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

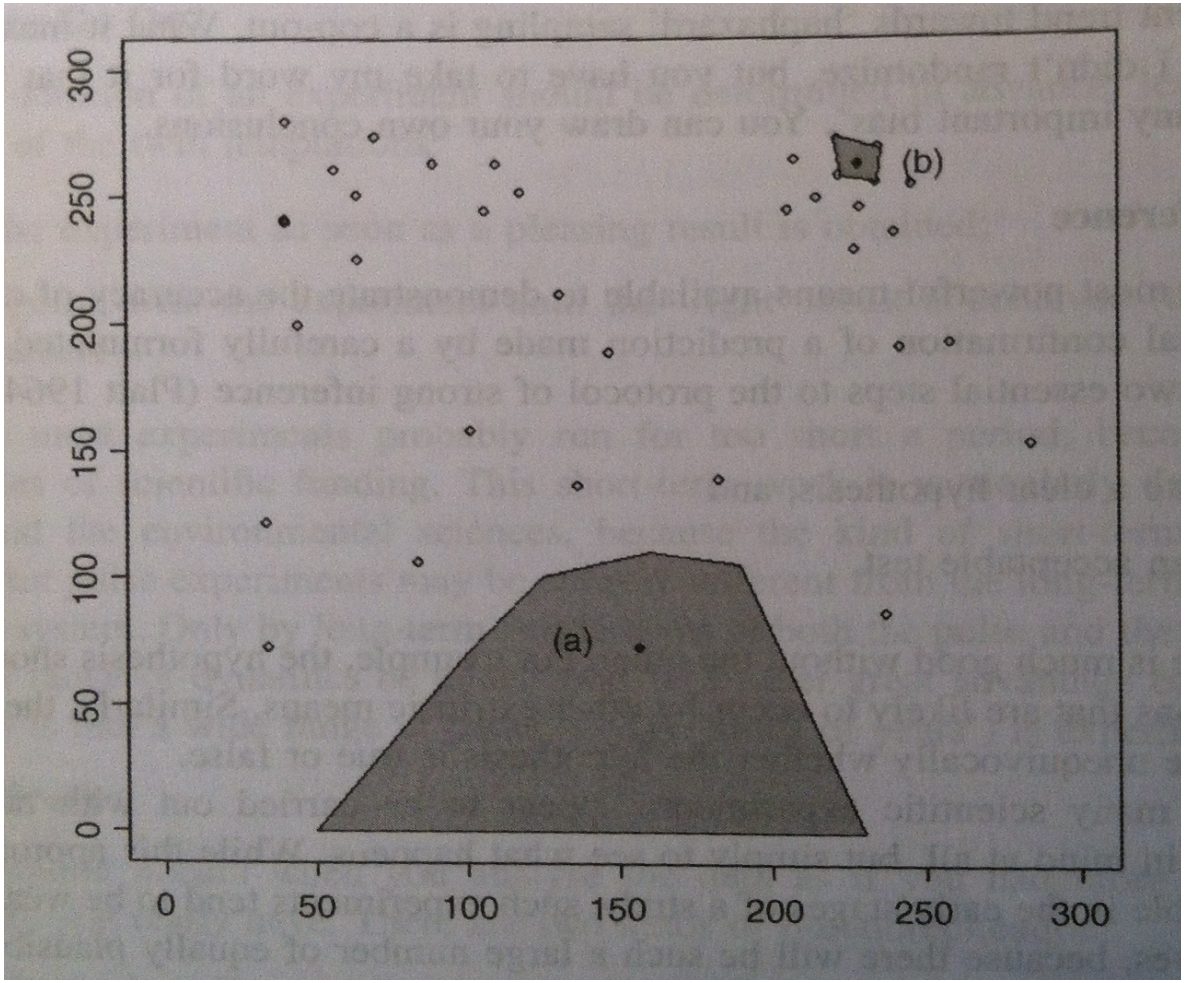
- The **independent** variables define the conditions of the experiment.
- In the Stroop experiment the independent variable was the relationship between the ink color and the word.
- There were two **levels** of the independent variable: congruent and incongruent.
- The dependent variable is what you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

