Hypothesis testing (in R)

Catalina Vallejos (The Alan Turing Institute and UCL)
Aaron Lun (Cancer Research UK - Cambridge Institute)

Recall: hypothesis testing

Often, the aim of the analysis is not only **estimation**.

For example,

- *Is the treatment effective?*
- *Is a gene differentially expressed?*
- *Is there an association between genotype and phenotype?*

These questions can be translated as **hypothesis testing** problems

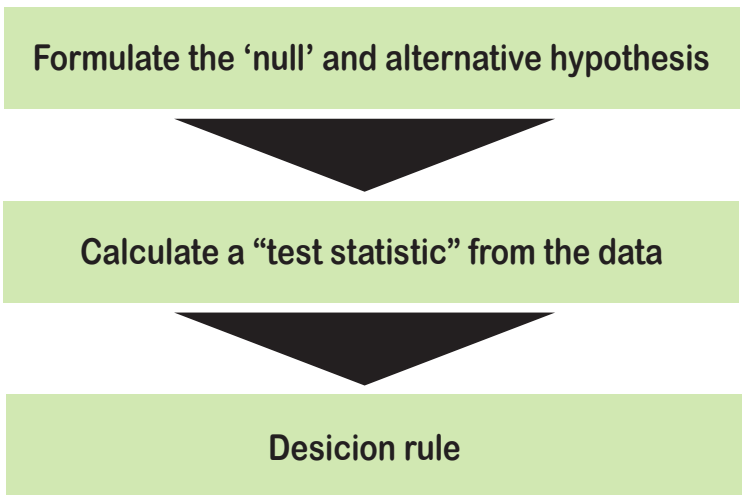
Comic taken from Significance Magazine, December 2008.

Outline for the course

- **Introduction to statistics**
- **Hypothesis testing (in R)**
  - Test for categorical variables
  - Tests for continuous variables
    - Parametric
    - Non-parametric
- **Linear regression (in R)**
- **Multiple testing (in R)**

Hypothesis testing

Basic setup



This afternoon:

Hypothesis testing in R

In particular,

- Test vs data type
- Hypothesis formulation
- Assumptions

Tests for categorical variables

Tests for categorical variables: an example

A trial was designed to assess the effectiveness of a new treatment versus a placebo in reducing tumour size (ovarian cancer patients)

The data contains the following information

- Patient group: treatment / placebo
- Tumour shrinkage: yes / no

How would you summarise the data?

Tests for categorical variables: an example

In the example, the data is as follows

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

Tests for categorical variables: an example

In the example, the data is as follows

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

This is called a 2×2 **contingency table**

Tests for categorical variables: an example

In the example, the data is as follows

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

This is called a 2×2 **contingency table**

Next steps:

1. Formulate the hypothesis
2. Calculate a "test statistic" (evidence)
3. Decision rule

Pearson's χ^2 test for independence

1. Formulate the hypothesis

Pearson's χ^2 test for **independence** is designed to contrast the following hypothesis

$$H_0: \text{ Independence} \quad \text{vs} \quad H_1: \text{ Not independence}$$

Pearson's χ^2 test for independence

1. Formulate the hypothesis

Pearson's χ^2 test for **independence** is designed to contrast the following hypothesis

$$H_0 : \text{ Independence} \quad \text{vs} \quad H_1 : \text{ Not independence}$$

In the example,

H_0 : no association between treatment and tumour shrinkage

H_1 : treatment and tumour shrinkage are associated

Pearson's χ^2 test for independence

2. Calculate a "test statistic"

Pearson's χ^2 test for independence

2. Calculate a "test statistic"
    2.1. Calculate expected frequencies under independence

Pearson's χ^2 test for independence

2. Calculate a "test statistic"
    2.1. Calculate expected frequencies under independence

Independence is equivalent to probabilities that factorise

For example, when *independently* tossing a coin twice:

What is the probability of observing $\{H, H\}$?

