UNIVERSITY OF BIRMINGHAM

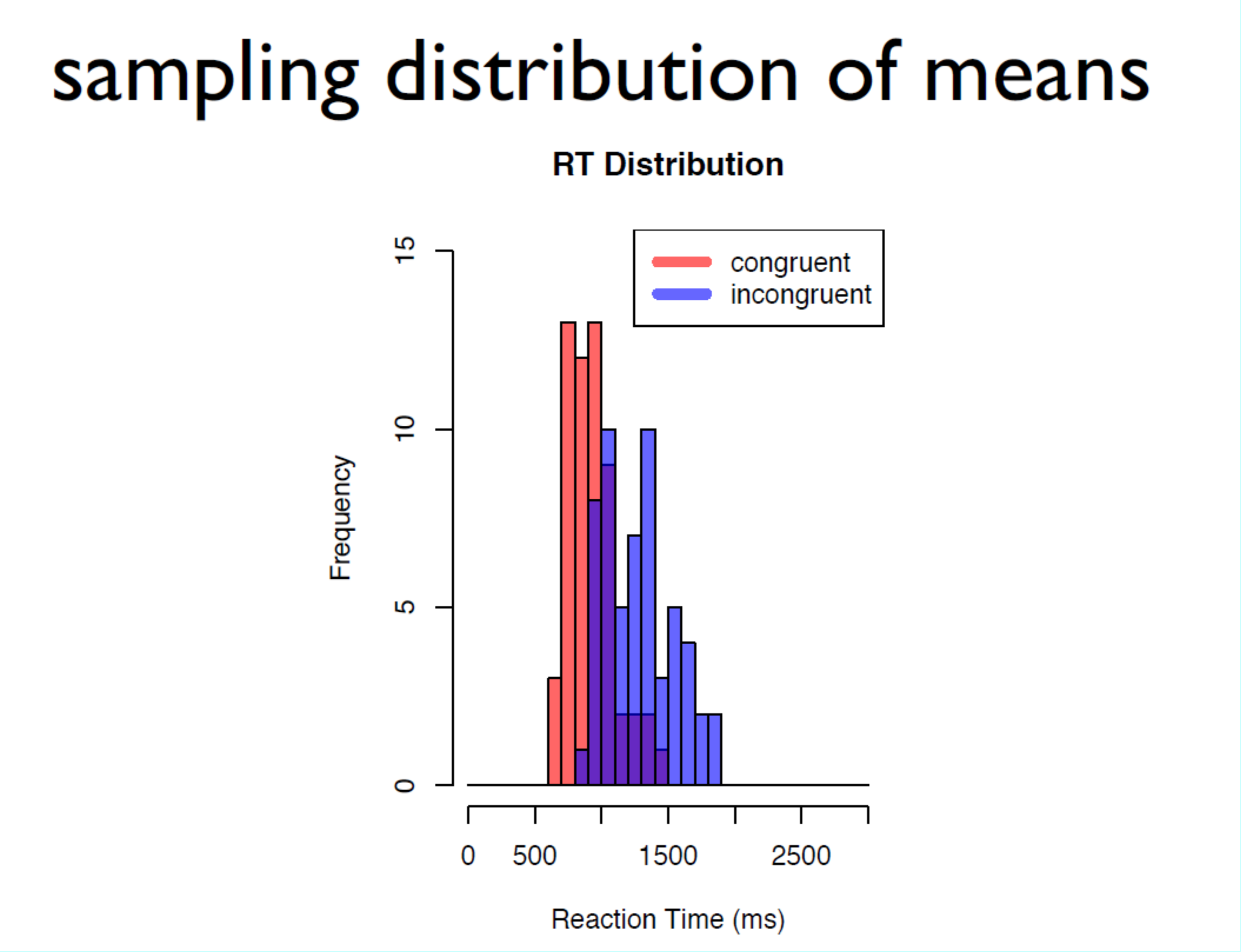
School of Computer Science

# EVALUATION METHODS AND STATISTICS

Prof. Chris Baber

Chair of Pervasive and Ubiquitous Computing

# Stroop Data (again): are the distributions different?

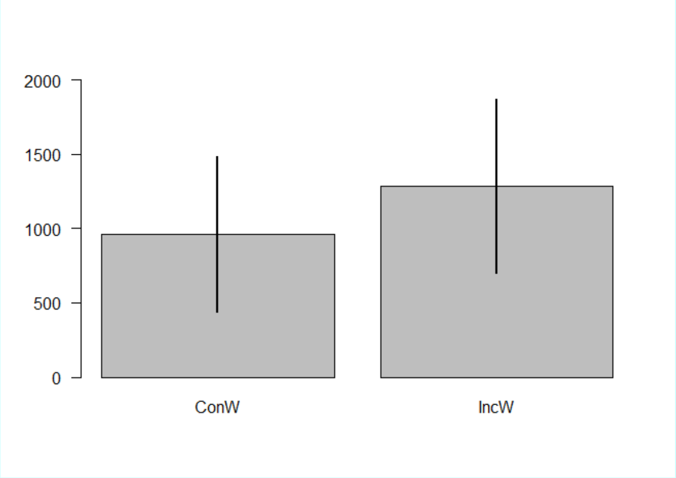



UNIVERSITY OF BIRMINGHAM

# Testing for Differences in Means

- If the sample mean difference is greater than we expect then
  - We have, by chance, collected samples that are not typical of the population

  Or
  - The two samples are from different populations




- We can compare the means to decide which explanation is most plausible

UNIVERSITY OF BIRMINGHAM

# t-distribution

- Assume that samples are drawn from a normal distribution
- A t-distribution is symmetrical but with more scores in the tails (leptokurtic)
- Developed by W.S. Gosset in 1908, working at the Guinness brewery in Dublin
- ‘Hypothesis test statistic’
- Published under pseudonym ‘student’
- A t-distribution is defined as a measure of central tendency divided by a measure of spread.

$$t = \frac{m - \mu}{\sqrt{\frac{s^2}{N}}}$$

$\mu = 0$, $m$ = mean of D

- There are numerous t-distributions, so we define the form required for a specific analysis by the degrees of freedom of a distribution.
- The areas under the normal distribution curve are calculated in terms of t rather than z values, e.g., 95% confidence interval is z = 1.96 but t = 2.045

