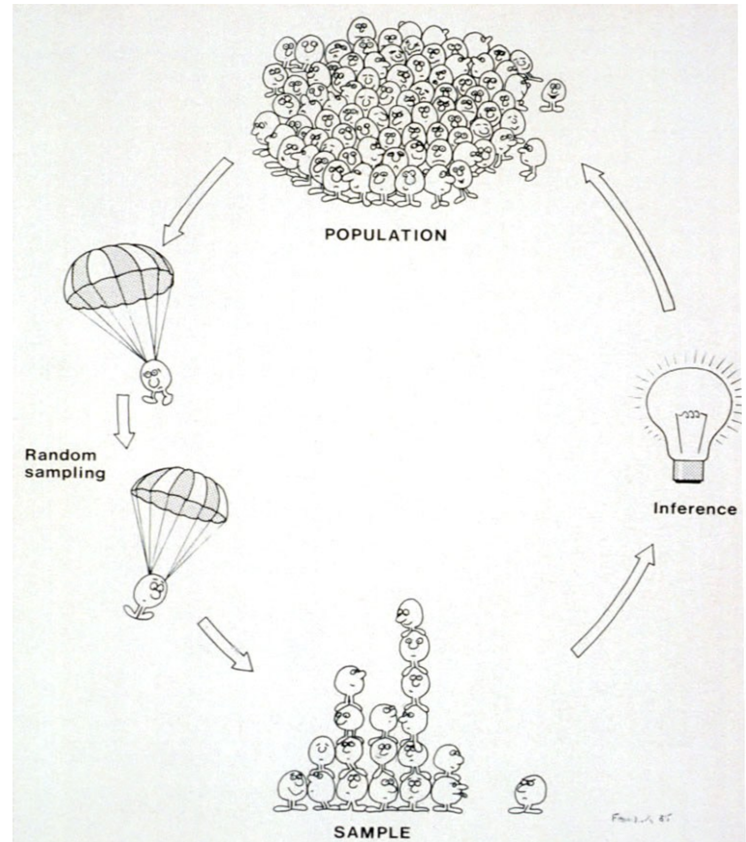
Statistical Tests for two (or more) groups

Tests used to compare two or more groups	▪ Interpret tests comparing means and percentages

Question of interest	population
Experiment	sample
Graphs and summary statistics	sample
Subjective answer to question	sample
Formal analysis (statistical inference) answer to question	population



You want to test your hypothesis

What statistical test should you use?

Students 1-sample t-test, Students 2-sample t-test, Students paired t-test, Wilcoxon signed rank test, Mann Whitney test, chi-square test, Fisher's test, McNemar's test.

Choice of test depends on whether -

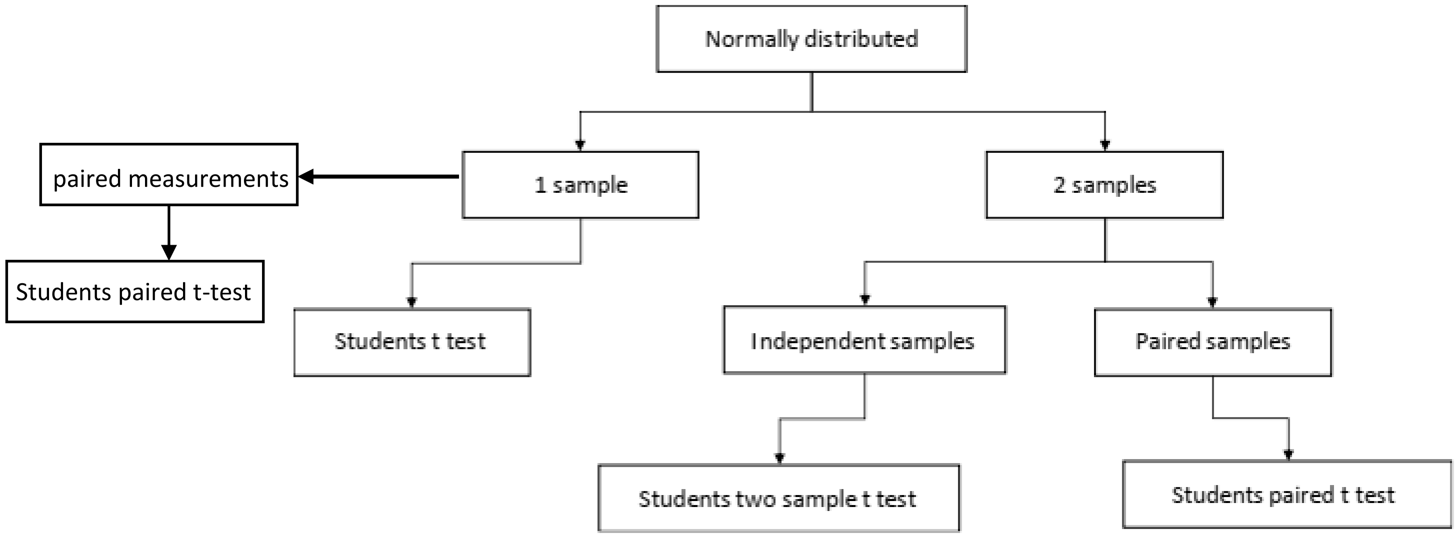
Data are numerical measurements or categorical counts

