Introduction to Statistical Analysis (using Shiny Apps)

CRUK:- Monday 28th November 2016

Mark Dunning, Aaron Lun & John Marioni

www.tiny.cc/crukStats

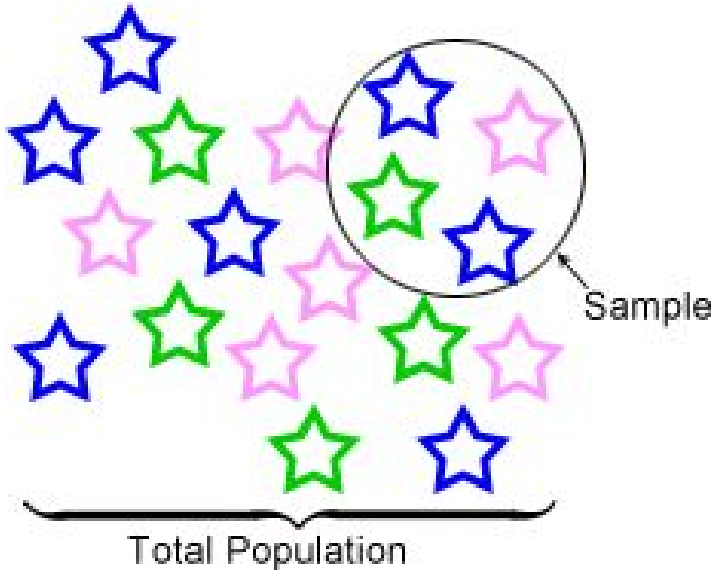
Acknowledgements:- Sarah Vowler, Sarah Dawson, Liz Merrell, Deepak Parashar, Rob Nicholls

Approximate Timetable

10.30 - 11.15 – Lecture: Introduction to Statistical analysis
11.15 - 11.30 – Quiz: Variables/Dependencies/Tests/Generalisability
11.30 - 12.00 – Lecture: Parametric Tests for Continuous Variables; t-tests
12.00 - 12.30 – Examples/Practicals (computer based)
12.30 - 13.30 – Lunch (not provided)
13.30 - 14.00 – Lecture: Non-parametric tests for continuous variable
14.00 - 14.30 – Examples/Practicals (computer based)
(14:30 COFFEE)
14.30 - 14.45 – Lecture: Tests for Categorical Variables
14.45 - 15.30 – Examples/Practicals/Solutions (computer based)
15.30 - 16.25 – Group based exercise: Choosing appropriate tests
16.25 - 16.30 – 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



Beginning a study

- Which samples to include?
  - Randomly selected?
  - Generalisability
- Always think about the statistical analysis
  - Randomised comparisons, or biased?
  - Any dependency between measurements?
  - Data type?
  - Distribution of data?
    - Normally distributed? Skewed? Bimodal?

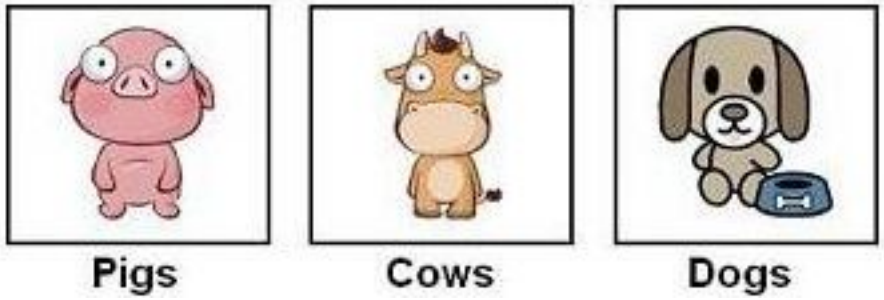
Generalisability

- How samples are selected affects interpretation
  - What is the population that the results apply to?
  - How widely applicable will the study be?
- Statistical methods assume random samples
- Do not extrapolate beyond range of the data
  - i.e. don't assume results apply to anything not represented in the data
- Examples:
  - Males only, no idea about females
  - Adults only, no idea about children

Data types

- Several different categorisations
- Simplest:
  - Categorical (nominal)
  - Categorical with ordering (ordinal)
  - Discrete
  - Continuous

