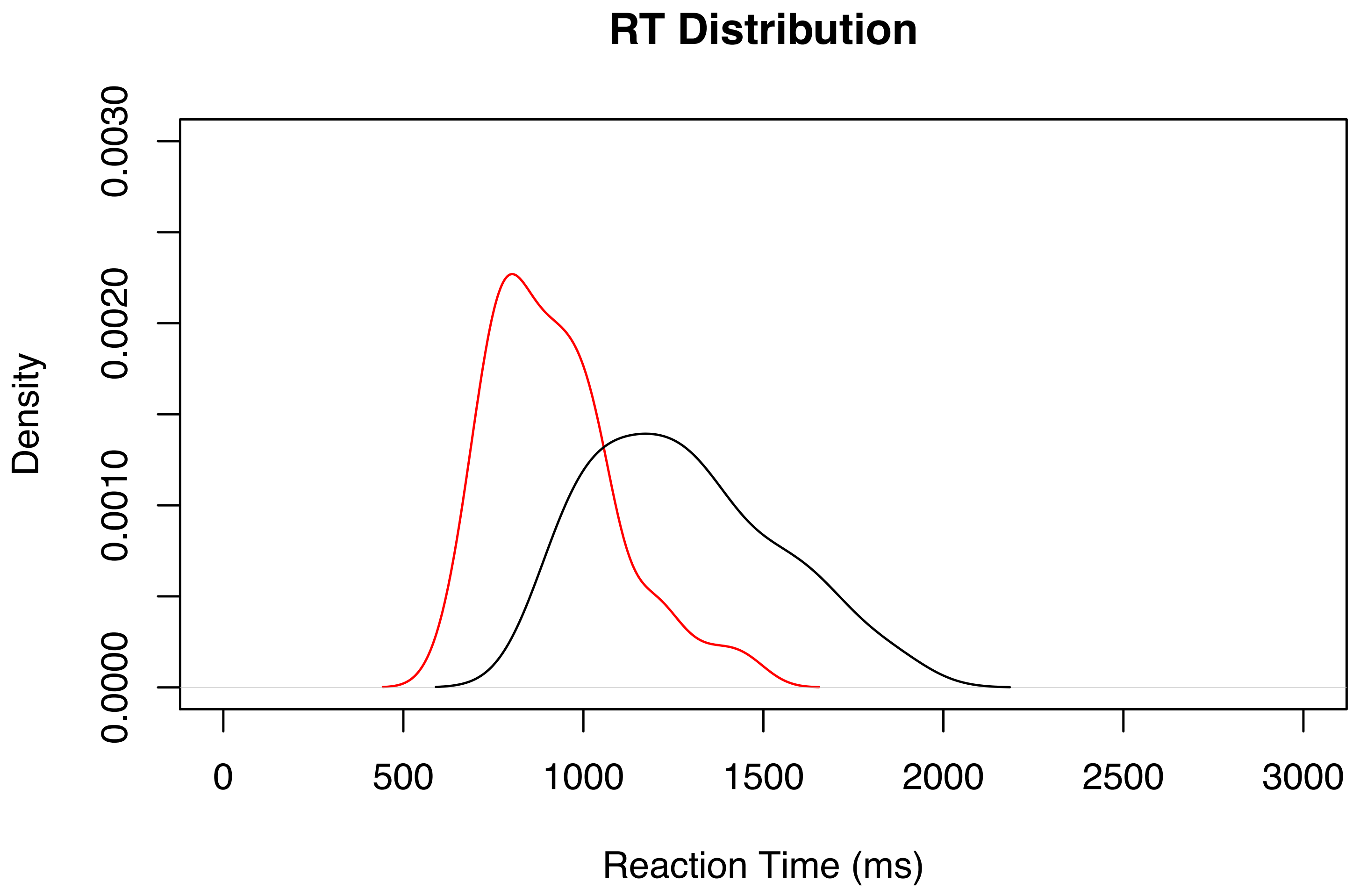
- The **mode** is a score that actually occurred,
- ... whereas the mean and sometimes the median may never have been present in the data.
- If 70% of your customers want "large" T-shirts and 30% want "small", then it does not make sense to stock the mean or "medium" size t-shirt.
- The major advantage 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.