Pearson's χ^2 test for independence

2. Calculate a "test statistic"
   2.1. Calculate expected frequencies under independence

Independence is equivalent to probabilities that factorise

For example, when *independently* tossing a coin twice:

What is the probability of observing {H, H}? $\frac{1}{2} \times \frac{1}{2}$

Pearson's χ^2 test for independence

2. Calculate a "test statistic"
   2.1. Calculate expected frequencies under independence

Independence is equivalent to probabilities that factorise

For example, when *independently* tossing a coin twice:

What is the probability of observing $\{H, H\}$? $\frac{1}{2} \times \frac{1}{2}$

If we repeat the experiment 5 times: how many times we expect to observe $\{H, H\}$?

Pearson's χ^2 test for independence

2. Calculate a "test statistic"
   2.1. Calculate expected frequencies under independence

Independence is equivalent to probabilities that factorise

For example, when *independently* tossing a coin twice:

What is the probability of observing $\{H, H\}$? $\frac{1}{2} \times \frac{1}{2}$

If we repeat the experiment 5 times: how many times we expect to observe $\{H, H\}$? $5 \times \frac{1}{2} \times \frac{1}{2}$

Pearson's χ^2 test for independence

2. Calculate a "test statistic"
   2.1. Calculate expected frequencies under independence

Independence is equivalent to probabilities that factorise

For example, when *independently* tossing a coin twice:

What is the probability of observing $\{H, H\}$? $\frac{1}{2} \times \frac{1}{2}$

If we repeat the experiment 5 times: how many times we expect to observe $\{H, H\}$? $5 \times \frac{1}{2} \times \frac{1}{2}$

This is used to compute **expected frequencies** under independence

Pearson's χ^2 test for independence: an example

2. Calculate a "test statistic"
   2.1. Calculate expected frequencies under independence

Patient group	Tumour shrinkage	
	Yes	No
Treatment	44 (46.1)	40 (37.9)
Placebo	24 (21.9)	16 (18.1)

In this example:

$$46.1 = 124 \times \frac{(44+40)}{124} \times \frac{(44+24)}{124}$$

Pearson's χ^2 test for independence

2. Calculate a "test statistic"

Pearson's χ^2 test for independence

2. Calculate a "test statistic"
   2.2. Compare observed versus expected frequencies

If the difference is *too* large, variables are likely not independent

Pearson's χ^2 test for independence

2. Calculate a "test statistic"
   2.2. Compare observed versus expected frequencies

If the difference is *too* large, variables are likely not independent

The test statistic to be computed is

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

Pearson's χ^2 test for independence

2. Calculate a "test statistic"
   2.2. Compare observed versus expected frequencies

If the difference is *too* large, variables are likely not independent

The test statistic to be computed is

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

If H_0 is true, it is known that

$$\chi^2 \sim \chi^2_{(r-1)\times(c-1)},$$

for a $r \times c$ contingency table.

Pearson's χ^2 test for independence: an example

2. Calculate a "test statistic" (evidence)
   2.2. Compare observed versus expected frequencies

Patient group	Tumour shrinkage	
	Yes	No
Treatment	44 (46.1)	40 (37.9)
Placebo	24 (21.9)	16 (18.1)