Nominal

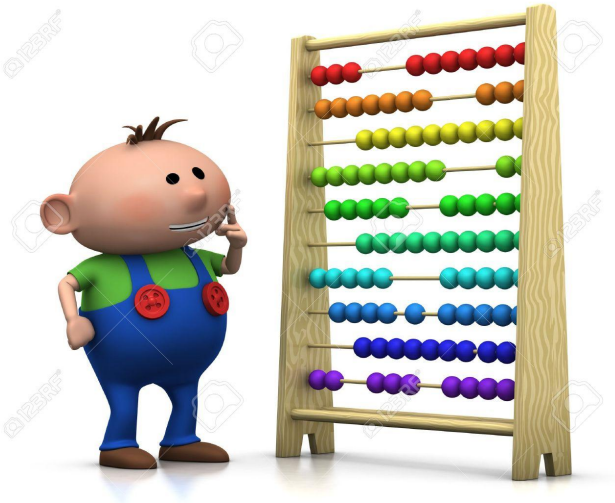


- Most basic type of data
- Three requirements:
  - Same value assigned to all the members of level
  - Same number not assigned to different levels
  - Each observation only assigned to one level
- E.g. gender: 1 = female, 2 = male
- Boils down to yes/no answer
- Others: Surgery type, cancer type, eye colour, dead/alive, ethnicity.

Images:http://www.restore.ac.uk

- Next type of data
- Mutually exclusive fixed categories
- Implicit order
- Can say one category higher than another
  - But not how much higher
- Example: stress level 1 = low … 7 = high
- Others: Grade, stage, treatment response, education level, pain level.

Discrete



- Third level of measurement
- Fixed categories, can only take certain values
- Like ordinal but with well-defined distances
  - Can be treated as continuous if range is large
- Anything counted (cardinal) is discrete
  - *how many*?
- Examples: number of tumours, shoe size, hospital admissions, number of side effects, medication dose, CD4 count, viral load, reads.

Continuous

- Final type of data
- Anything that is measured, can take any value
- May have finite or infinite range
- Zero may be meaningful: ratios, differences
  - Care required with interpretation
- Given any two observations, one fits between
- Examples: Height, weight, blood pressure, temperature, operation time, blood loss, age.

Data types

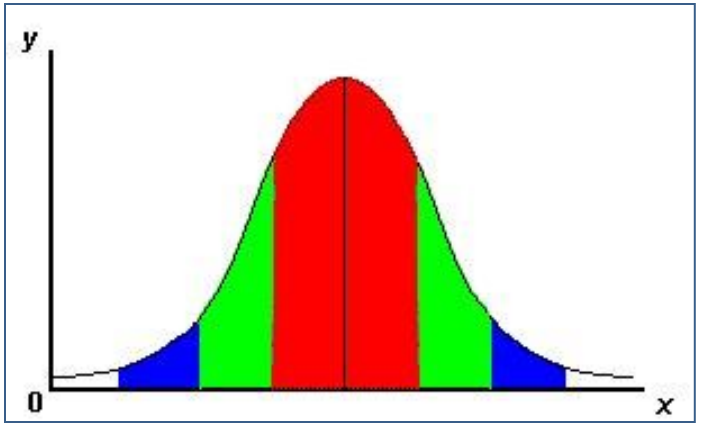
- Several different categorisations
- Simplest:
  - Categorical (nominal) – yes/no
  - Categorical with ordering (ordinal) – implicit order
  - Discrete – only takes certain values; counts (cardinal)
  - Continuous – measurements; finite/infinite range