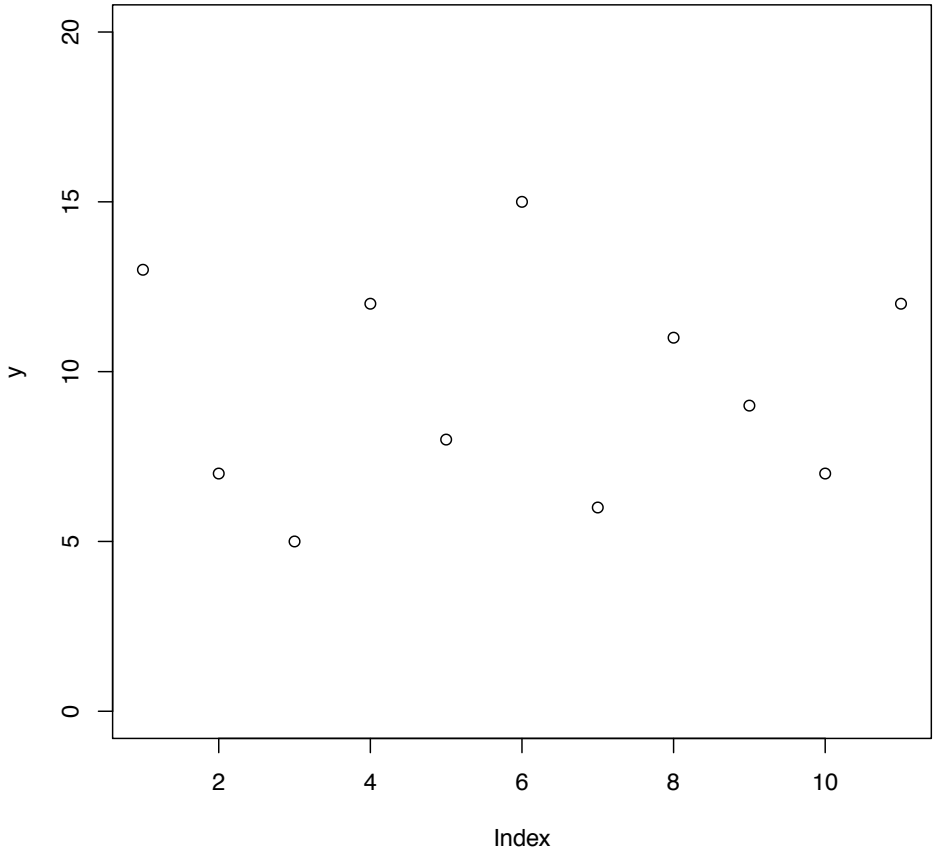
- When we are studying human behaviour we need to select participants randomly if we wish to draw inferences about the population.
- Full randomization from a global population is difficult to achieve.
- There are accepted limits on randomization.
- Student participants introduce age and IQ biases. They 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