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

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

Pearson's χ^2 test for independence: an example

3. Decision rule

Pearson's χ^2 test for independence: an example

3. Decision rule

   Reject H_0 if observed is far from expected: if χ^2 is large

Pearson's χ^2 test for independence: an example

3. Decision rule

Reject H_0 if observed is far from expected: if χ^2 is large

In the example, we have
$\chi^2 = 0.64$, $r = 2$ and $c = 2$

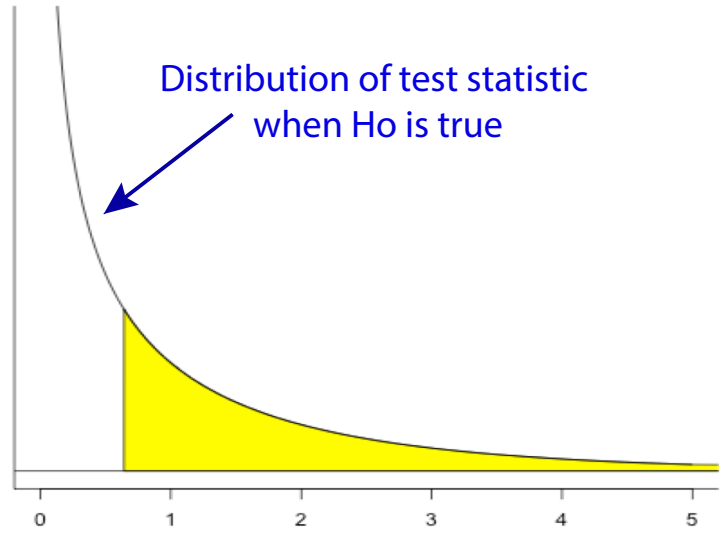
Pearson's χ^2 test for independence: an example

3. Decision rule

Reject H_0 if observed is far from expected: if χ^2 is large

In the example, we have
$\chi^2 = 0.64$, $r = 2$ and $c = 2$

Hence, we can calculate that
the **p-value** is equal to 0.43



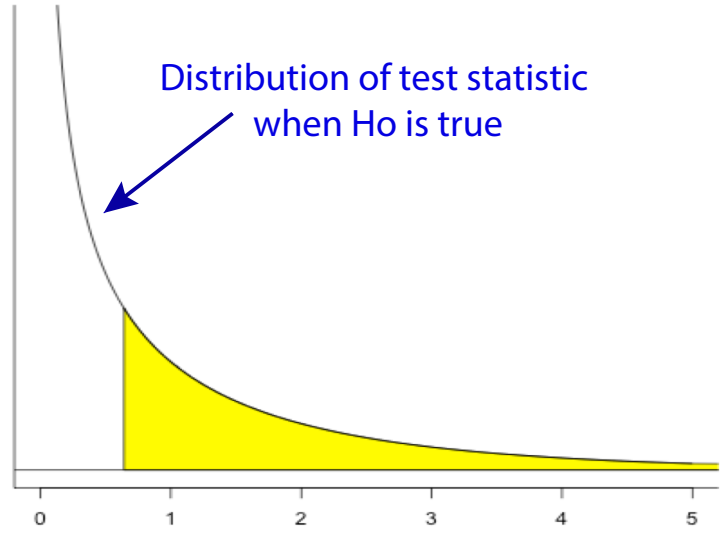
Pearson's χ^2 test for independence: an example

3. Decision rule

Reject H_0 if observed is far from expected: if χ^2 is large

In the example, we have
$\chi^2 = 0.64$, $r = 2$ and $c = 2$

Hence, we can calculate that the **p-value** is equal to 0.43



Do not reject H_0: there is no evidence of an association between treatment group and tumour shrinkage (with 95% significance)

Pearson's χ^2 test for independence

Can we always use this test?

Pearson's χ^2 test for independence

Can we always use this test?

In general, a chi-square test is appropriate when:

- Observations are **independent**, randomly sampled from a population
- None of the cells have an **expected** frequency less than 1
- At least 80% of the cells have an **expected** frequency of 5 or greater

Pearson's χ^2 test for independence

Can we always use this test?

In general, a chi-square test is appropriate when:

- Observations are **independent**, randomly sampled from a population
- None of the cells have an **expected** frequency less than 1
- At least 80% of the cells have an **expected** frequency of 5 or greater