UNIVERSITY OF BIRMINGHAM

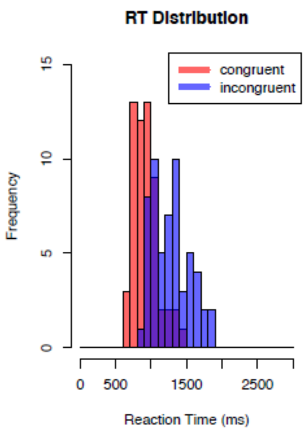
# Types of t-test

- Dependent means / Paired Samples
  - Each participant completes all conditions

- Independent Samples
  - Different participants are allocated to each condition

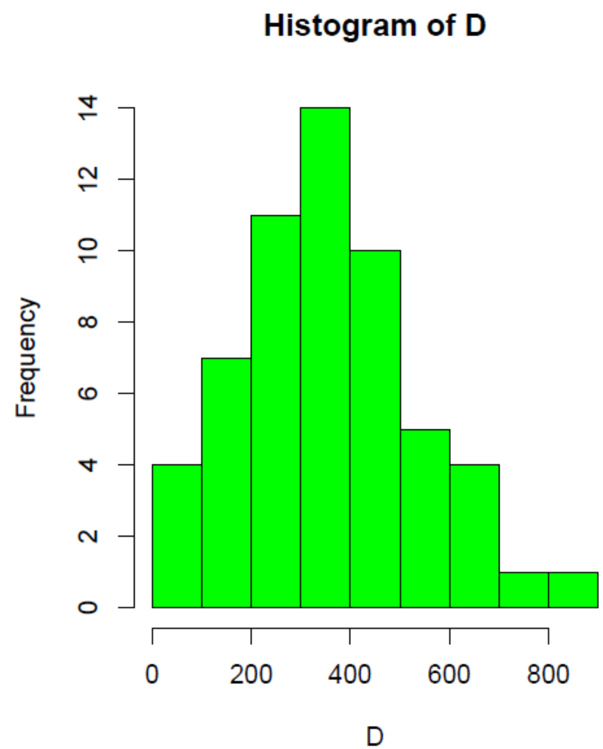
UNIVERSITY OF
BIRMINGHAM

# What is the differences (between the reaction time for the congruent and incongruent conditions)?




| | ConW | IncW | D |
|---|---|---|---|
| 1 | 806 | 1087 | 281 |
| 2 | 709 | 1099 | 390 |
| 3 | 935 | 1086 | 151 |
| 4 | 1004 | 1548 | 544 |
| 5 | 1225 | 1616 | 391 |
| 6 | 1391 | 1855 | 464 |
| 7 | 1051 | 1667 | 617 |
| 8 | 917 | 1231 | 314 |
| 9 | 1216 | 1306 | 90 |

...




UNIVERSITY OF BIRMINGHAM

# Central Tendency and Distribution

- We can assume that the histogram shows that difference, D, follows a normal distribution
- So, we can calculate central tendency as the mean
- If the null hypothesis is true, the mean = 0
- We don't (as yet) know the spread, i.e., variance, of the population.
- But, we assume that the population variance is the same as the sample variance.

# t-statistic (paired samples)

$$t = \frac{\left(\sum D\right)/N}{\sqrt{\dfrac{\sum D^2 - \left(\dfrac{\left(\sum D\right)^2}{N}\right)}{(N-1)(N)}}}$$

D = difference between trials;
N = number of samples

UNIVERSITY OF BIRMINGHAM

# Data Set (adapted from the source data to be small enough to work by hand)

| IncW | ConW |
|---|---|
| 1087 | 806 |
| 1099 | 709 |
| 1086 | 935 |
| 1548 | 1004 |
| 1616 | 1225 |
| 1855 | 1391 |
| 1667 | 1051 |
| 1231 | 917 |
| 1306 | 1216 |
| 896 | 1548 |
| 1472 | 840 |
| 1008 | 640 |
| 1023 | 752 |
| 1056 | 815 |
| 1040 | 800 |
| 1548 | 736 |

In this experiment, participants completed both conditions.
Each row represents one person’s mean reaction time for Incongruent words (different colour ink and word) and Congruent words (same colour ink and word)

UNIVERSITY OF BIRMINGHAM

Calculating a Paired Samples t-statistic by hand…

# 1. Calculate difference in time for each condition

| | IncW | ConW | Diff |
|---|---|---|---|
| | 1087 | 806 | |
| | 1099 | 709 | |
| | 1086 | 935 | |
| | 1548 | 1004 | |
| | 1616 | 1225 | |
| | 1855 | 1391 | |
| | 1667 | 1051 | |
| | 1231 | 917 | |
| | 1306 | 1216 | |
| | 896 | 1548 | |
| | 1472 | 840 | |
| | 1008 | 640 | |
| | 1023 | 752 | |
| | 1056 | 815 | |
| | 1040 | 800 | |
| | 1548 | 736 | |
| Sum(Σ) | | | |