Whether numerical measurements can be assumed to have a normal distribution (parametric)

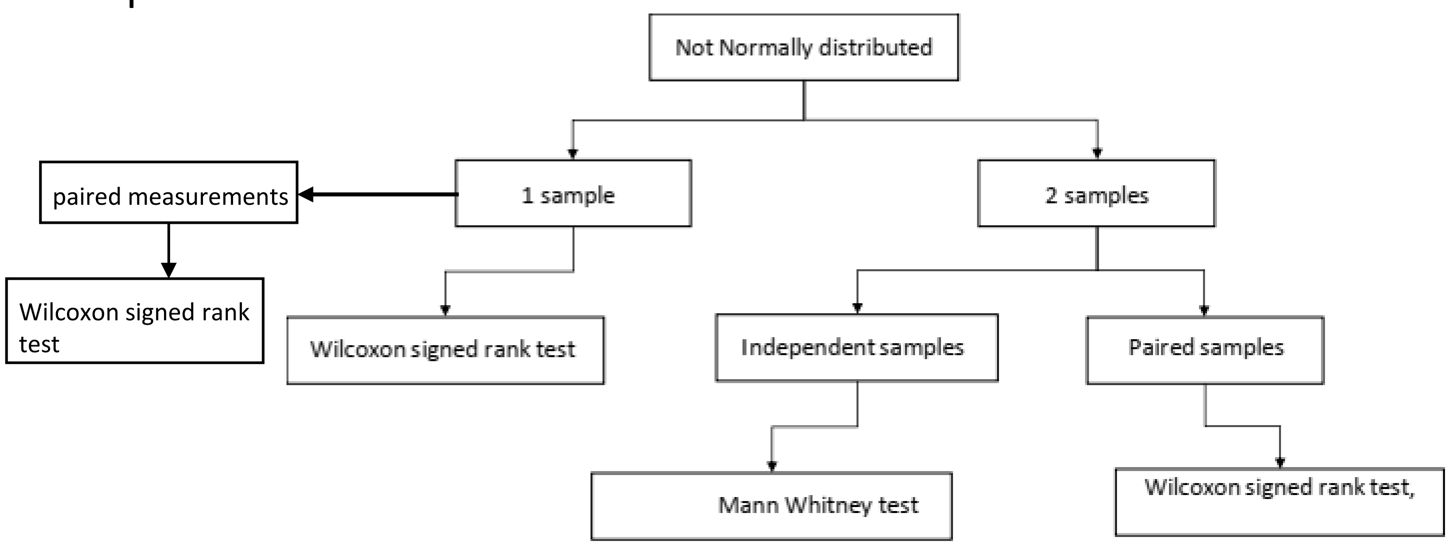
Whether the observations come from 1 sample, or 2 samples, or the observations are paired (i.e measured in the same individuals twice)

Numerical data

Parametric tests



Non-parametric tests



Numerical measurement data

One-sample

one-sample tests usually compare the mean of a sample to a specific value (usually a known population mean value)

it seeks to test whether the population mean value is equal to a specific value

H_0: population mean = specific value

H_1: population mean ≠ specific value

Numeric data

One-sample

H_0: population mean = specific value

H_1: population mean ≠ specific value

Check data

Do the data come from a **Normal** distribution?

Are the data independent

Numeric

One-sample

Data from Normal distribution

Use parametric test

Student's one sample t-test

H_0: population mean = specific value

H_1: population mean ≠ specific value

William Gosset (1876-1937)

Forced to use pseudonym **Student** by his employer, Guinness

Numeric

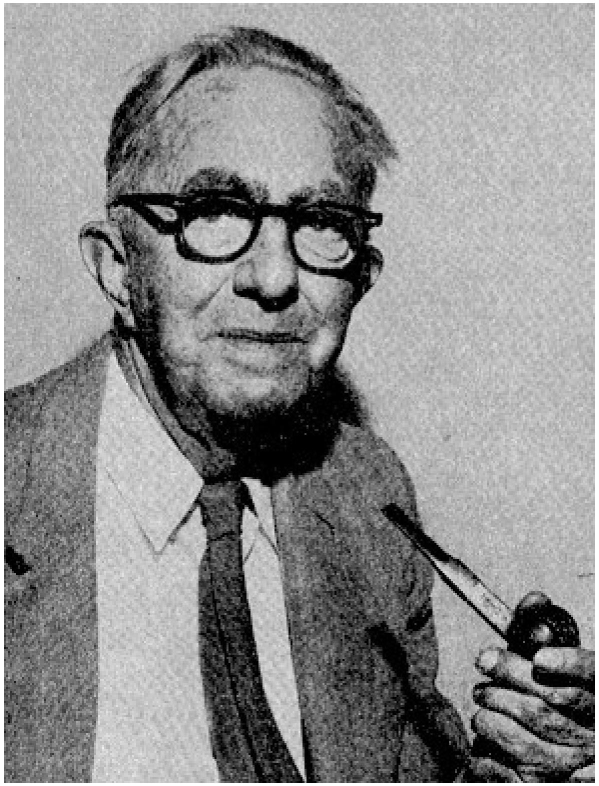
One-sample - Data not Normal distribution

Use non-parametric test

Wilcoxon signed rank test

H_0: population median = specific value

H_1: population median ≠ specific value



Frank Wilcoxon (1892-1965)

Wilcoxon and Mann and Whitney described rank sum tests, which have been shown to be the same.