If the last 2 conditions aren't met, **Fisher's exact test** should be used

Questions + practical

Tests for continuous variables

Tests for continuous variables

The χ^2 uses expected frequencies for all possible data values

Tests for continuous variables

The χ^2 uses expected frequencies for all possible data values

This is no longer possible when dealing with **continuous variables**

Tests for continuous variables

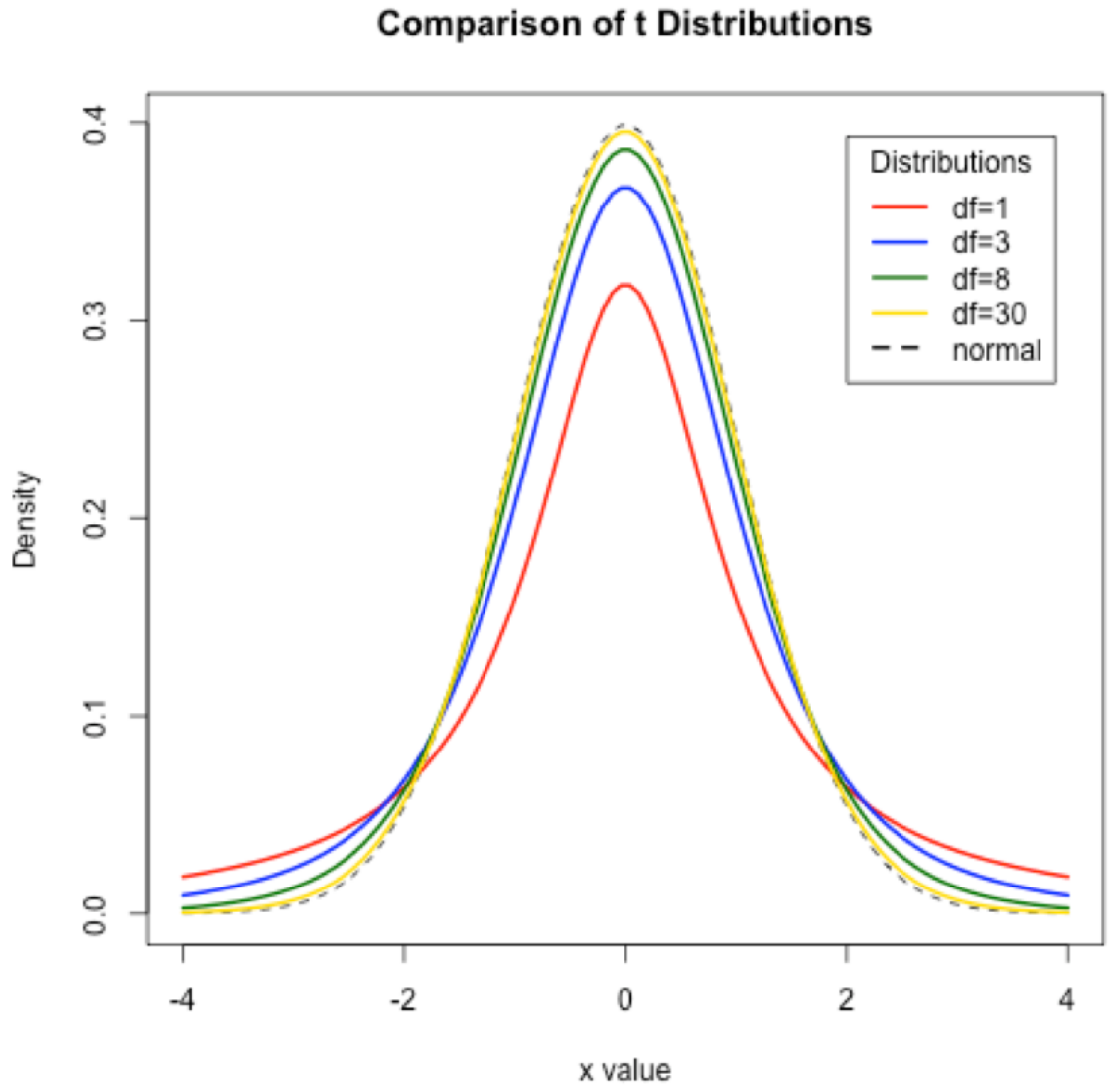
The χ^2 uses expected frequencies for all possible data values

This is no longer possible when dealing with **continuous variables**

Instead, tests for continuous variables focus on global features:

- Parametric tests
- Non-parametric tests

Before we continue: Student's t distribution



Parametric tests

Tests for continuous variables: an example

Published data suggests that the microarray failure rate for a particular supplier is 2.1%.

Genomics Core want to know if this holds true?

