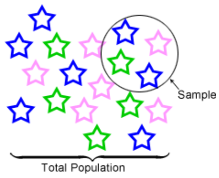
Introduction to Statistical Analysis
(using R Commander)

Sarah Vowler and **Mark Dunning**

9.30-10:30	- Lecture: introduction to stats analysis, tests for continuous variables
10:30-11:50	- Examples & Practicals
11:50-12:05	- Lecture: tests for categorical variables
12:05-12:50	- Examples & Practicals
12:50-13:00	- Summary

The point of statistics

- Rarely feasible to study the whole population that we are interested in, so we take a sample instead
- Assume that data collected represents a larger population
- Use sample data to make conclusions about the overall population

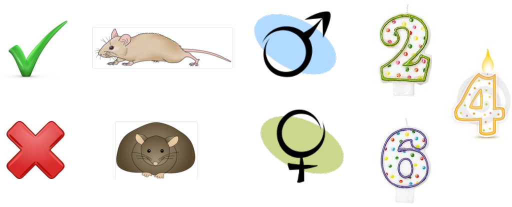


Data

- Type?
  - Categorical (nominal) , e.g. Gender
  - Categorical with ordering (ordinal), e.g. Tumour grade
  - Discrete, e.g. Shoe size
  - Continuous, e.g. Body weight in kg
- **Independent** or **dependent** measurements
- Representative of which population?
- Distribution
  - Normally distributed? Skewed? Bimodal?

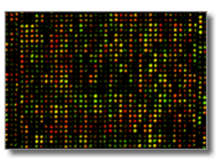
Data type – examples

- Success/ failure of achieving a task for a mouse which may be wild-type or knock-out, male or female, 2, 4 or 6 months old.



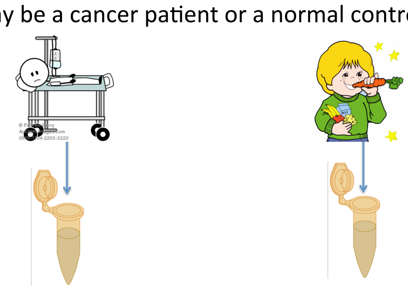
Data type – examples

- Gene expression in each cell sample which may be one of five cell-types (A, B, C, D, E)



Data type – examples

- The number of bacteria for each subject which may be a cancer patient or a normal control

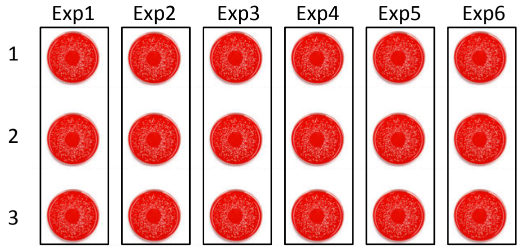


Measurements: Dependent / Independent?

- Measurements of gene expression taken from each of 20 individuals
- Are any measurements more closely related than others?
  - Siblings/littermates?
  - Same individual measured twice?
  - Batch effects?
- If no reason – **independent observations**

Measurements: Dependent / Independent?

- 18 measurement: from repeating an experiment 6 times, each time in triplicate



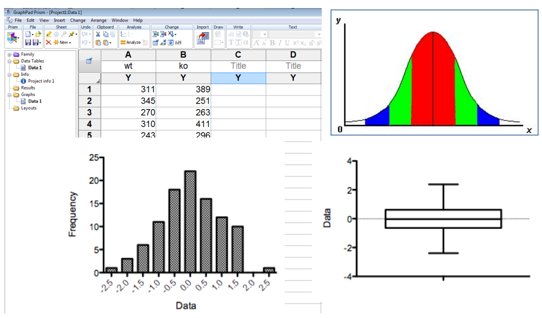
Measurements: Dependent / Independent?

- Measuring blood pressure before and after treatment for 30 patients

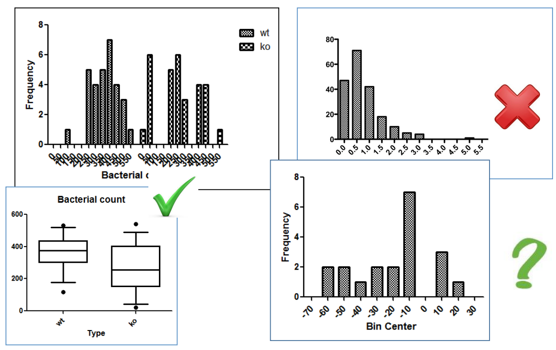
Measurements: Dependent / Independent?