Convention ascribes the Wilcoxon test to 1-sample or paired data

and the Mann-Whitney U test to unpaired data.

Parametric tests are based on the means of the data

Non-parametric tests are based on the medians

Non-parametric tests are not as powerful as parametric tests if the assumption of normality holds

Look at data and use parametric tests if possible

Large sample sizes - you can get away with the use parametric tests (central limit theorem)

One-sample

Tests provide …

95% Confidence interval for mean (can be produced for median)

(Does it contain the specific value?)

H_0: mean (median) = specific value

H_1: mean (median) ≠ specific value

A p-value<0.05 shows evidence against the null hypothesis.

Example: plasma calcium in Everley's syndrome

Sample of 18 patients aged 20-44 with Everley's syndrome; a rare congenital disease.

Mean plasma calcium of 3.2 mmol/l (SD 1.1) in the sample

Population mean is 2.5 mmol/l in healthy people of similar age

Is the population mean in patients with Everley's syndrome abnormally high?

Data are assumed to be from a Normal distribution.

H_0: mean = 2.5

H_1: mean ≠ 2.5

t-statistic=-2.69; $p<0.02$; 95% Confidence interval (2.65 to 3.75)

Reject H_0 in favour of H_1

Summary Numeric data – one sample

Distribution	Appropriate test	Test of
Normal	one-sample t-test	Mean
Non-normal / outliers	Wilcoxon signed rank test	Median

Numerical measurement data - 2 independent samples/groups

Numeric data

two-samples/groups

Used when you want to see if a quantitative measurement differs between 2 groups

H_0: mean (group 1) = mean (group 2)

H_0: mean (group 1) - mean (group 2) = 0

H_1: mean (group 1) ≠ mean (group 2)

H_1: mean (group 1) - mean (group 2) ≠ 0

Numeric

two-sample - independent

H_0: mean group 1 = mean group 2

H_1: mean group 1 ≠ mean group 1

Check data

Do the data come from a **Normal** distribution?

Numeric

two-sample - independent

Data from Normal distribution

Use parametric test

Students two sample t-test

H_0: population mean group 1 = population mean group 2

H_1: population mean group 1 ≠ population mean group 1

Numeric

two-sample - independent

Data not from Normal distributions

Use non-parametric test

Mann-Whitney test

H_0: population median group 1 = population median group 2

H_1: population median group 1 ≠ population median group 2

Two-sample independent

Tests provide …

95% Confidence interval for ***difference*** in population means

(Does CI contain 0? i.e. could the difference in population means be 0?)

H_0: population mean (median) group 1 = population mean (median) group 2

H_1: population mean (median) group 1 ≠ population mean (median) group 2

A p-value<0.05 shows evidence against the null hypothesis.

or weak evidence in support of the null hypothesis.

Example

2 groups of 10 renal transplant patients.
One group given Placebo. One group given Fluvastatin.

% Change in LDL is measured at 6 weeks.

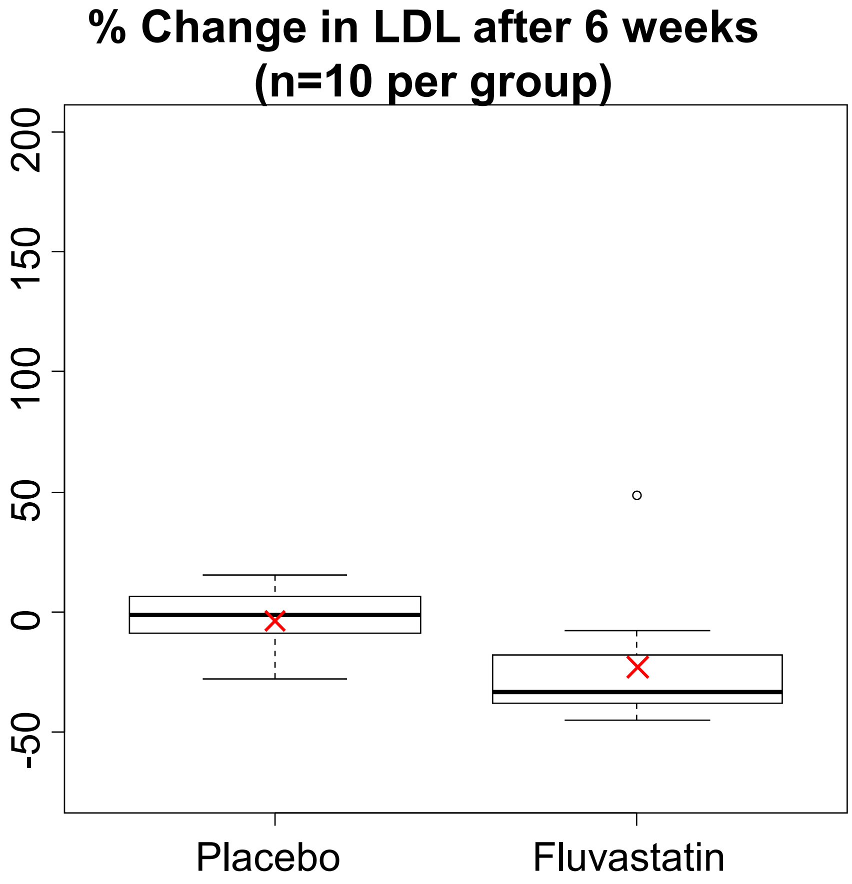
Is population mean %Δ in LDL different for Placebo and Fluvastain?