UNIVERSITY OF BIRMINGHAM

Calculating a Paired Samples t-statistic by hand…

# 1. Calculate difference in time for each condition

| | IncW | ConW | Diff |
|---|---|---|---|
| | 1087 | 806 | 281 |
| | 1099 | 709 | 390 |
| | 1086 | 935 | 151 |
| | 1548 | 1004 | 544 |
| | 1616 | 1225 | 391 |
| | 1855 | 1391 | 464 |
| | 1667 | 1051 | 616 |
| | 1231 | 917 | 314 |
| | 1306 | 1216 | 90 |
| | 896 | 1548 | -652 |
| | 1472 | 840 | 632 |
| | 1008 | 640 | 368 |
| | 1023 | 752 | 271 |
| | 1056 | 815 | 241 |
| | 1040 | 800 | 240 |
| | 1548 | 736 | 812 |
| Sum(Σ) | | | |

UNIVERSITY OF BIRMINGHAM

Calculating a Paired Samples t-statistic by hand…

# 2. Sum differences and Square the product

| | IncW | ConW | Diff |
|---|---|---|---|
| | 1087 | 806 | 281 |
| | 1099 | 709 | 390 |
| | 1086 | 935 | 151 |
| | 1548 | 1004 | 544 |
| | 1616 | 1225 | 391 |
| | 1855 | 1391 | 464 |
| | 1667 | 1051 | 616 |
| | 1231 | 917 | 314 |
| | 1306 | 1216 | 90 |
| | 896 | 1548 | -652 |
| | 1472 | 840 | 632 |
| | 1008 | 640 | 368 |
| | 1023 | 752 | 271 |
| | 1056 | 815 | 241 |
| | 1040 | 800 | 240 |
| | 1548 | 736 | 812 |
| Sum(Σ) | | | 5153 |
| $(\Sigma D)^2$ | | | 26553409 |

UNIVERSITY OF BIRMINGHAM

Calculating a Paired Samples t-statistic by hand…

# 3. Square the differences

| | IncW | ConW | Diff | Diff^2 |
|---|---|---|---|---|
| | 1087 | 806 | 281 | |
| | 1099 | 709 | 390 | |
| | 1086 | 935 | 151 | |
| | 1548 | 1004 | 544 | |
| | 1616 | 1225 | 391 | |
| | 1855 | 1391 | 464 | |
| | 1667 | 1051 | 616 | |
| | 1231 | 917 | 314 | |
| | 1306 | 1216 | 90 | |
| | 896 | 1548 | -652 | |
| | 1472 | 840 | 632 | |
| | 1008 | 640 | 368 | |
| | 1023 | 752 | 271 | |
| | 1056 | 815 | 241 | |
| | 1040 | 800 | 240 | |
| | 1548 | 736 | 812 | |
| Sum(Σ) | | | 5153 | |
| $(\Sigma D)^2$ | | | 26553409 | |

UNIVERSITY OF BIRMINGHAM

Calculating a Paired Samples t-statistic by hand…

# 3. Square the differences

| | IncW | ConW | Diff | Diff^2 |
|---|---|---|---|---|
| | 1087 | 806 | 281 | 78961 |
| | 1099 | 709 | 390 | 152100 |
| | 1086 | 935 | 151 | 22801 |
| | 1548 | 1004 | 544 | 295936 |
| | 1616 | 1225 | 391 | 152881 |
| | 1855 | 1391 | 464 | 215296 |
| | 1667 | 1051 | 616 | 379456 |
| | 1231 | 917 | 314 | 98596 |
| | 1306 | 1216 | 90 | 8100 |
| | 896 | 1548 | -652 | 425104 |
| | 1472 | 840 | 632 | 399424 |
| | 1008 | 640 | 368 | 135424 |
| | 1023 | 752 | 271 | 73441 |
| | 1056 | 815 | 241 | 58081 |
| | 1040 | 800 | 240 | 57600 |
| | 1548 | 736 | 812 | 659344 |
| Sum(Σ) | | | 5153 | |
| $(\Sigma D)^2$ | | | 26553409 | |