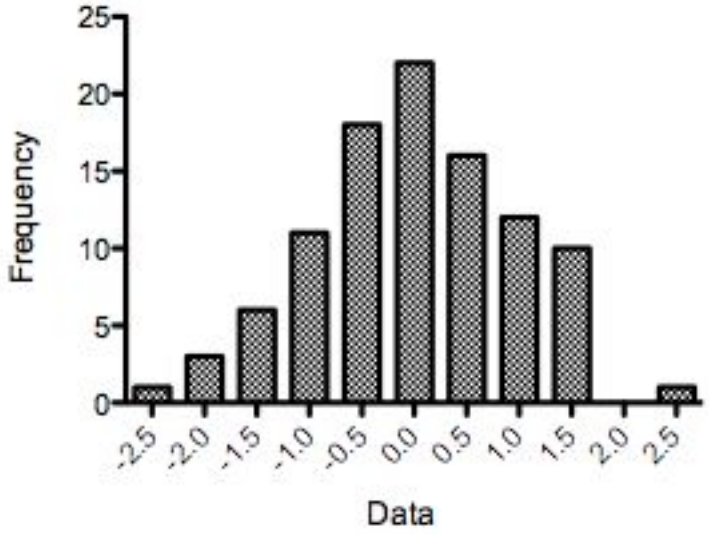
Measurements: Dependent / Independent?

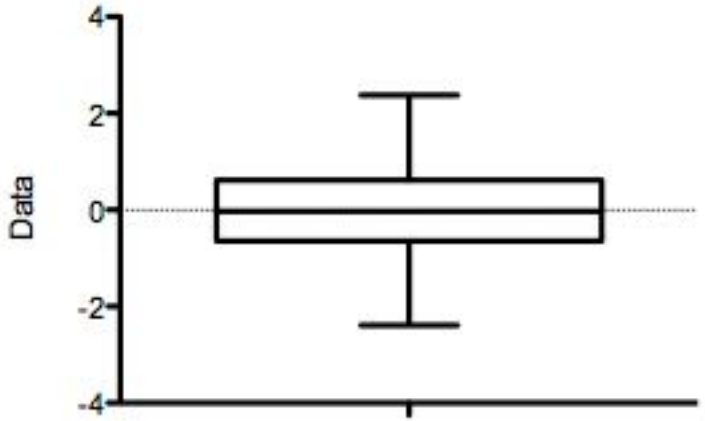
- Measurements of gene expression taken from each of 20 individuals
- Are any measurements more closely related than others?
  - Siblings/littermates?
  - Same individual measured twice?
  - Batch effects?
- If no reason, assume **independent observations**

Continuous Data – Distribution

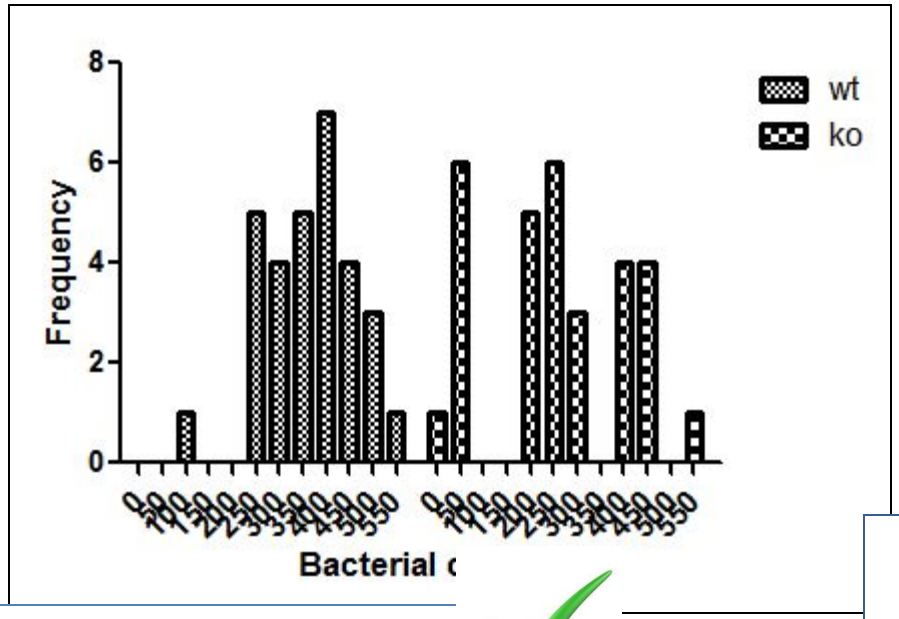
	Y
1	311
2	345
3	270
4	310
5	243
6	530
7	118
8	343
9	277
10	472