Normal distribution is assumed.

Use two-sample t-test

$\mathbf{H_0}$**:**Population mean %Δ LDL same for Fluvastatin & Placebo

$\mathbf{H_1}$**:**Population mean %Δ LDL not same for Fluvastatin & Placebo


% Change in LDL after 6 weeks
(n=10 per group)
200
150
100
50
0
-50
Placebo
Fluvastatin

Placebo patients mean %Δ LDL -3.36 (SD 13.33)

Fluvastatin patients mean %Δ LDL -22.68 (SD 27.49)

H_0: Population mean %Δ LDL same for Fluvastatin & Placebo

H_1: Population mean %Δ LDL not same for Fluvastatin & Placebo

Use two-sample t-test

t-statistic=2.00; p=0.061; 95% Confidence interval (-0.97 to 39.62)

95% Confidence that the true population difference could be as low as -0.97 or as high as 39.62

Fail to Reject H_0

Example (full data)

2 groups of 500 renal transplant patients.
One group given Placebo. One group given Fluvastatin.

% Change in LDL is measured at 6 weeks

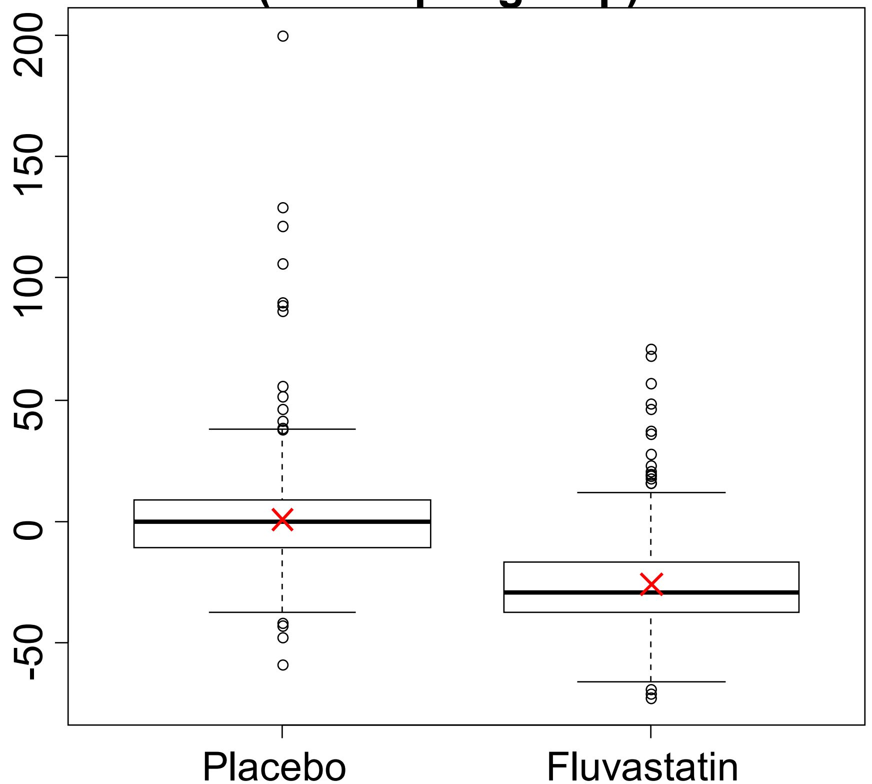
Normal distribution is assumed – sample size is large.

Use two-sample t-test

H_0: Population mean %Δ LDL same for Fluvastatin & Placebo

H_1: Population mean %Δ LDL not same for Fluvastatin & Placebo

% Change in LDL after 6 weeks (n=500 per group)



Example

Placebo group mean %Δ LDL 0.89 (SD 21.60)

Fluvastatin group mean %Δ LDL mean -25.74 (SD 18.37)