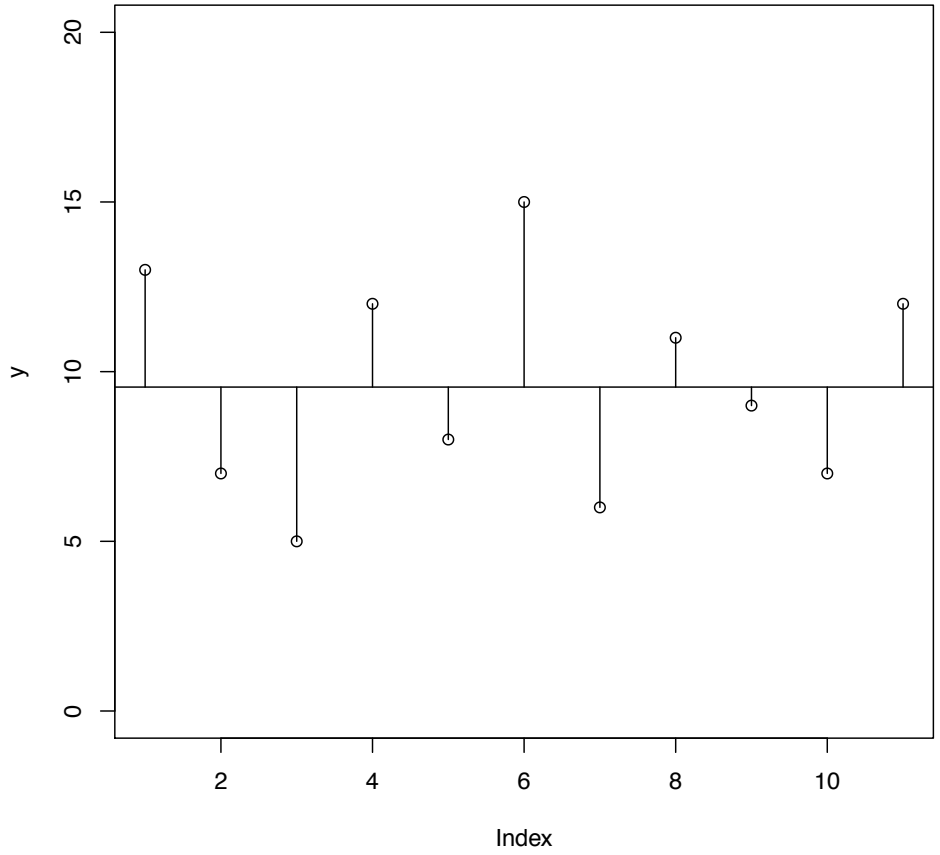
- Perhaps the most important concepts in statistics are those concerned with the representation of variability.
- Range is one simple measure of variation but it fails to capture the extent to which values are clustered, e.g. in a normal distribution.
- What would be a better measure of variability?

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)

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

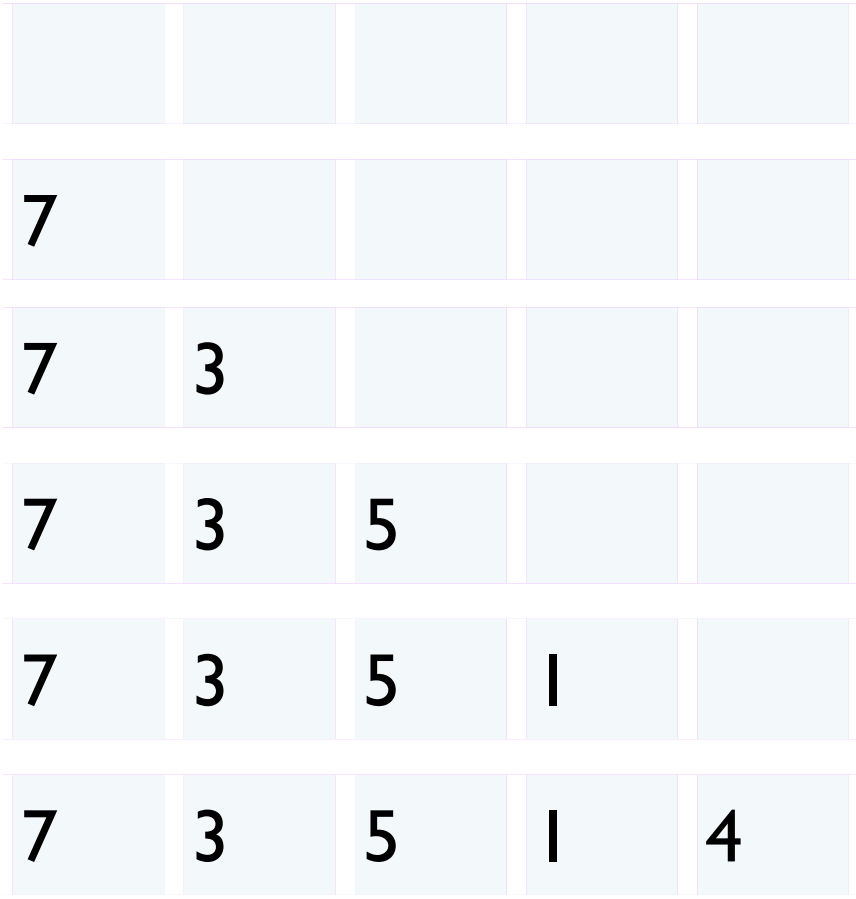
- [1] 104.7273
- The bigger the variability (the further points are from the mean) then the higher the sum of squares.