- Measuring gene expression in each cell sample which may be one of five cell-types from cancer patients and normal subjects

Continuous Data – Distribution

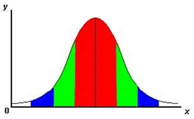


Continuous Data – Distribution?



Continuous Data – Descriptive Statistics

- Measures of location and spread

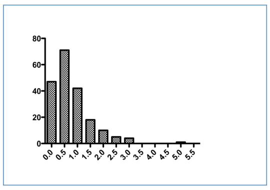


  - Mean and standard deviation

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

$$s.d. = \sqrt{\frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \ldots + (X_n - \bar{X})^2}{n}}$$

Continuous Data – Descriptive Statistics



- Median: middle value
- Lower quartile: median bottom half of data
- Upper quartile: median top half of data

Continuous Data – Descriptive Statistics (Example)

E.g. No. of Facebook friends for 7 colleagues

311, 345, 270, 310, 243, 5300, 11

- Measures of location and spread
  - Mean and standard deviation

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = 970;$$

$$s.d. = \sqrt{\frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \ldots + (X_n - \bar{X})^2}{n}} = 1912.57$$

  - Median and interquartile range

    11, **243**, 270, **310**, 311, **345**, 5300

Continuous Data – Descriptive Statistics (Example)

E.g. No. of facebook friends for 7 colleagues

311, 345, 270, 310, 243, **530**, 11

- Measures of location and spread
  - Mean and standard deviation

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = 289;$$

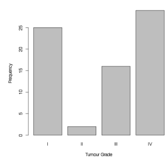
$$s.d. = \sqrt{\frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \ldots + (X_n - \bar{X})^2}{n}} = 153.79$$

  - Median and interquartile range

    11, **243**, 270, **310**, 311, **345**, 530

Categorical Data

- Summarised by counts and percentages
- Examples
  - 19/82 (23%) subjects had Grade IV tumour
  - 48/82 (58%) subjects had Diarrhoea as an Adverse Event.



Standard Deviation and Standard Error

- Commonly confused
- Standard deviation:
  - Measure of spread of the data
  - Used for describing population
- Standard error:
  - Variability of the mean from repeated sampling
  - Precision of mean
  - Used to calculate confidence interval
- SD: How widely scattered measurements are
- SE: Uncertainty in estimate of sample mean

Confidence intervals for the mean

- Confidence interval (CI) is a random interval
- In repeated experiments
  - 95% of time cover the mean
- Looser interpretation 95% of time mean in CI

$$95\%\ CI : \left(\bar{X} - 1.96 \times \text{standard error}, \bar{X} + 1.96 \times \text{standard error}\right)$$

$$\text{Standard error} = \frac{\text{Standard deviation}}{\sqrt{n}} = \frac{154}{\sqrt{7}} = 58$$

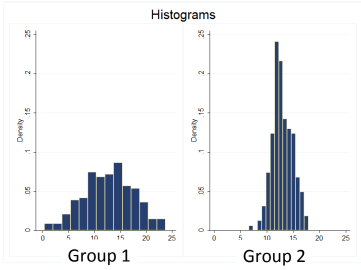
Mean 289, 95% CI (175, 402)

Confidence intervals

No. of samples/ observations

Standard deviation

Standard error of mean



Hypothesis tests – basic set-up

- Formulate a null hypothesis, H_0

The difference in gene expression before and after treatment = 0

- Calculate a test statistic from the data under the null hypothesis $t_{n-1} = t_{29} = \frac{\bar{X}_{After-Before}}{s.e.\left(\bar{X}_{After-Before}\right)}$
- Determine whether the test statistic is more extreme than expected under the null hypothesis (p-value)
- Reject or do not reject the null hypothesis

Absence of evidence is not evidence of absence (Bland and Altman, 1995)

- Correction for multiple testing

Hypothesis tests – Example

Lady Tasting Tea

Randomised Experiment by Fisher

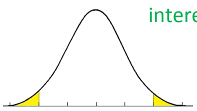
- Randomly ordered 8 cups of tea
  - 4 were prepared by first adding milk
  - 4 were prepared by first adding tea
- Task: Lady had to select the 4 cups of one particular method
- H_0: Lady had no such ability
- Test Statistic: number of successes in selecting the 4 cups.
- Result: Lady got all 4 cups right!