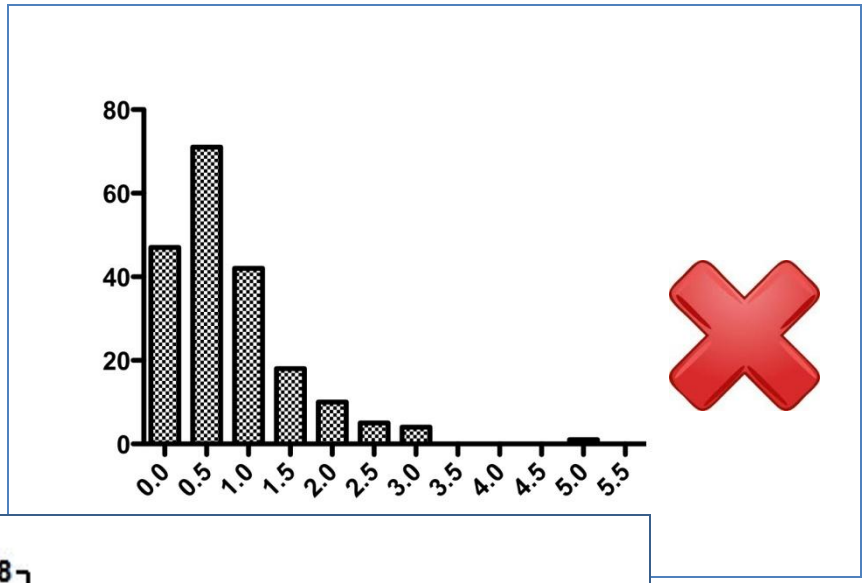


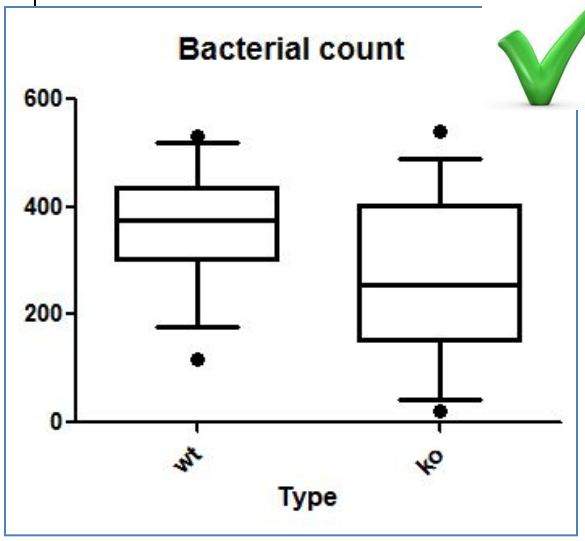


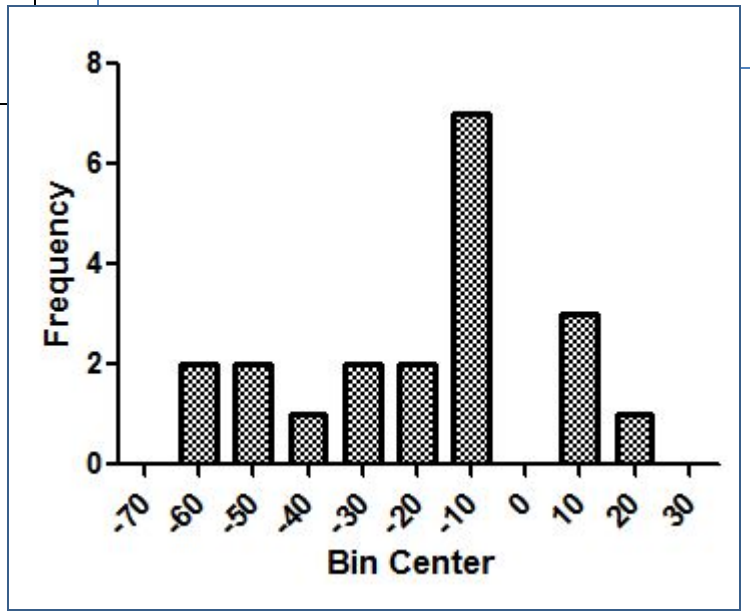


Continuous Data – Distribution?