UNIVERSITY OF BIRMINGHAM

Calculating a Paired Samples t-statistic by hand…

# 4. Sum the Squares

| | IncW | ConW | Diff | Diff^2 |
|---|---|---|---|---|
| | 1087 | 806 | 281 | 78961 |
| | 1099 | 709 | 390 | 152100 |
| | 1086 | 935 | 151 | 22801 |
| | 1548 | 1004 | 544 | 295936 |
| | 1616 | 1225 | 391 | 152881 |
| | 1855 | 1391 | 464 | 215296 |
| | 1667 | 1051 | 616 | 379456 |
| | 1231 | 917 | 314 | 98596 |
| | 1306 | 1216 | 90 | 8100 |
| | 896 | 1548 | -652 | 425104 |
| | 1472 | 840 | 632 | 399424 |
| | 1008 | 640 | 368 | 135424 |
| | 1023 | 752 | 271 | 73441 |
| | 1056 | 815 | 241 | 58081 |
| | 1040 | 800 | 240 | 57600 |
| | 1548 | 736 | 812 | 659344 |
| Sum($\Sigma$) | | | 5153 | |
| $(\Sigma D)^2$ | | | 26553409 | |

UNIVERSITY OF BIRMINGHAM

Calculating a Paired Samples t-statistic by hand…

# 4. Sum the Squares

| | IncW | ConW | Diff | Diff^2 |
|---|---|---|---|---|
| | 1087 | 806 | 281 | 78961 |
| | 1099 | 709 | 390 | 152100 |
| | 1086 | 935 | 151 | 22801 |
| | 1548 | 1004 | 544 | 295936 |
| | 1616 | 1225 | 391 | 152881 |
| | 1855 | 1391 | 464 | 215296 |
| | 1667 | 1051 | 616 | 379456 |
| | 1231 | 917 | 314 | 98596 |
| | 1306 | 1216 | 90 | 8100 |
| | 896 | 1548 | -652 | 425104 |
| | 1472 | 840 | 632 | 399424 |
| | 1008 | 640 | 368 | 135424 |
| | 1023 | 752 | 271 | 73441 |
| | 1056 | 815 | 241 | 58081 |
| | 1040 | 800 | 240 | 57600 |
| | 1548 | 736 | 812 | 659344 |
| Sum(Σ) | | | 5153 | 3212545 |
| $(\Sigma D)^2$ | | | 26553409 | |

UNIVERSITY OF BIRMINGHAM

Calculating a Paired Samples t-statistic by hand…

# 4. Calculate t-statistic (paired samples)

$$t = \frac{\left(\sum D\right)/N}{\sqrt{\dfrac{\sum D^2 - \left(\dfrac{\left(\sum D\right)^2}{N}\right)}{(N-1)(N)}}}$$

UNIVERSITY OF BIRMINGHAM

# 4. Calculate t-statistic

$$t = \frac{(\sum D)/N}{\sqrt{\frac{\sum D^2 - \left(\frac{(\sum D)^2}{N}\right)}{(N-1)(N)}}}$$

$$t = \frac{5153/16}{\sqrt{\frac{3212545 - \left(\frac{26553409}{16}\right)}{(16-1)(16)}}}$$

$t = 4.00741909$

UNIVERSITY OF BIRMINGHAM

# 5. Use t-table to check significance

- For 95% confidence interval, $\alpha = 0.05$
- D.F. (degrees of freedom) = N-1 = 15
- From the t-table, t should be >2.131
- The calculated t = 4.007, is larger than the value in the table
- Therefore, we reject the null hypothesis and state that the difference between the mean reaction times in these data is less than 5% likely to occur by chance.
- We write the result as **$t(15) = 4.007$, $p<0.05$**