Reject the null hypothesis

Hypothesis tests – Errors

	Null hypothesis does not hold	Null hypothesis holds
Reject null hypothesis	Correct True positive	Wrong False positive
Do not reject null hypothesis	Wrong False negative	Correct True negative

significance level, sample size, difference of interest, variability of the observations.



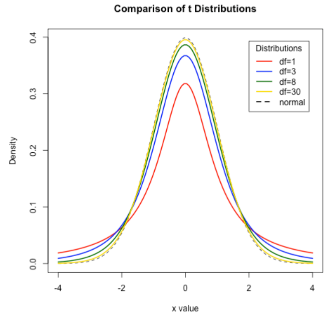
Be aware of issues of multiple testing!

Tests for continuous variables
T-tests

Statistical tests – continuous variables

- T-test:
  - **One-sample t-test**
    (e.g. H_0: mean = 5)
  - **Independent two-sample t-test**
    (e.g. H_0: mean of sample 1 = mean of sample 2)
  - **Paired two-sample t-test**
    (e.g. H_0: mean difference between pairs = 0)

T-distributions



One-sample t-test: does mean = X?

E.g. Research question: Published data suggests that the microarray failure rate for a particular supplier is 2.1%.

Genomics Core want to know if this holds true in their own lab?

One-sample t-test: does mean = X?

- **Null hypothesis, H_0** :
  Mean monthly failure rate = 2.1%.
- **Alternative hypothesis, H_1** :
  Mean monthly failure rate ≠ 2.1%.
- **Tails**: two-tailed.
- Either reject or do not reject the **null hypothesis** – <u>never accept the alternative hypothesis</u>

One-sample t-test – the data

Month	Monthly failure rate
January	2.90
February	2.99
March	2.48
April	1.48
May	2.71
June	4.17
July	3.74
August	3.04
September	1.23
October	2.72
November	3.23
December	3.40

The **mean** is the sum of all observations divided by the number of observations.

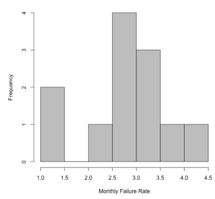
Mean = (2.90 +...+ 3.40)/12
= 2.84

Standard deviation = 0.84

Test value: 2.1

One-sample t-test – key assumptions

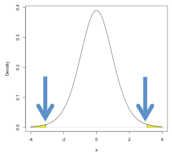
- Observations are independent
- Observations are normally distributed



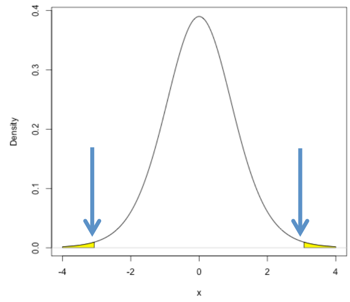
One-sample t-test - results

Test statistic:

$$t_{n-1} = t_{11} = \frac{\bar{x} - \mu_0}{s.d./\sqrt{n}} = \frac{2.84 - 2.10}{s.e.(\bar{x})} = 3.07$$



One-sample t-test - results



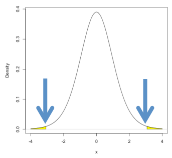
One-sample t-test - results

Test statistic:

$$t_{n-1} = t_{11} = \frac{\bar{x} - \mu_0}{s.d./\sqrt{n}} = \frac{2.84 - 2.10}{s.e.(\bar{x})} = 3.07$$

df = 11

P = 0.01



Reject H_0

(Evidence that mean monthly failure rate ≠ 2.1%.)

One-sample t-test results

- The mean monthly failure rate of microarrays in the Genomics core is 2.84 (95% CI: 2.30, 3.37).
- It is not equal to the hypothesized mean proposed by the company of 2.1.
- t=3.07, df=11, p=0.01

Two-sample t-test

- Two types of two-sample t-test:
  - Independent:

    e.g. the weight of two different breeds of mice.
  - Paired:

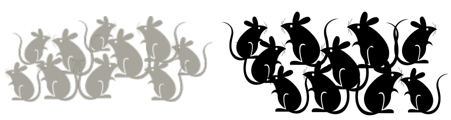
    e.g. a measurement of disease at two different parts of the body in the same patient/animal.

Independent two-sample t-test

Does mean of group A = mean of group B?