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

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

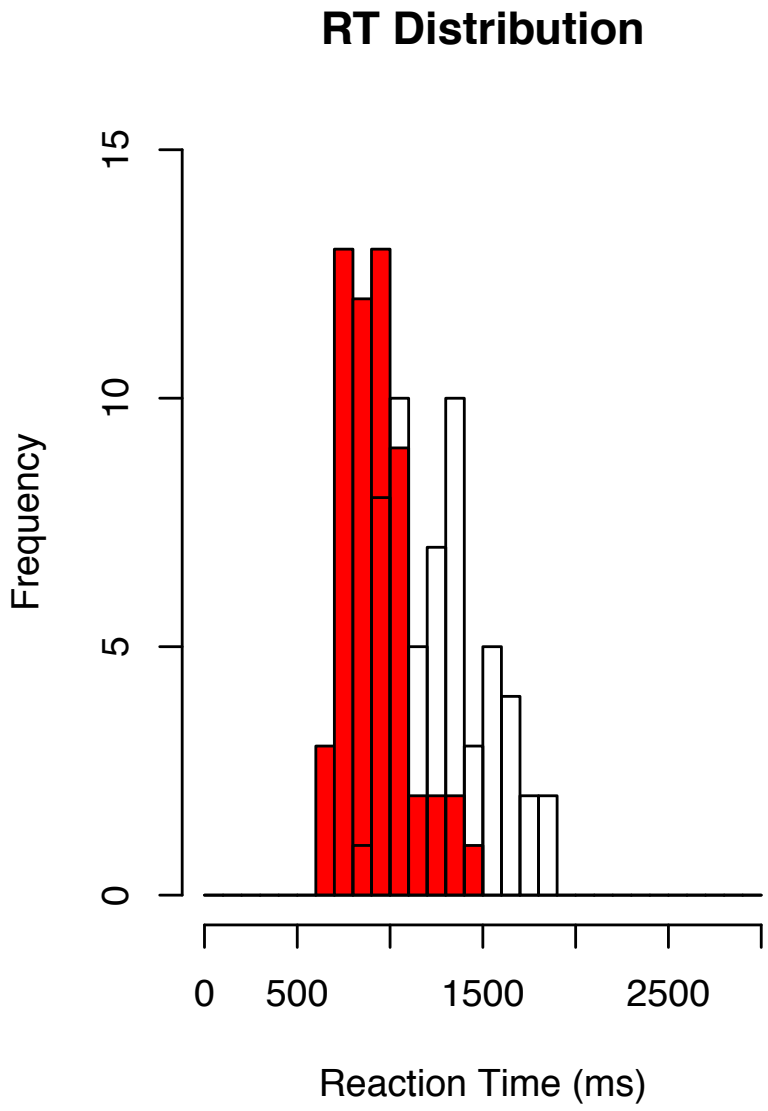
variance

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