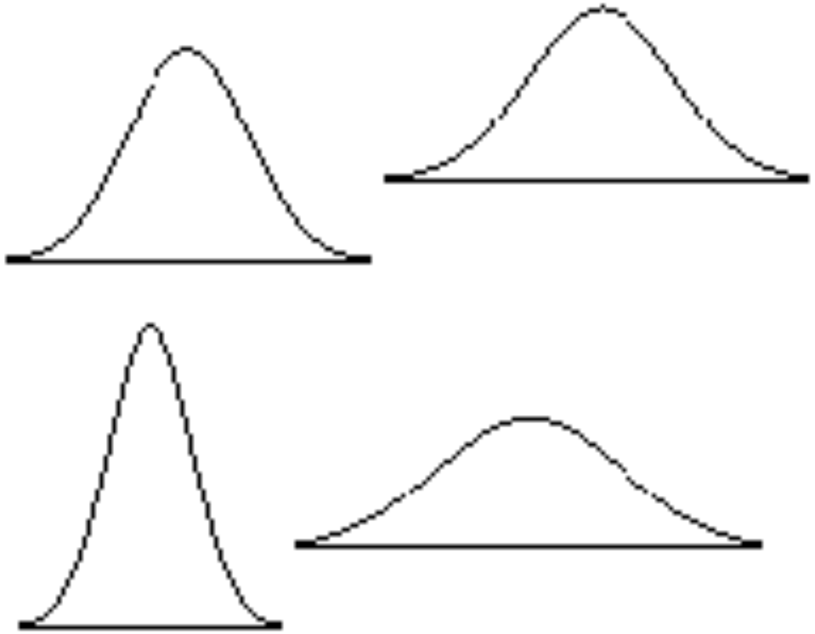


density plots



the Normal distribution

- Normal distributions have a distinctive bell shape density.

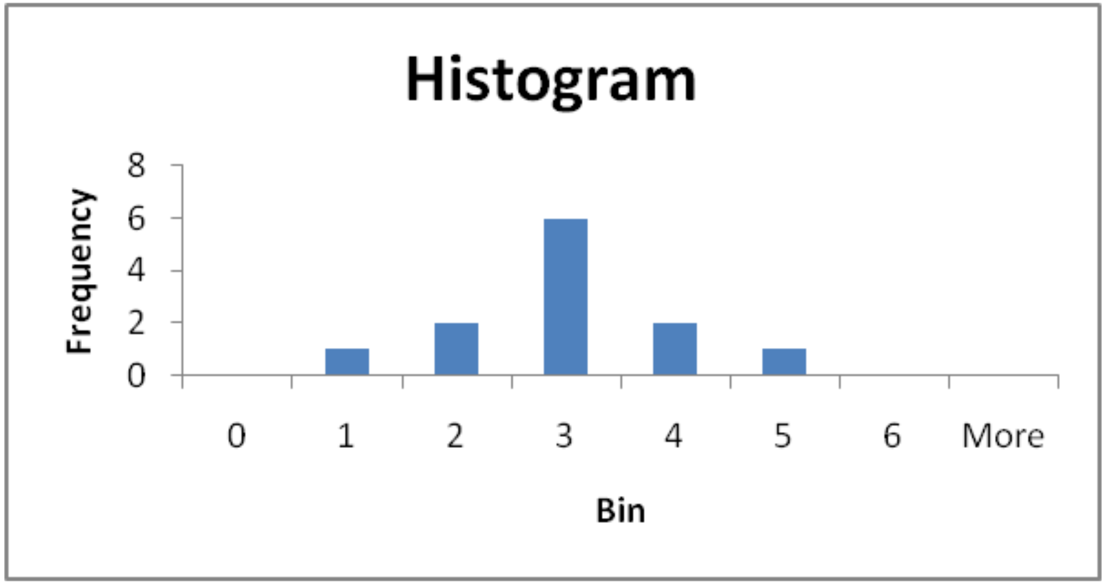


how can we analyse the distributions of means?