E.g. Research question: 40 male mice (20 of breed A and 20 of breed B) were weighed at 4 weeks old.

Does the weight of 4 week old male mice depend on breed?



Independent two-sample t-test
Does mean of group A = mean of group B?

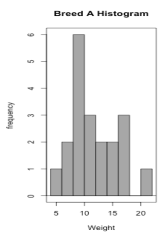
- **Null hypothesis, H_0** :
  Mean weight of breed A = Mean weight of breed B.
- **Alternative hypothesis, H_1** :
  Mean weight of breed A ≠ Mean weight of breed B.
- **Tails**: two-tailed.
- Either reject or do not reject the **null hypothesis** – never accept the alternative hypothesis

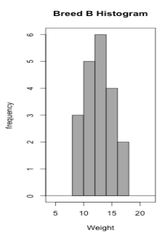
Independent two-sample t-test – the data

Breed A		Breed B	
Subject	**Weight at 4 weeks (g)**	**Subject**	**Weight at 4 weeks (g)**
1	20.77	21	15.51
2	9.08	22	12.93
3	9.80	23	11.50
4	8.13	24	16.07
5	16.54	25	15.51
6	11.36	26	17.66
7	11.47	27	11.25
8	12.10	28	13.65
9	14.04	29	14.28
10	16.82	30	13.21
11	6.32	31	10.28
12	17.51	32	12.41
13	9.87	33	9.63
14	12.41	34	14.75
15	7.39	35	9.81
16	9.23	36	13.02
17	4.06	37	12.33
18	8.26	38	11.90
19	10.24	39	8.98
20	14.64	40	11.29
Mean	11.50	**Mean**	12.80
Standard deviation	4.18	**Standard deviation**	2.33

Independent two-sample t-test – key assumptions

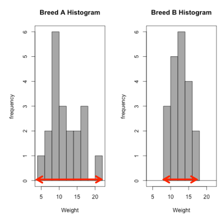
- Observations are independent
- Observations are normally distributed





Independent two-sample t-test -More key assumptions...

- Equal variance in the two comparison groups
  - Use Welch's correction if variances are different
    - » Alters the t-value and degrees of freedom



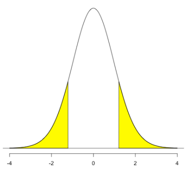
Standard deviation 4.18 2.33

Independent two-sample t-test - results

Test statistic: $$t_{df} = \frac{\bar{X}_A - \bar{X}_B}{s.e.\left(\bar{X}_A - \bar{X}_B\right)} = 1.21$$

df = 29.78
(Welch's correction)

P-value: **0.24**



Do not reject H_0
(No evidence that mean weight of breed A ≠ mean weight of breed B)

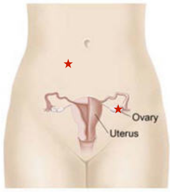
Independent two-sample t-test - results

- The difference in mean weight between the two breeds is -1.30 (95% CI: -3.48, 0.89)
  - [NB this is negative breed B weights tend to be bigger than breed A weights].
- There is no evidence of a difference in weights between breed A and breed B.
- t=1.21, df= 29.78 (Welch's correction), p=0.24.

Paired two-sample t-test: Does the mean difference = 0?

E.g. Research question: 20 patients with ovarian cancer were studied using MRI imaging. Cellularity was measured for each patient at two sites of disease.

Does the cellularity differ between two different sites of disease?



Paired two-sample t-test: Does the mean difference = 0?

- **Null hypothesis, H_0 :**
  Cellularity at site A = Cellularity at site B
- **Alternative hypothesis, H_1 :**
  Cellularity at site A ≠ Cellularity at site B
- **Tails**: two-tailed.
- Either reject or do not reject the **null hypothesis** – never accept the alternative hypothesis

Paired two-sample t-test – Null hypothesis

H_0 : Cellularity at site A = Cellularity at site B

OR

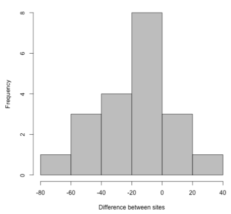
H_0 : Cellularity at site A - Cellularity at site B = 0