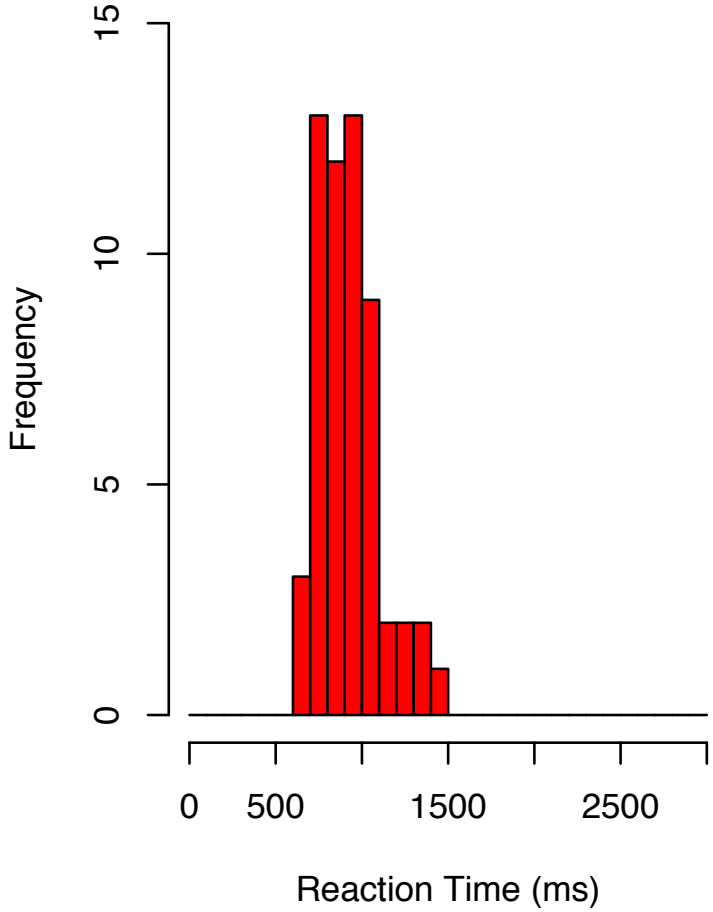
variance

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

variance of Stroop data



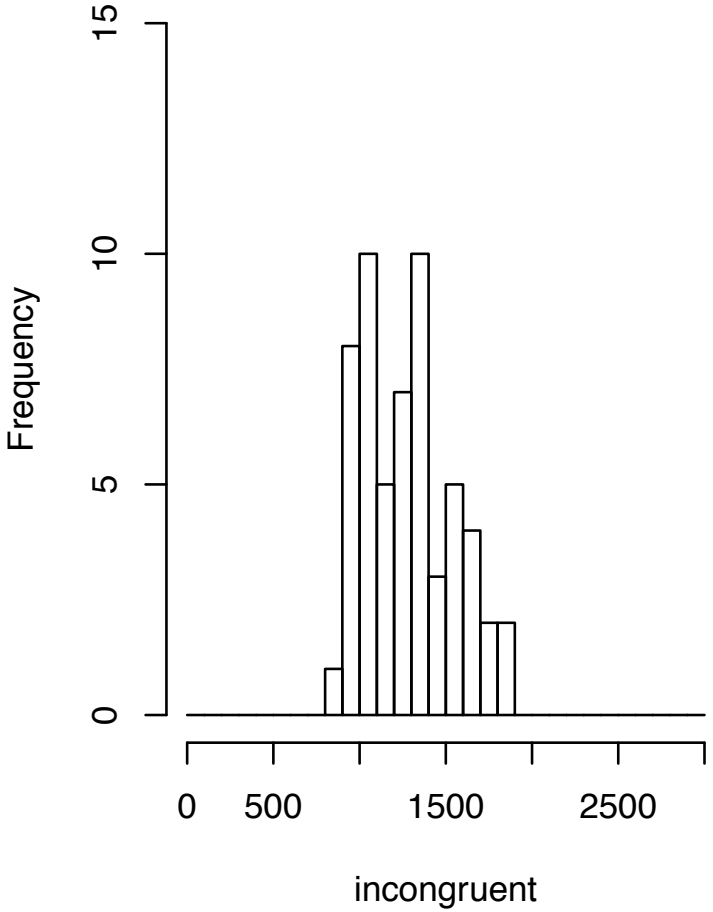
Congruent