Next steps:

1. Formulate the hypothesis
2. Calculate a "test statistic"
3. Decision rule

One-sample t-test

1. Formulate the hypothesis

The one-sample t-test is designed to test hypothesis about the **mean** of a distribution. For example,

$$H_0 : \mu = \mu_0 \text{ vs } H_1 : \mu \neq \mu_0$$

One-sample t-test

1. Formulate the hypothesis

The one-sample t-test is designed to test hypothesis about the **mean** of a distribution. For example,

$$H_0 : \mu = \mu_0 \text{ vs } H_1 : \mu \neq \mu_0$$

In the example:

One-sample t-test

1. Formulate the hypothesis

The one-sample t-test is designed to test hypothesis about the **mean** of a distribution. For example,

$$H_0 : \mu = \mu_0 \text{ vs } H_1 : \mu \neq \mu_0$$

In the example:

H_0: Mean failure rate $= 2.1\%$ vs H_1: Mean failure rate $\neq 2.1\%$

One-sample t-test

2. Calculate a "test statistic" (evidence)

- Idea: how far is $\bar{x} = (\sum_{i=1}^{n} x_i)/n$ from μ_0?
- To quantify this, the following test statistic is defined:

$$t = \frac{\bar{x} - \mu_0}{se(\bar{x})}$$

Recall: $se(\bar{x}) = sd(x)/\sqrt{n}$

If H_0 is true, it is known that $t \sim t_{n-1}$.

One-sample t-test: an example

2. Calculate a "test statistic" (evidence)

Genomics Core recorded the following 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