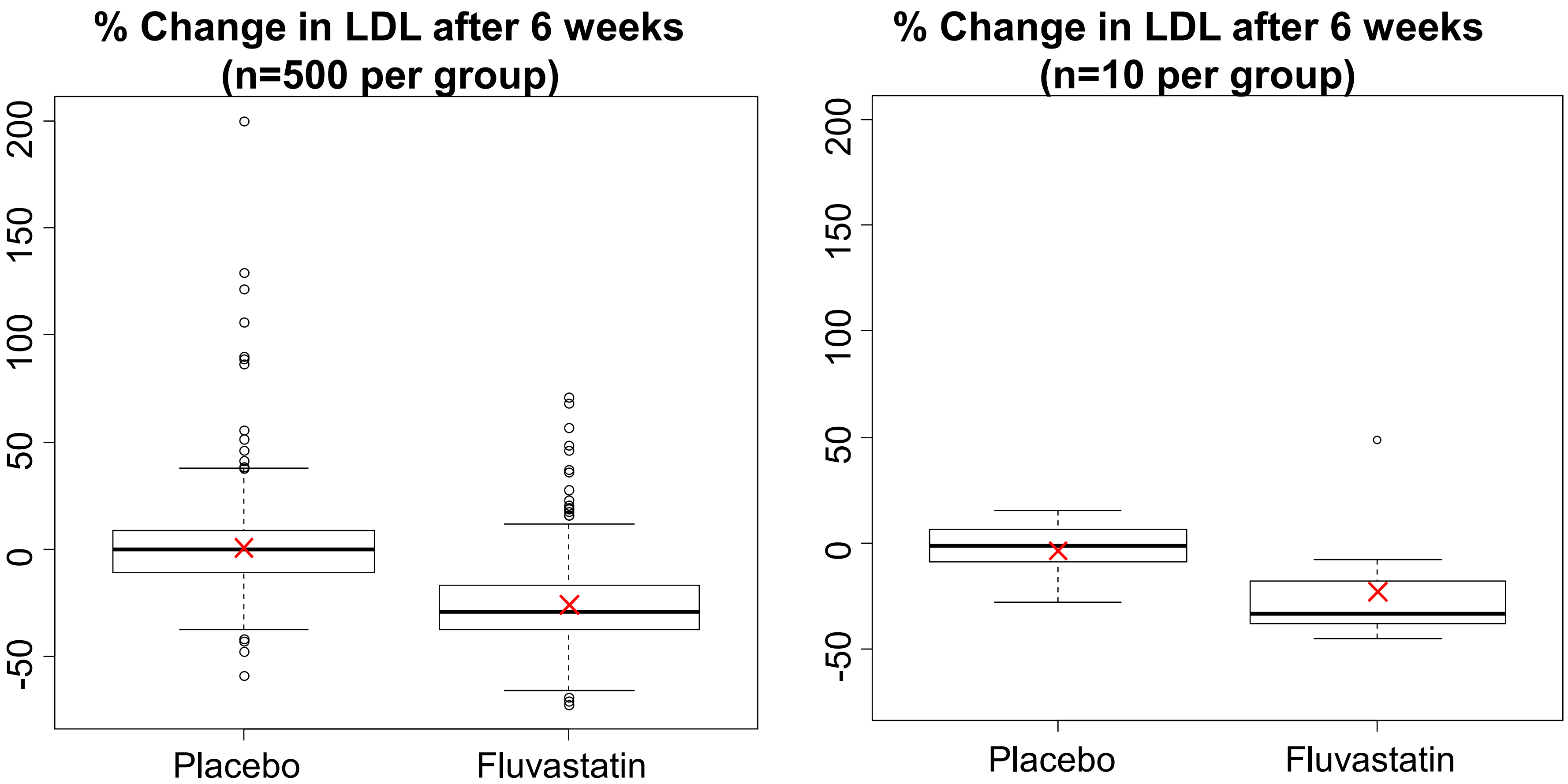
H_0: Population mean %Δ LDL same for Fluvastatin & Placebo

H_1: Population mean %Δ LDL not same for Fluvastatin & Placebo

t-statistic=20.3; p<0.0001; 95% Confidence interval (24.1to 29.2)

95% Confidence that the true population difference could be as low as 24.1 or as high as 29.2

Reject H_0



% Change in LDL after 6 weeks
(n=500 per group)
200
150
100
50
0
-50
Placebo
Fluvastatin
% Change in LDL after 6 weeks
(n=10 per group)
200
150
100
50
0
-50
Placebo
Fluvastatin

Numerical measurement data –

1 sample with 2 measurements

Or

2 non-independent samples

Numeric

two-sample - paired

H_0: population mean group 1 = population mean group 2

H_1: population mean group 1 ≠ population mean group 2

Check data

Are there **significant outliers**? Histograms/Box plots/dot plots

Do the data come from a **Normal** distribution?

Numeric

two-sample – paired

Data from Normal distribution

Use parametric test

Students paired t-test

H_0: population mean group 1 = population mean group 2

H_1: population mean group 1 ≠ population mean group 1

Numeric

two-sample - paired

Data not for Normal distributions

Use non-parametric test

Wilcoxon signed rank test

H_0: population median group 1 = population median group 2

H_1: population median group 1 ≠ population median group 2

Two-sample paired

Tests provide …

95% Confidence interval for ***difference*** in means

(Does CI contain 0? i.e. could the difference in population means be 0?)

H_0: mean (median) group 1 = mean (median) group 2

H_1: mean (median) group 1 ≠ mean (median) group 1

A p-value<0.05 shows evidence against the null hypothesis.