mean = 918ms

variance = 32657ms^2

Incongruent



mean = 1278ms

variance = 66022ms^2

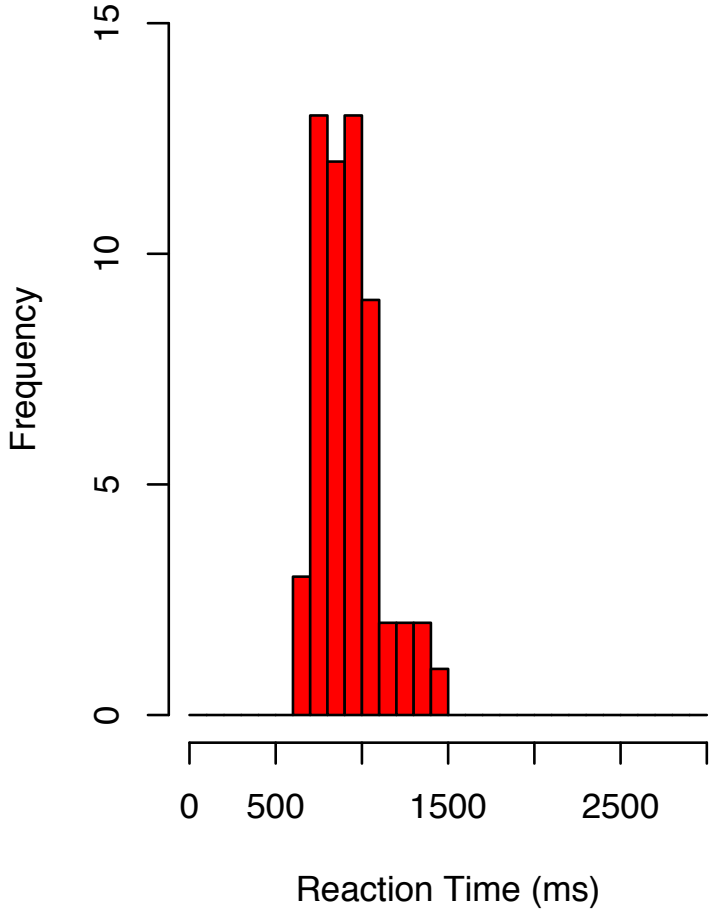
unequal variance

- 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.