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

Standard deviation $= 0.84$

Therefore,

$$t = \frac{2.84 - 2.10}{0.84/\sqrt{12}} = 3.07$$

One-sample t-test: an example

3. Decision rule

   Reject H_0 if $\bar{x}$ is far from μ_0: if t is too large or too small

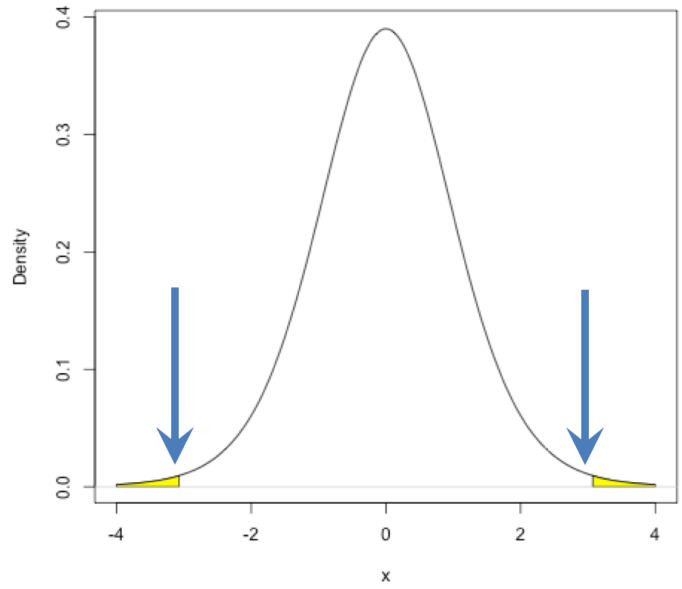
One-sample t-test: an example

3. Decision rule

Reject H_0 if $\bar{x}$ is far from μ_0: if t is too large or too small

In the example, $t = 3.07$

The **p-value** is equal to 0.01



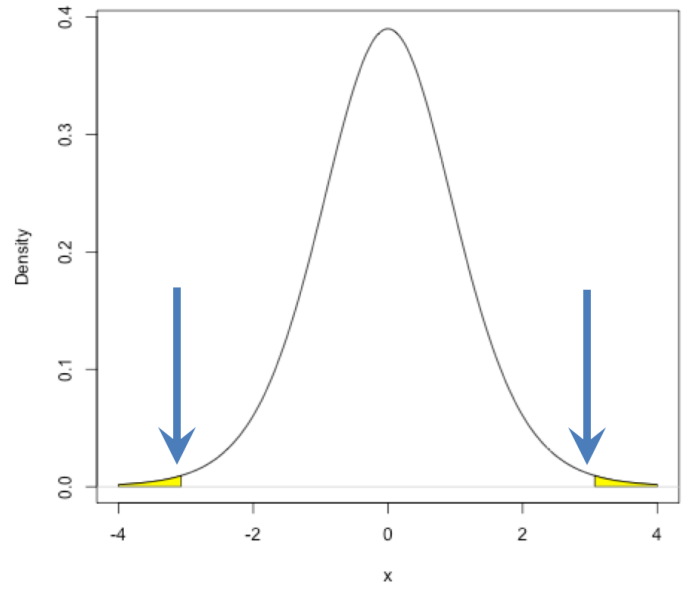
One-sample t-test: an example

3. Decision rule

Reject H_0 if $\bar{x}$ is far from μ_0: if t is too large or too small

In the example, $t = 3.07$

The **p-value** is equal to 0.01



Reject H_0: the data recorded by the Genomics Core suggest the supplier's claim is not true (95% significance)

One-sample t-test

The previous test is a **2-sided** test

The rejection area has *2 tails*

One-sample t-test

The previous test is a **2-sided** test

The rejection area has *2 tails*

In some situations a **1-sided** test is appropriate:

$$H_0 : \mu \geq \mu_0 \text{ vs } H_1 : \mu < \mu_0$$

$$H_0 : \mu \leq \mu_0 \text{ vs } H_1 : \mu > \mu_0$$

One-sample t-test

The previous test is a **2-sided** test

The rejection area has *2 tails*

In some situations a **1-sided** test is appropriate:

$$H_0 : \mu \geq \mu_0 \text{ vs } H_1 : \mu < \mu_0$$

$$H_0 : \mu \leq \mu_0 \text{ vs } H_1 : \mu > \mu_0$$

In the example,

*Is the microarray failure rate **lower** than 2.1%?*

One-sample t-test

The previous test is a **2-sided** test

The rejection area has *2 tails*

In some situations a **1-sided** test is appropriate:

$$H_0 : \mu \geq \mu_0 \text{ vs } H_1 : \mu < \mu_0$$

$$H_0 : \mu \leq \mu_0 \text{ vs } H_1 : \mu > \mu_0$$

In the example,

*Is the microarray failure rate **lower** than 2.1%?*

*Is the microarray failure rate **higher** than 2.1%?*

One-sample t-test

Can we always use this test?

One-sample t-test

Can we always use this test?

The one-sample t-test assumes that:

- Observations are **independent**, randomly sampled from a population
- Observations are **normally distributed**

One-sample t-test

Can we always use this test?

The one-sample t-test assumes that:

- Observations are **independent**, randomly sampled from a population
- Observations are **normally distributed**

What can we do if the data is not normally distributed?

One-sample t-test

Can we always use this test?

The one-sample t-test assumes that:

- Observations are **independent**, randomly sampled from a population
- Observations are **normally distributed**

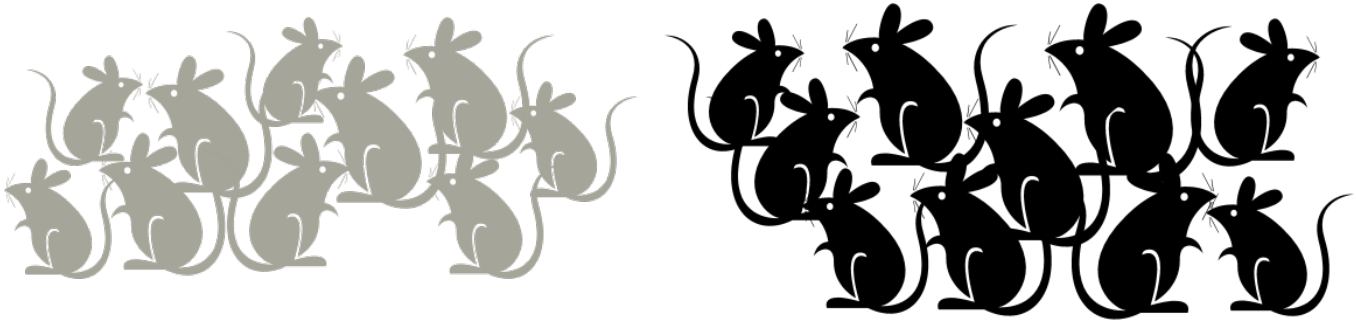
What can we do if the data is not normally distributed?