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

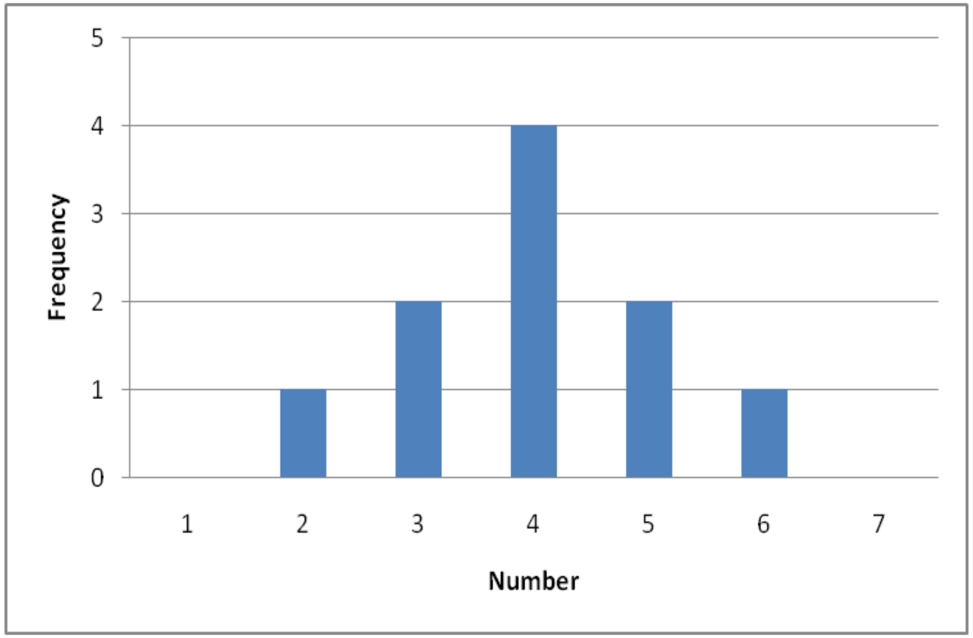
mean, mode, median for a Normal distribution

- 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

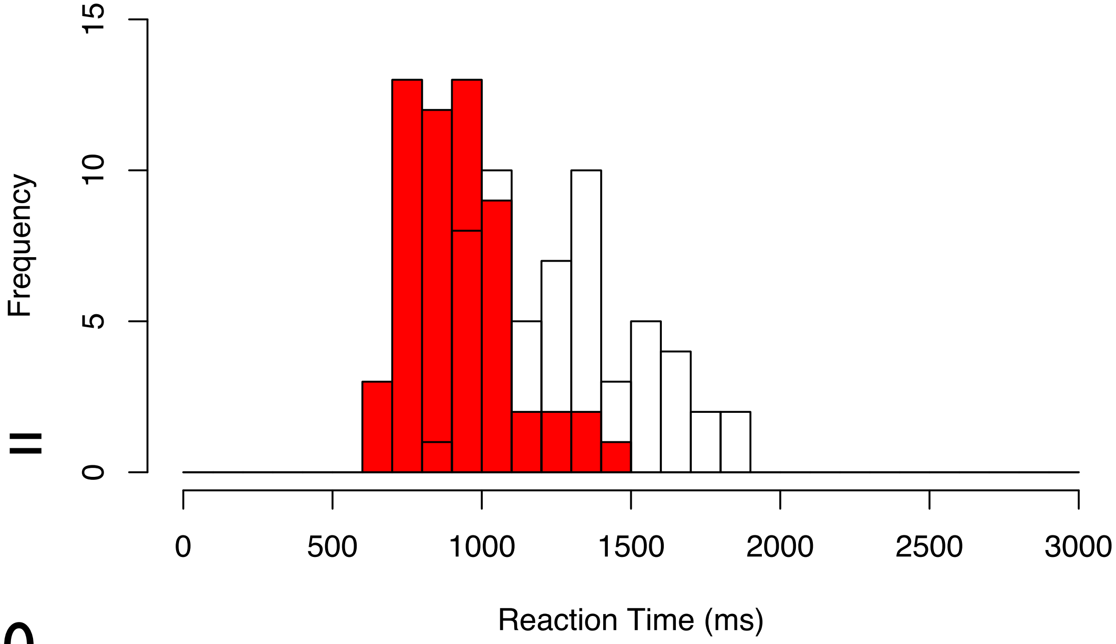
- Range is one measure of Spread/Variability
- DATA: {2, 3, 3, 4, 4, 4, 4, 5, 5, 6}
- (number of years since left school)