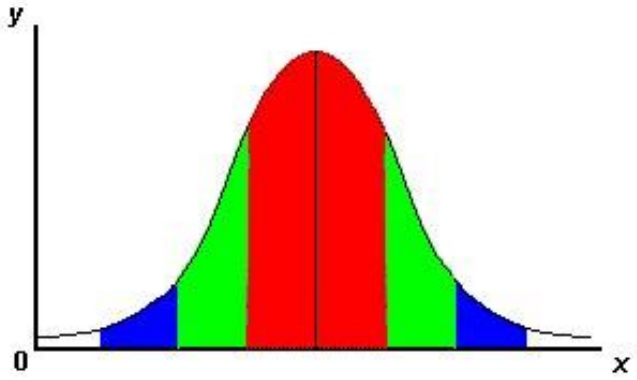






Continuous Data Descriptive Statistics

- Measures of location and spread

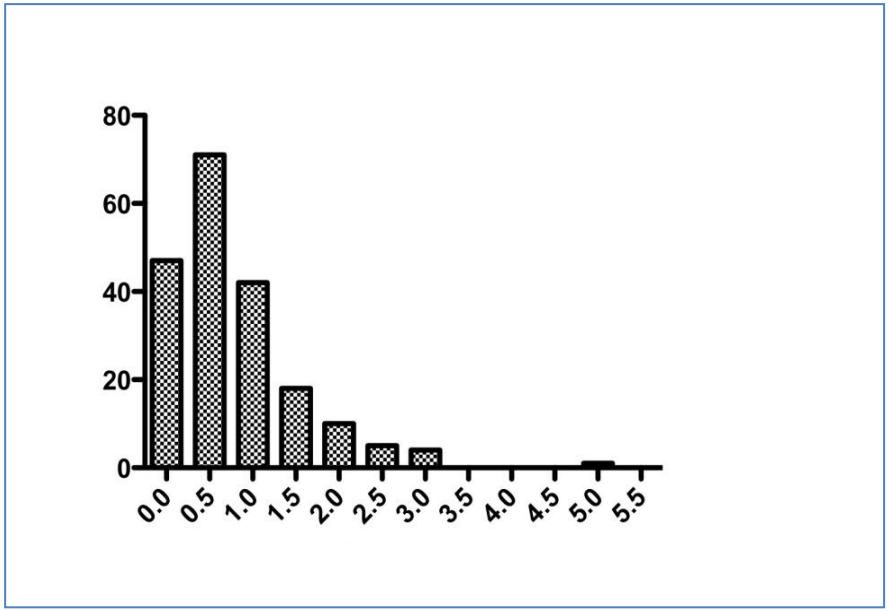


Mean and standard deviation

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

$$s.d. = \sqrt{\frac{\left(X_1 - \overline{X}\right)^2 + \left(X_2 - \overline{X}\right)^2 + \cdots + \left(X_n - \overline{X}\right)^2}{n}}$$

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

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = 970;$$

$$s.d. = \sqrt{\frac{\left(X_1 - \overline{X}\right)^2 + \left(X_2 - \overline{X}\right)^2 + \cdots + \left(X_n - \overline{X}\right)^2}{n}} = 1912.57$$

  - 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

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = 289;$$

$$s.d. = \sqrt{\frac{\left(X_1 - \overline{X}\right)^2 + \left(X_2 - \overline{X}\right)^2 + \cdots + \left(X_n - \overline{X}\right)^2}{n}} = 153.79$$