## Two Tails T Distribution Table

| DF | A = 0.2 | 0.10 | 0.05 | 0.02 | 0.01 | 0.002 | 0.001 |
|---|---|---|---|---|---|---|---|
| ∞ | $t_a$ = 1.282 | 1.645 | 1.960 | 2.326 | 2.576 | 3.091 | 3.291 |
| 1 | 3.078 | 6.314 | 12.706 | 31.821 | 63.656 | 318.289 | 636.578 |
| 2 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 | 22.328 | 31.600 |
| 3 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 | 10.214 | 12.924 |
| 4 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 | 7.173 | 8.610 |
| 5 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 | 5.894 | 6.869 |
| 6 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 | 5.208 | 5.959 |
| 7 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 | 4.785 | 5.408 |
| 8 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 | 4.501 | 5.041 |
| 9 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 | 4.297 | 4.781 |
| 10 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 | 4.144 | 4.587 |
| 11 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 | 4.025 | 4.437 |
| 12 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 | 3.930 | 4.318 |
| 13 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 | 3.852 | 4.221 |
| 14 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 | 3.787 | 4.140 |
| 15 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 | 3.733 | 4.073 |
| 16 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 | 3.686 | 4.015 |
| 17 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 | 3.646 | 3.965 |
| 18 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 | 3.610 | 3.922 |
| 19 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 | 3.579 | 3.883 |
| 20 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 | 3.552 | 3.850 |

# Is this is a ‘good’ result?

- We consider the generalisability of the result by its Effect Size
- Effect size indicates the size of the difference between samples independent of sample size
  - This could relate to proportion of explained variance, e.g., using $R^2$ or (partial) eta squared
  - Or could relate to difference in averages, e.g., using Cohen’s d

| | **Small** | **Medium** | **Large** | **Very large** |
|---|---|---|---|---|
| $R^2$ | 1% | 9% | 25% | |
| Partial eta$^2$ | 0.02 | 0.13 | 0.26 | |
| d | 0.2 | 0.5 | 0.8 | >1 |

Here is a nice paper about Effect Size:
https://www.leeds.ac.uk/educol/documents/00002182.htm

Here is an online calculator for Effect size:
https://www.ai-therapy.com/psychology-statistics/effect-size-calculator

UNIVERSITY OF BIRMINGHAM

# Power and Effect Size

- Power is the probability that the study produces significant difference between groups when a difference actually exits. Power is 1-$\beta$, and relates to Type II error.
- Power relies on sample size and effect size.
- Effect size is the magnitude of difference between groups.

UNIVERSITY OF BIRMINGHAM

# Paired t-test effect size

- There are several formulae, I prefer Cohen's *d*…
  - For a paired t-test…

$$d = \frac{|m_1 - m_2|}{\sqrt{{s_1}^2 + {s_2}^2 - (2rs_1s_2)}}$$

where r is the correlation coefficient between groups

For this dataset, r = 0.3277 (we'll calculate this in a later lecture)

So, the Effect size for this result is 1.001

(meaning that the difference is larger than 1 standard deviation)

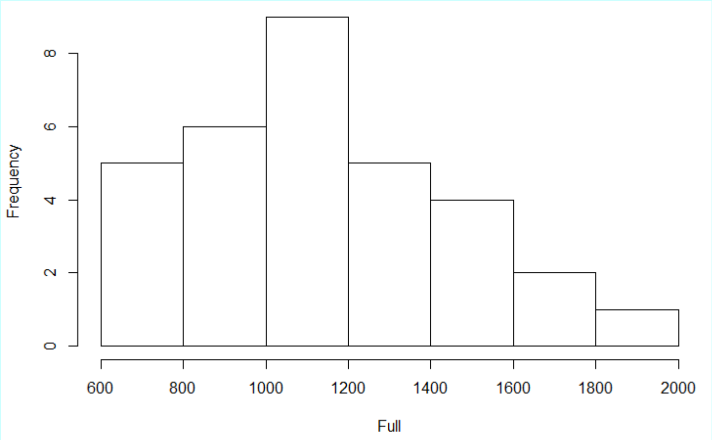
Here is an online calculator for Effect size:
https://www.ai-therapy.com/psychology-statistics/effect-size-calculator