- 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.

$$\sum_{1}^{N} X_i$$

Mean $\overline{X}$

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

central limit theorem

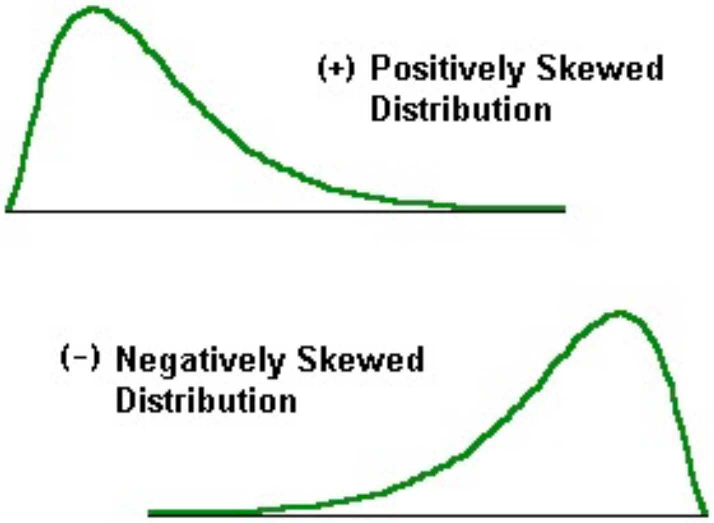
- We have seen that, despite skewed individual distributions, the distribution of mean Stroop RTs appears approximately Normal.
- In fact, there is a mathematical theorem, the Central Limit Theorem, that states that conditions under which the distribution of means will be approximately normal.
- In other 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.


(+) Positively Skewed Distribution
(−) Negatively Skewed Distribution

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.

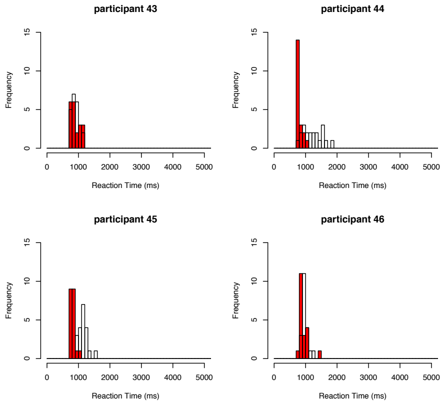
sampling distribution of the median

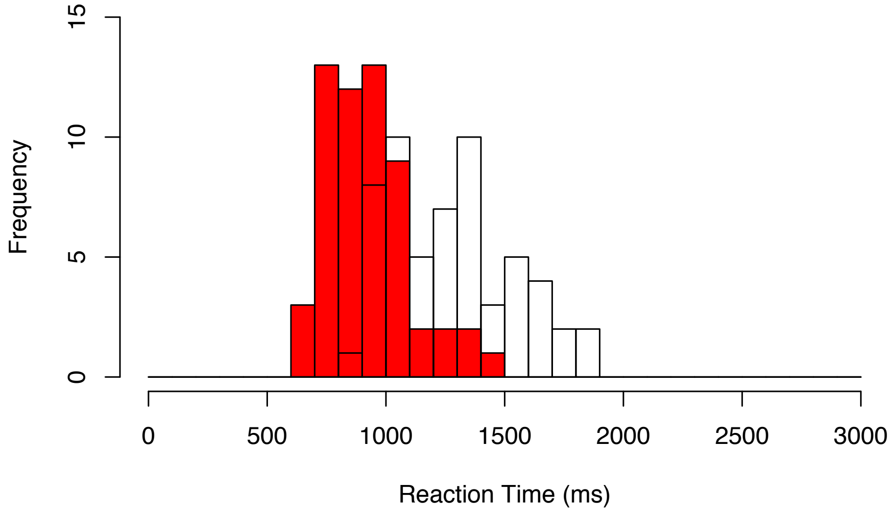
- Is not generally normal except, perhaps, for very large sample sizes.

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) { }