62 lung cancer patients. Boxplots show tumour volume before and after SABR

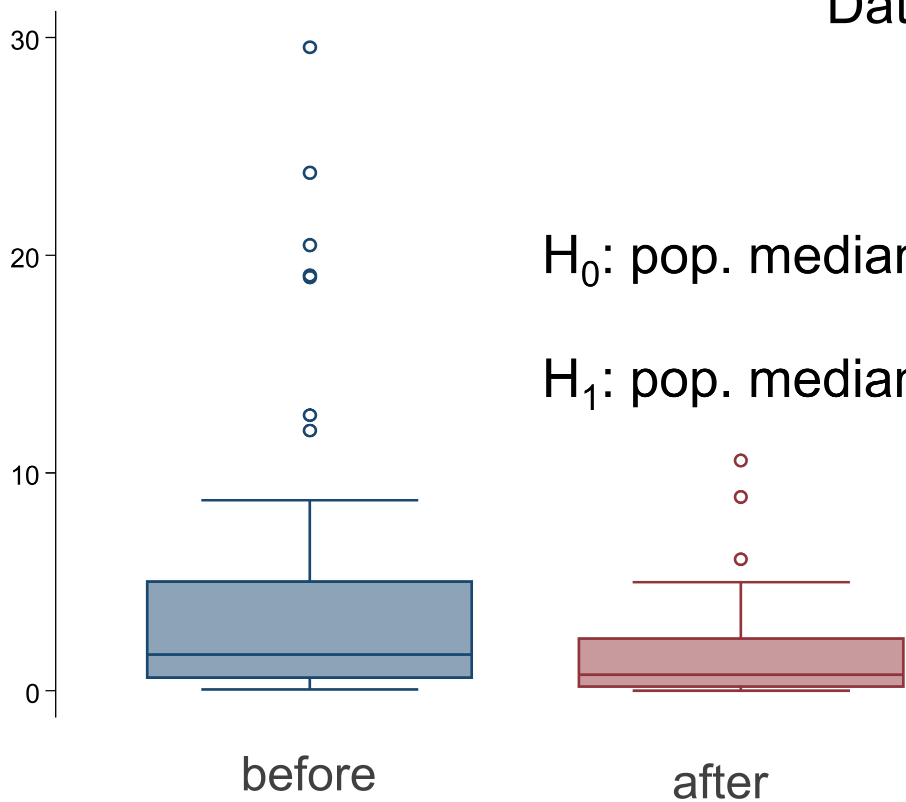
Data are skewed & not Normal

Data are paired

Wilcoxon signed-rank test

H_0: pop. median vol before = pop. median vol after

H_1: pop. median vol before ≠ pop. median vol after



Median=1.64

Median=0.73

Z=6.5, p<0.0001

H_0 is rejected

Summary Numeric data – two samples

Distribution	Groups are	Appropriate test	Test of
Normal	independent	two-sample t-test	Means
	paired	paired sample t-test	
Non-normal / outliers	independent	Mann-Whitney test	Medians
	paired	Wilcoxon signed rank test	

Comparing more than 2 samples/groups

Groups **A**, **B**, **C**, **D** …. etc

If data are Normally distributed, differences between the groups can be tested using analysis of variance (**ANOVA**).

The non-parametric equivalent is **Kruskal-Wallis.**

Multiple comparisons

Imagine 3 treatment groups **A**, **B**, **C**

How many comparisons can we make?

A with **B**, **A** with **C**, **B** with **C** (3 comparisons)

4 treatment groups **A**, **B**, **C**, **D**

How many comparisons can we make?

A with **B**, **A** with **C**, **A** with **D**, **B** with **C**, **B** with **D**, **C** with **B** (6 comparisons)

5 treatment groups **A**, **B**, **C**, **D, E**

Multiple comparisons

	Probability of observing at least one significant result due to chance
1 comparison	5%
2 comparisons	10%
3 comparisons	14%
4 comparisons	19%
10 comparisons	40%
20 comparisons	64%

Bonferroni correction.

If multiple tests are performed then the significance level has to be adjusted.

Without adjustment the overall type I error will be higher than 5%.

This is achieved using a **Bonferroni correction**

Adjusted significance level = α**/n** where n=number of tests

Subgroup analyses

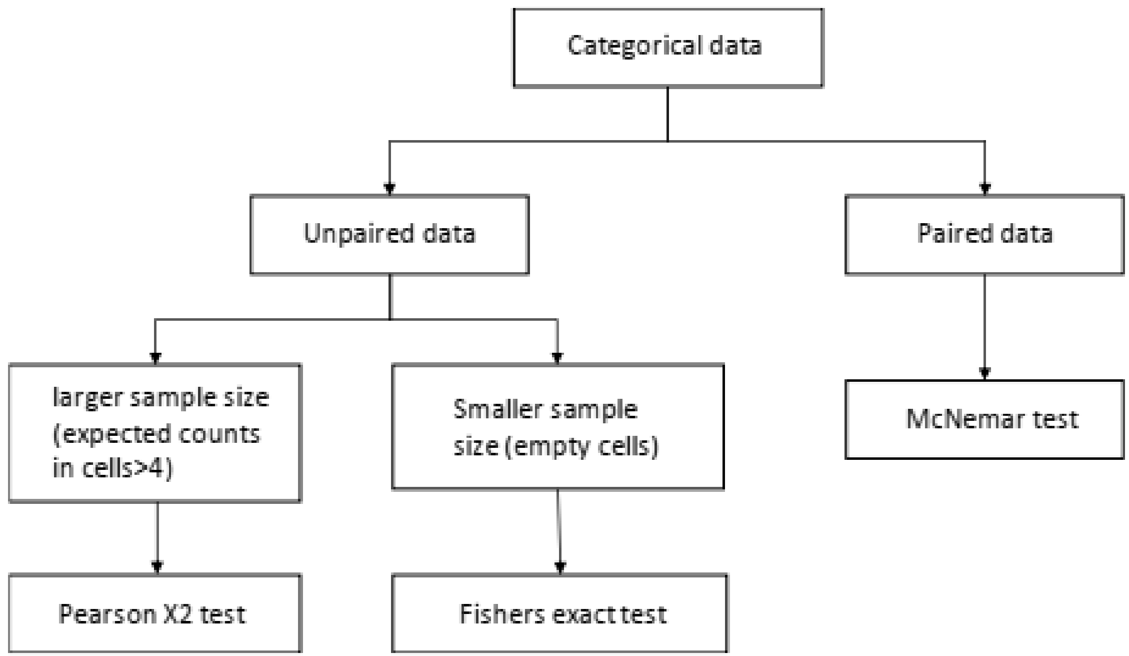
Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17 187 cases of suspected acute myocardial infarction. Lancet. 1988; ii: 349–360

A table reporting subgroup analyses of death after streptokinase, aspirin, both, or neither for acute myocardial infarction.