  - Median and interquartile range

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

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 : low breakdown point

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = 289;$$

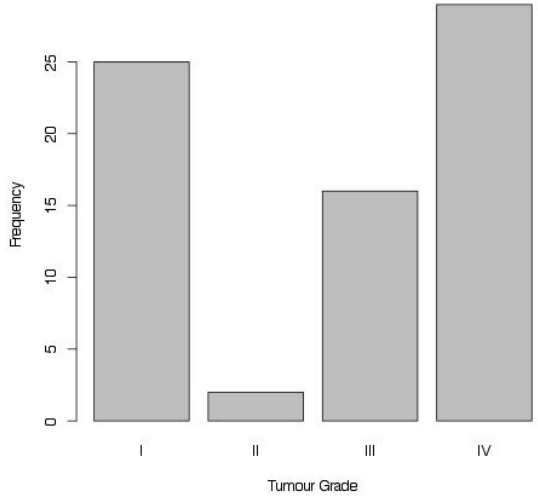
$$s.d. = \sqrt{\frac{\left(X_1 - \overline{X}\right)^2 + \left(X_2 - \overline{X}\right)^2 + \cdots + \left(X_n - \overline{X}\right)^2}{n}} = 153.79$$

  - Median and interquartile range : robust to outliers

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 (SD):
  - Measure of spread of the data
- Standard error (SE):
  - 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 CI covers the mean
- The mean should be in the CI 95% of the time

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

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

Confidence intervals for the mean

- Confidence interval (CI) is a random interval
- In repeated experiments…

  95% of time CI covers the mean
- The mean should be in the CI 95% of the time

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

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

Facebook data: 11, **243**, 270, **310**, 311, **345**, 530

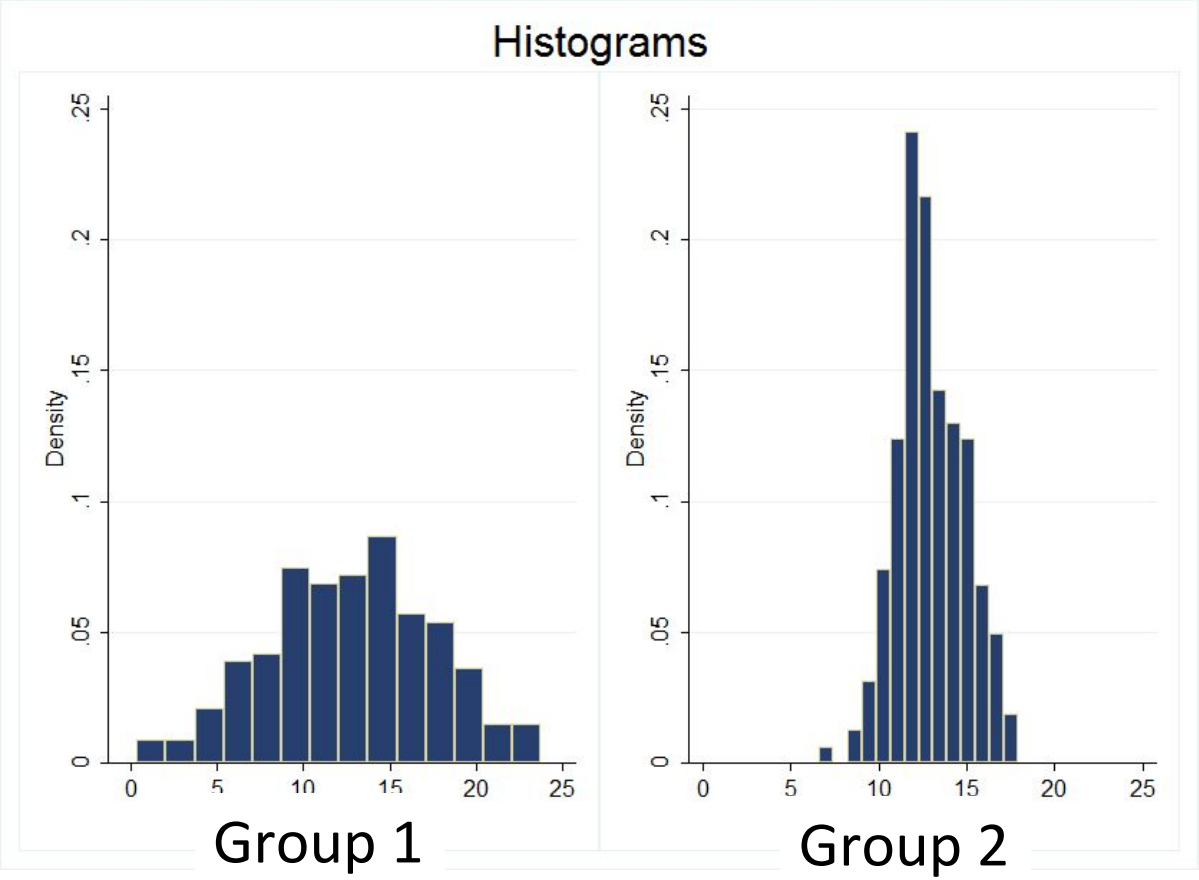
Mean 289, 95% CI (175, 402)