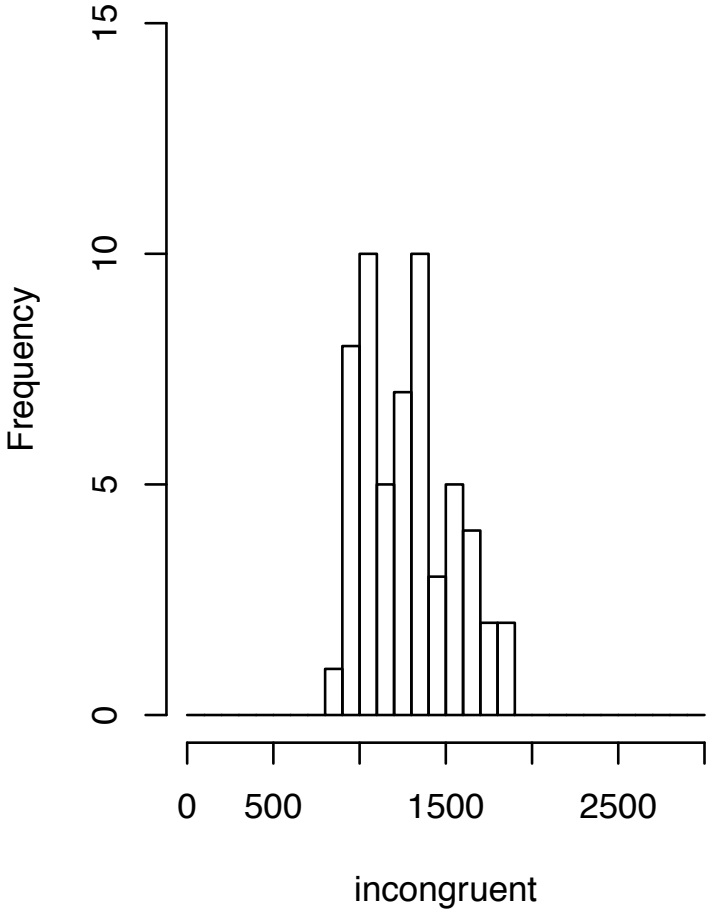
Congruent



mean = 918ms

standard deviation = 180.71ms

Incongruent



mean = 1278ms

standard deviation = 256.95ms

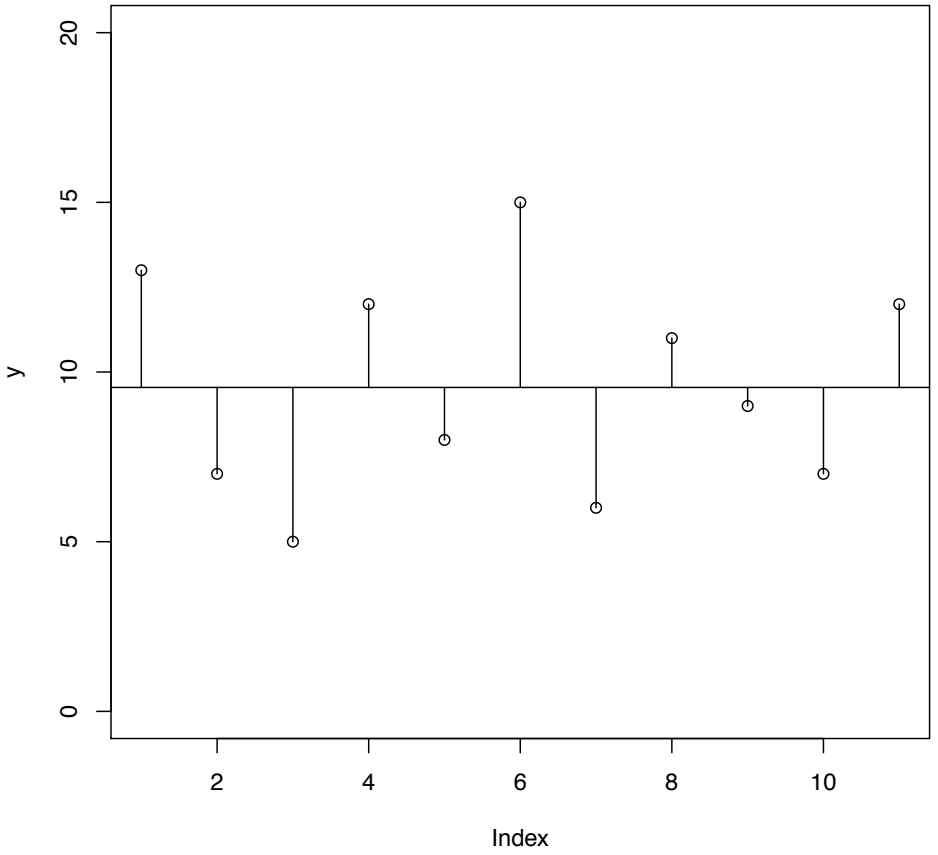
R

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
-