UNIVERSITY OF BIRMINGHAM

# Explore the data set in R

- Copy the data on slide 8 into a spreadsheet, such as Excel and save as .txt or .tsv
- In R user interface: File – Import dataset – From text…

```
>Attach (stroop)
>stroop <-data.frame(IncW, ConW)
>hist(IncW)
>hist(ConW)
> t.test(Incw, ConW, paired=TRUE, conf.level=0.95)
```




Paired t-test data: IncW and ConW (assuming null hypothesis, $\mu = 0$)

t = 4.0037, df = 15, p-value = 0.001151

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

150.6079 493.5171

sample estimates:

mean of the differences

322.0625

UNIVERSITY OF BIRMINGHAM

# Calculating Power in R

- Note…Different tests calculate Power in different ways…t-tests use

  Pwr.t.test (n =  , d =  , sig.level =  , power =  , type = c(“two.sample”, “one.sample”, “paired”))

- For Paired t-test, we take the mean for each group…

  > mean(IncW)

  [1] 1293.625

  > mean(ConW)

  [1] 961.562

- …and we need to hypothesise a difference to detect (let’s say 200) and the power that we wish to use (let’s say, .8)

```
> library(pwr)
> pwr.t.test(d=(1283.625-961.5625)/200, power=0.8,sig.level=0.05, type ="paired",alternative="two.sided")
```

```
Paired t test power calculation
n = 5.242336
d = 1.610313
sig.level = 0.05
power = 0.8
alternative = two.sided
NOTE: n is number of *pairs*
```

UNIVERSITY OF BIRMINGHAM

```
pwr.t.test(d=(1283.625-961.5625)/200, power=0.8,sig.lieve=0.05,type="paired", alternative="two.sided")
```

# We can explore this further…

- When we set the difference to 500, Cohen's d = 0.64, n = 20.9 (or 21)
- If we see beta to .95 (and keep difference at 200), d = 1.61, n = 7.2 (7)
- And beta .95, d = 500, d= 0.64, n = 33.3 (33)

- Where n = participants in paired (repeated measures) design

UNIVERSITY OF BIRMINGHAM

# Independent t-test

- Participants are assigned to *only* one trial
- If you have the same number of participants per trial (equal sample sizes):

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\left(\frac{{s_1}^2}{N_1} + \frac{{s_2}^2}{N_2}\right)}}$$

UNIVERSITY OF BIRMINGHAM

# Cohen's d for independent samples

$$d = \frac{|m_1 - m_2|}{pooled\ variance}$$

$$pooled\ variance = \sqrt{\left(\frac{(s_1{}^2 + \ s_2{}^2)}{2}\right)}$$

- Why you don't use this for paired samples
  - The effect size assumes a differences between sets of independent scores
  - The paired t-test corrects for correlation, r, into account to correct for pooled variance

UNIVERSITY OF BIRMINGHAM

# Calculating an Odds Ratio

- An odds ratio indicates the strength of effect for one condition relative to another
- We are calculating the probability of a result given condition and contrasting this with a the probability of a contrasting result
- The ‘odds’ are the likelihood an event, or p(e) / 1-p(e)
- So, we need to describe our experiment in a 2 x 2 matrix
- Odd Ratio = (a*d) / (b*c)

| DV | IV1 | IV2 |
|---|---|---|
| i | A | B |
| ii | C | D |

UNIVERSITY OF BIRMINGHAM

# A Cautionary tale – or completely made-up example

- In a class of 60 students, 37 regularly attended lectures and 23 did not
- Of the regular attenders: 32 passed the exam and 5 failed
- Of the non-attenders, 6 passed the exam and 17 failed
- Odds ratio = (17*32) / (5*6) = 18.13
- So, 18.13 more likely to pass exams if you attend lectures

| | **No lectures** | **Lectures** | |
|---|---|---|---|
| fail | 17 | 5 | 22 |
| pass | 6 | 32 | 38 |
| | 23 | 37 | |

UNIVERSITY OF BIRMINGHAM