For people with the star signs Gemini and Libra, aspirin was no better than placebo.

For others, aspirin had a strongly beneficial effect.

Tests for categorical data


Categorical data
Unpaired data
Paired data
larger sample size (expected counts in cells>4)
Smaller sample size (empty cells)
McNemar test
Pearson X2 test
Fishers exact test

Categorical data

We may wish to test whether or not there is an association between sex (categorical: male, female) and stage at presentation (categorical: I, II, III, IV, V)

H_0: There is no association between the categorical variables

(i.e. % of men in each stage category is the same as the % of women in each stage category)

H_1: There is an association between the categorical variables

(i.e. % of men in each stage category is not the same as the % of women in each stage category)

Are the data paired or unpaired?

Categorical data - unpaired

Variable 1

Variable 2		1	2
	1	O_{11}	O_{12}
	2	O_{21}	O_{22}

Simple case of categorical variables with 2 categories 2x2

O is number of people in each category

Can be extended to NxM categories

Categorical data- unpaired

		Variable 1	
		1	2
Variable 2	1	O_{11}	O_{12}
	2	O_{21}	O_{22}

		Sex	
		Men	Women
Outcome	Died	13	12
	Alive	29	48

2 year outcome among 102 SABR lung cancer patients

2 year outcome among 102 (42 men and 60 women) SABR lung cancer patients

Outcome		Sex: Men	Sex: Women
	Died	13	12
	Alive	29	48

25% (25/102) died

31% (13/42) men died
20% (12/60) women died

Is there an association between outcome at 2 years and sex?

Is the population % who die the same for men and women?

Categorical data - unpaired

There are sufficient people in each cell of the cross-tabulation

Pearsons χ^2

Chi-square test

Outcome	Sex: Men	Sex: Women
Died	13	12
Alive	29	48

$$= \sum \frac{(O - E)^2}{E}$$

E = expected number in each cell

O = observed number in each cell

Categorical data - unpaired

Observed numbers

Outcome	Sex: Men	Women
Died	13	12
Alive	29	48

Expected numbers assuming H_0 true

	Men	Women
Died	10	15
Alive	32	45

$$\text{Chi-square} = \sum \frac{(0 - E)^2}{E} = 1.60$$

P=0.21 do not reject H_0

Karl Pearson

1857-1936

Mathematician & Statistician

$$\chi^2 = \sum \frac{(0 - E)^2}{E}$$

Categorical - unpaired

There are insufficient numbers in each cell of the cross-tabulation

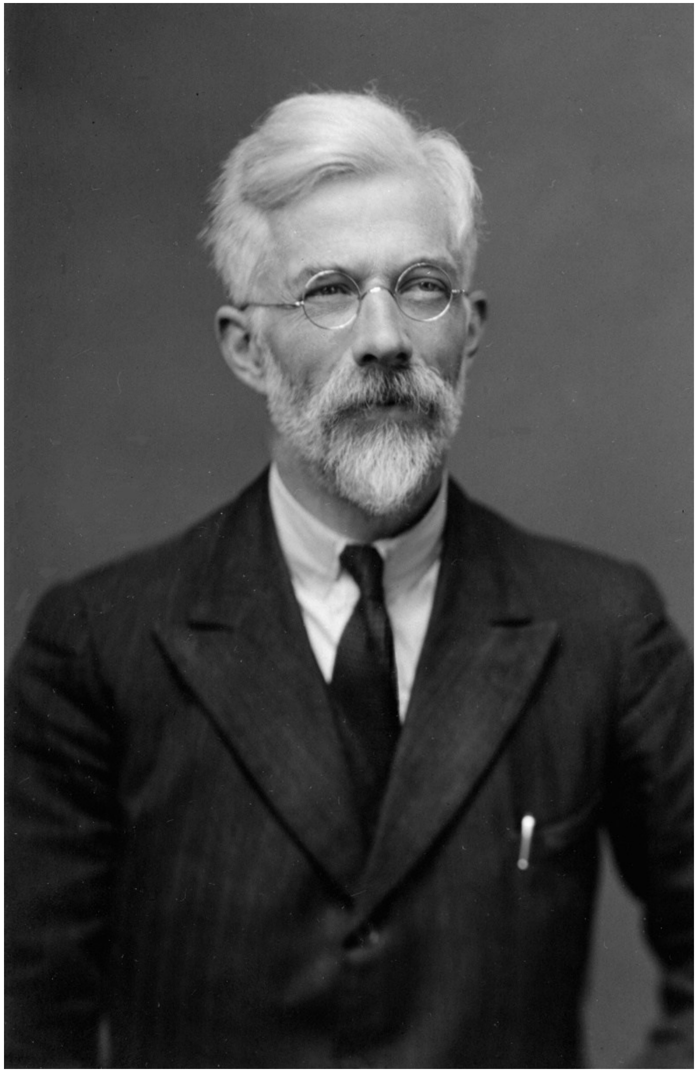
Outcome \ Sex	Men	Women
Died	3	1
Alive	9	8