Confidence intervals

No. of observations (samples)

Standard deviation

Standard error of mean



Hypothesis tests – basic set-up

- Formulate a null hypothesis, H_0

Example: the difference in gene expression before and after treatment = 0

- Calculate a test statistic from the data under the null hypothesis

$$t_{n-1} = t_{29} = \frac{\overline{X}_{After-Before}}{s.e.\left(\overline{X}_{After-Before}\right)}$$

Hypothesis tests – basic set-up

- Formulate a null hypothesis, H_0

Example: the difference in gene expression before and after treatment = 0

- Calculate a test statistic from the data under the null hypothesis

$$t_{n-1} = t_{29} = \frac{\overline{X_{After-Before}}}{s.e.\left(\overline{X_{After-Before}}\right)}$$

- Compare the test statistic to theoretical values
  Is it more extreme than expected? (p-value)

- Either reject or do not reject the null hypothesis

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

Hypothesis tests – basic set-up

- Formulate a null hypothesis, H_0

Example: the difference in gene expression before and after treatment = 0

- Calculate a test statistic from the data under the null hypothesis

$$t_{n-1} = t_{29} = \frac{\overline{X_{After-Before}}}{s.e.\left(\overline{X_{After-Before}}\right)}$$

- Compare the test statistic to theoretical values
  Is it more extreme than expected? (p-value)

- Either 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

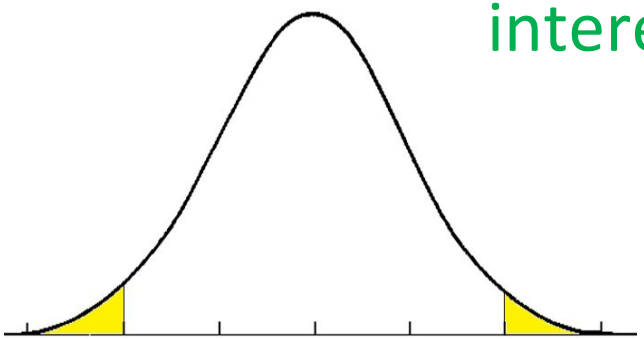
- Randomly ordered 8 cups of tea
  - 4 were prepared by first adding milk
  - 4 were prepared by first adding tea
- Task: Lady had to select the 4 cups of one particular method
- H_0: Lady had no such ability
- Test Statistic: number of successes in selecting the 4 cups.
- Result: Lady got all 4 cups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

When to use which test (see course site for "cheat sheet")

		RESPONSE		
NO OF SAMPLES		**NOMINAL**	**ORDINAL OR NON-NORMAL**	**NORMALLY DISTRIBUTED**
ONE SAMPLE		χ^2-test, Z-test	Kolmogorov-Smirnov Sign test	t-test
TWO SAMPLE	**INDEPENDENT**	χ^2-test (r x c), Fisher's exact test	Mann-Whitney U Median test	Unpaired t-test
	PAIRED	McNemar's test Stuart-Maxwell test	Wilcoxon signed rank Sign test	Paired t-test
MULTIPLE SAMPLES (K>2)	**INDEPENDENT**	χ^2-test (r x k) Fisher-Freeman-Halton	Kruskal-Wallis test Median Test Jonckheere-Terpstra test	Analysis of variance (ANOVA)
	PAIRED	Cochran Q test	Friedman test Page test Quade test	Repeated measures ANOVA
ASSOCIATION BETWEEN TWO VARIABLES		Contingency coefficient Phi, r_ϕ Cramér, C	Spearman's rank Kendall's tau	Pearson product moment correlation
AGREEMENT BETWEEN TWO VARIABLES		Simple kappa	Weighted kappa	Limits of agreement