- If n is large, Central Limit Theorem can allow its use
- We can transform the data (e.g. $\log(X)$)
- We can use a **non-parametric** test

Tests for continuous variables: an example

The weight of 40 male mice (20 of breed A and 20 of breed B) was recorded at 4 weeks old.

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



Next steps:

1. Formulate the hypothesis
2. Calculate a "test statistic"
3. Decision rule

Two-sample t-test

1. Formulate the hypothesis

The two-sample t-test is designed to test hypothesis about the **mean** of two distributions. For example,

$$H_0 : \mu_1 = \mu_2 \text{ vs } H_1 : \mu_1 \neq \mu_2$$

Two-sample t-test

1. Formulate the hypothesis

The two-sample t-test is designed to test hypothesis about the **mean** of two distributions. For example,

$$H_0 : \mu_1 = \mu_2 \text{ vs } H_1 : \mu_1 \neq \mu_2$$

In the example:

Two-sample t-test

1. Formulate the hypothesis

The two-sample t-test is designed to test hypothesis about the **mean** of two distributions. For example,

$$H_0 : \mu_1 = \mu_2 \text{ vs } H_1 : \mu_1 \neq \mu_2$$

In the example:

H_0: Mean weight breed A = Mean weight breed B

H_1: Mean weight breed A $\neq$ Mean weight breed B

Two-sample t-test

2. Calculate a "test statistic"

- Idea: how far are the observed means?
- To quantify this, the following test statistic is defined:

$$t = \frac{\bar{x}_A - \bar{x}_B}{se(\bar{x}_A - \bar{x}_B)},$$

where

- $\bar{x}_A$, $\bar{x}_B$ are the observed means for each group
- $se(\bar{x}_A - \bar{x}_B)$ is the observed standard error for $\bar{x}_A - \bar{x}_B$

Important: to calculate $se(\bar{x}_A - \bar{x}_B)$ and to derive the distribution of t (under H_0), we need to consider whether the groups have:

- Equal or unequal variances

One-sample t-test: an example

2. Calculate a "test statistic" (evidence)

The following data was recorded

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

Assumming **unequal variances** we obtain $t = 1.21$

Two-sample t-test: an example

3. Decision rule

   Reject H_0 if $\bar{x}_A$ is far from $\bar{x}_B$: if t is too large or too small

Two-sample t-test: an example

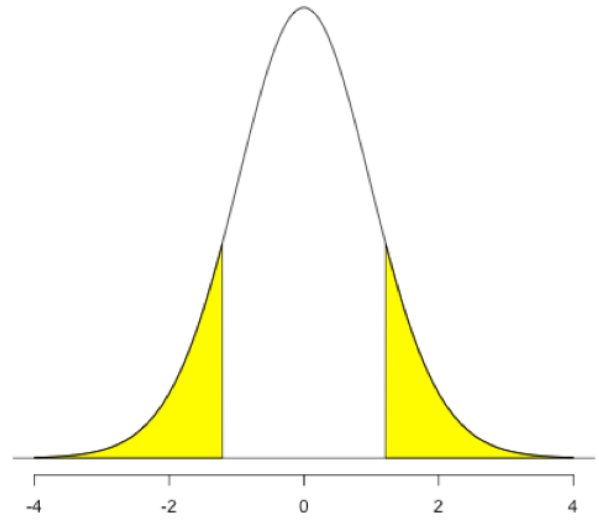
3. Decision rule

Reject H_0 if $\bar{x}_A$ is far from $\bar{x}_B$: if t is too large or too small

In the example, $t = 1.21$

The **p-value** is equal to 0.24

(Assumes **unequal variances**)