25% of men and
11% of women died

Fishers exact test

$p=0.60$

Do not reject H_0



Ronald Fisher

1890-1962

Statistician & biologist

Categorical – paired data

Skin improvement among 85 patients with severe psoriasis treated with topical therapy A (on left side of trunk) and B (on right side)

Skin improvement	treatment A		treatment B	
No	25		15	
Yes	60	71%	70	82%

Chi-square test=3.3 $p = 0.07$ Do not reject H_0

Categorical – paired data

Skin improvement among 85 patients with severe psoriasis treated with topical therapy A (on left side) and treatment B (on right side)

		Skin improvement treatment B: No	Skin improvement treatment B: Yes
Skin improvement treatment A	No	10	15
	Yes	5	55

McNemar's test

P=0.025

Reject H_0

Test	Appropriate when
Chi-square test	Data are unpaired and there are sufficient numbers in each cell of a cross-tabulation
Fisher’s exact test	Data are unpaired and there are sufficient numbers in each cell of a cross-tabulation
McNemar’s test	Paired data

A p-value <0.05 in these tests is indicative of an association between the variables.

Sample size considerations

How many subjects should be included in a study

Too few people, the power to detect a clinically significant effect will be low

A sample size that is larger than required is difficult to achieve and expensive.

Recruiting patients to a study which will be too small to detect the minimum effect we are looking for or recruiting more patients than necessary (over-powered) can be considered unethical.

The **power** of a test is chance of detecting, as statistically significant, a real treatment effect.

Power (1-β) is the probability of rejecting the null hypothesis when it is false;

α is the **significance level.**

It **is the probability of rejecting the null hypothesis when it is true**;

i.e. it is the chance (usually expressed as a percentage) of detecting, as statistically significant, a treatment effect when none exists.

To establish the sample size needed for a study the following factors should be known …

1) The minimum size of the effect to be detected

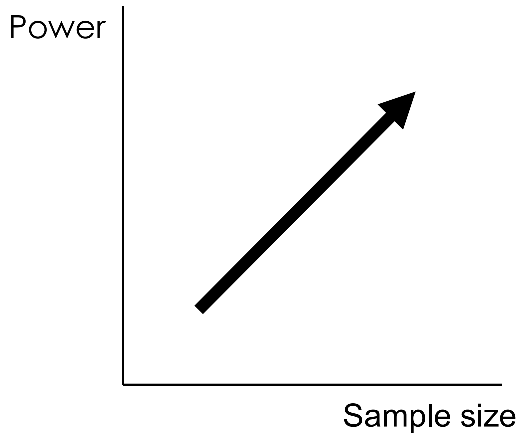
2) The variability (standard deviation)

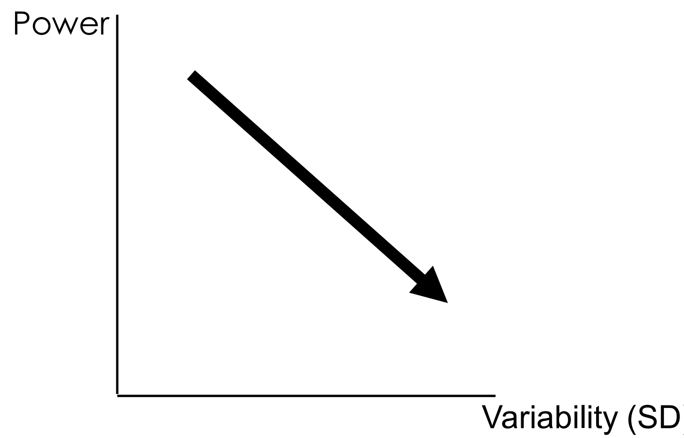
3) The power required

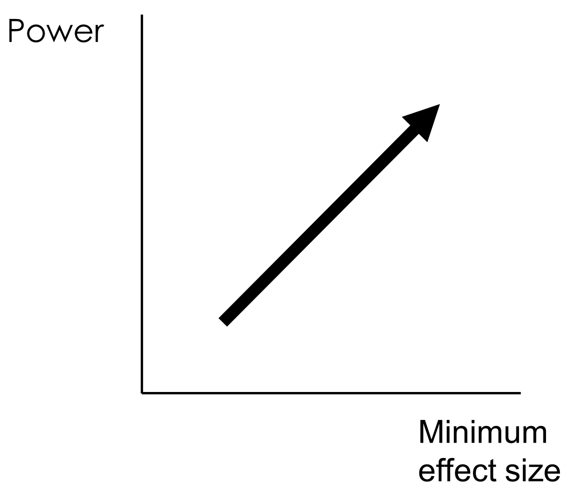
4) The significance level

For a chosen significance level, power, minimum size of effect to be detected and standard deviation the sample size needed can be calculated.

Power is the probability of rejecting the null hypothesis when it is false







For a chosen significance level, power, minimum effect size, and SD the sample size can be calculated.