H_0 : Cellularity at site A - Cellularity at site B = 0

Subject	Cellularity		
	Site A: Primary ovarian mass	Site B: Peritoneal deposits	Difference
1	1201.33	1155.98	-45.35
2	1029.64	1020.82	-8.82
3	895.57	881.21	-14.37
4	842.14	830.78	-11.36
5	903.07	897.06	-6.01
6	1311.57	1262.73	-48.84
7	833.52	823.06	-10.46
8	1007.66	951.01	-56.65
9	1465.51	1450.98	-14.53
10	967.82	978.15	10.33
11	812.72	778.26	-34.46
12	884.08	823.57	-60.51
13	1358.56	1335.78	-22.78
14	1280.10	1293.91	13.80
15	942.38	925.75	-16.63
16	884.33	891.34	7.01
17	930.09	892.02	-38.07
18	1146.75	1132.80	-13.95
19	881.50	847.78	-33.72
20	1315.22	1337.80	22.58
		Mean difference	19.14
		Standard deviation	23.37

Paired two-sample t-test – key assumptions

- Observations are independent
- The **paired differences** are normally distributed

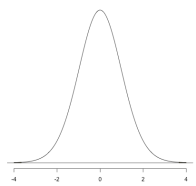


Paired two-sample t-test - results

Test statistic $$t_{n-1} = t_{19} = \frac{\overline{X_{A-B}}}{s.e.\left(\overline{X_{A-B}}\right)} = 3.66$$

df = 19

P-value: **0.002**



Reject H_0
(Evidence that cellularity at site A ≠ Cellularity at site B)

Paired two-sample t-test - results

- The difference in cellularity between the two sites is 19.14 (95% CI: 8.20, 30.08).
- There is evidence of a difference in cellularity between the two sites.
- t=3.66, df=19, p=0.0017.

What if normality is not reasonable?

- Transform your data, e.g. Ln transformation
- Non-parametric tests:

Parametric test	Non-parametric test
One-sample t-test	One-sample Wilcoxon signed rank test
Independent two-sample t-test	Mann-Whitney U test/ Wilcoxon rank sum test
Paired two-sample t-test	Matched-pairs Wilcoxon signed rank test

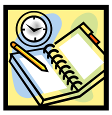
Summary – continuous variables

- **One-sample t-test**
  Use when we have one group.
- **Independent two-sample t-test**
  Use when we have two independent groups. A Welch correction may be needed if the two groups have different spread.
- **Paired two-sample t-test**
  Use when we have two non-independent groups.
- **Non-parametric tests or transformations**
  Use when we cannot assume normality.

Summary – t-test

- Turn scientific question to null and alternative hypothesis
- Think about test assumptions
- Calculate summary statistics
- Carry out t-test if appropriate

T-tests practical



- Work through examples on manual pages 18 - 36
- Complete the t-test practical
- We will start the next lecture at 11:30pm
- Feel free to take a short break if you want to

Tests for categorical variables

Associations between categorical variables

- All about frequencies!
- Row x Column table (2 x 2 simplest)
- Categorical data

Treatment group	Tumour shrinkage	
	No	Yes
Treatment	44	40
Placebo	24	16

2 x 2

- Look for association (relationship) between row variable and column variable

Chi-square test

- **E.g. Research question:** A trial to assess the effectiveness of a new treatment versus a placebo in reducing tumour size in patients with ovarian cancer.

Treatment group	Tumour shrinkage	
	No	Yes
Treatment	44	40
Placebo	24	16

- Is there an association between treatment group and tumour shrinkage?
- **Null hypothesis, H_0 :** No association
- **Alternative hypothesis, H_1 :** Some association

Chi-square test

Calculating expected frequencies:

Treatment group	Tumour shrinkage		Total
	No	Yes	
Treatment	44 46.1	40 37.9	84
Placebo	24 21.9	16 18.1	40
Total	68	56	124

$$E = \frac{\text{row total} \times \text{col total}}{\text{overall total}}$$

e.g. $\frac{84}{124} \times \frac{68}{124} \times 124 = \frac{84 \times 68}{124} = 46.1$

Chi-square test

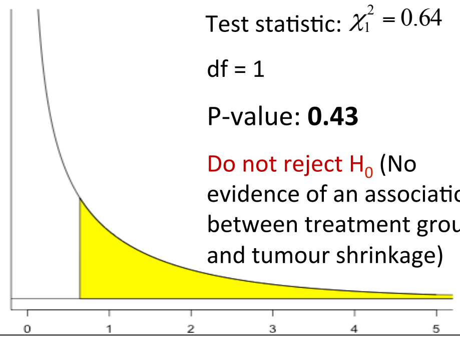
Calculating the chi-square statistic:

Treatment group	Tumour shrinkage		Total
	No	Yes	
Treatment	44 46.1	40 37.9	84
Placebo	24 21.9	16 18.1	40
Total	68	56	124

$$\chi^2_{(r-1)\times(c-1)} = \sum \frac{(O-E)^2}{E}$$

$$\chi^2_{(r-1)\times(c-1)} = \sum \frac{(O-E)^2}{E} = \frac{(44-46.1)^2}{46.1} + \frac{(40-37.9)^2}{37.9} + \frac{(24-21.9)^2}{21.9} + \frac{(16-18.1)^2}{18.1} = 0.64$$

Chi-square test



Test statistic: $\chi^2_1 = 0.64$

df = 1

P-value: **0.43**

Do not reject H_0 (No evidence of an association between treatment group and tumour shrinkage)

Limitations of the Chi-square test

- In general, a Chi-square test is appropriate when:
  - ➢ at least 80% of the cells have an <u>expected</u> frequency of 5 or greater
  - ➢ none of the cells have an <u>expected</u> frequency less than 1
- If these conditions aren't met, <u>Fisher's exact test</u> should be used.

Same question, smaller sample size

- **E.g. Research question**: Is there an association between treatment group and tumour shrinkage?

Treatment group	Tumour shrinkage		Total
	No	Yes	
Treatment	8	3	11
Placebo	9	4	13
Total	17	7	24

- **Null hypothesis, H_0** : No association
- **Alternative hypothesis, H_1 :** Some association

Expected frequencies

$$E = \frac{\text{row total} \times \text{col total}}{\text{overall total}}$$

Treatment group	Tumour shrinkage		Total
	No	Yes	
Treatment	8 7.8	3 3.2	11
Placebo	9 9.2	4 3.8	13
Total	17	7	24

Expected frequency less than 5

Only 50% of cells have an expected frequency greater than 5 → use Fisher's exact test

e.g. $\frac{11}{24} \times \frac{17}{24} \times 24 = \frac{11 \times 17}{24} = 7.8$

Fisher's exact test - results

Treatment group	Tumour shrinkage		Total
	No	Yes	
Treatment	8 7.8	3 3.2	11
Placebo	9 9.2	4 3.8	13
Total	17	7	24

- Test statistic: **N/A**
- P-value: **1.00**
- Interpretation: Do not reject H_0 (No evidence of an association between treatment group and tumour shrinkage).

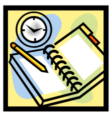
Summary – categorical variables

- **Chi-square test**
  Use when we have two categorical variables, each with two or more levels, and our expected frequencies **are not** too small.
- **Fishers exact test**
  Use when we have two categorical variables, each with two levels, and our expected frequencies **are** small.
- **Chi-square test for trend**
  Use when we have two categorical variables, where one or both are naturally ordered and the ordered variable has at least three levels, and our expected frequencies **are not** too small.

Summary – contingency tables

- Turn scientific question to null and alternative hypothesis
- Calculate expected frequencies
- Think about test assumptions
- Carry out chi-square or Fishers test if appropriate

Contingency table practical



- Work through examples on manual pages 38 – 43
- Complete contingency table practical
- We will have solutions and a summary at 12:30pm

Summary

- For independent observations
- For normally distributed continuous outcomes - T-tests
- For categorical outcomes - Chi-squared tests
- Confidence interval tell us more of story than p-value
- Limitations
  - Confounding – can adjust for important factors by stratification or regression
  - Come and see us!

Sarah.Vowler@cruk.cam.ac.uk
Mark.Dunning@cruk.cam.ac.uk

Statistics Clinic

The Statistics Clinic is held every Wednesday afternoon. Come and get advice in the following areas:

- Study design
- Sample size and replicates
- Grant applications
- Data collection and analysis
- Statistics packages (including R, Stata, SPSS and GraphPad Prism)
- Presentation and interpretation of statistical results
- Paper writing and reviewers' comments
- General questions on statistics

Four 30-minute slots are available each week from 2pm and should be booked in advance.

Please contact **CRIStatsClinic@cruk.cam.ac.uk** to book an appointment.

Finally…

- Course Materials:-
- http://tiny.cc/crukStats
- Course Feedback:-
- http://tiny.cc/stats-june23