Two-sample t-test: an example

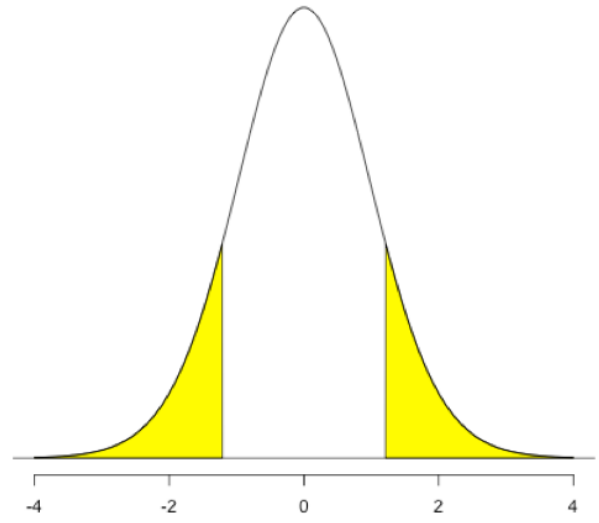
3. Decision rule

Reject H_0 if $\bar{x}_A$ is far from $\bar{x}_B$: if t is too large or too small

In the example, $t = 1.21$

The **p-value** is equal to 0.24

(Assumes **unequal variances**)



Do not reject H_0: there is no evidence to believe that the mean weights of breed A and breed B differ (95% significance)

Two-sample t-test

The previous test is a **2-sided** test

The rejection area has *2 tails*

Two-sample t-test

The previous test is a **2-sided** test

The rejection area has *2 tails*

In some situations a **1-sided** test is appropriate:

$$H_0 : \mu_A \geq \mu_B \text{ vs } H_1 : \mu_A < \mu_B$$

$$H_0 : \mu_A \leq \mu_B \text{ vs } H_1 : \mu_A > \mu_B$$

Two-sample t-test

The previous test is a **2-sided** test

The rejection area has *2 tails*

In some situations a **1-sided** test is appropriate:

$$H_0 : \mu_A \geq \mu_B \text{ vs } H_1 : \mu_A < \mu_B$$

$$H_0 : \mu_A \leq \mu_B \text{ vs } H_1 : \mu_A > \mu_B$$

In the example,

*Is the weight of breed A **lower** than of breed B?*

Two-sample t-test

The previous test is a **2-sided** test

The rejection area has *2 tails*

In some situations a **1-sided** test is appropriate:

$$H_0 : \mu_A \geq \mu_B \text{ vs } H_1 : \mu_A < \mu_B$$

$$H_0 : \mu_A \leq \mu_B \text{ vs } H_1 : \mu_A > \mu_B$$

In the example,

*Is the weight of breed A **lower** than of breed B?*

*Is the weight of breed A **higher** than of breed B?*

Two-sample t-test

Can we always use this test?

Two-sample t-test

Can we always use this test?

The two-sample t-test assumes that

- **Within** each group, observations are **independent** and randomly drawn from a population
- Observations are independent **between** the groups
- Observations are normally distributed **within** each group

Two-sample t-test

Can we always use this test?

The two-sample t-test assumes that

- **Within** each group, observations are **independent** and randomly drawn from a population
- Observations are independent **between** the groups
- Observations are normally distributed **within** each group

What can we do if the data is not normally distributed?

Two-sample t-test

Can we always use this test?

The two-sample t-test assumes that

- **Within** each group, observations are **independent** and randomly drawn from a population
- Observations are independent **between** the groups
- Observations are normally distributed **within** each group

What can we do if the data is not normally distributed?

- If n is large, Central Limit Theorem can allow its use
- We can transform the data (e.g. $\log(X)$)
- We can use a **non-parametric** test

Two-sample t-test

What can we do if there is no independence between groups?

Two-sample t-test

What can we do if there is no independence between groups?

- We should used a **paired** two sample t-test

Assumption: pairs of samples are randomly drawn from a population

Two-sample t-test

What can we do if there is no independence between groups?

- We should used a **paired** two sample t-test

Assumption: pairs of samples are randomly drawn from a population

Examples where a paired test should be used:

Two-sample t-test

What can we do if there is no independence between groups?

- We should used a **paired** two sample t-test

Assumption: pairs of samples are randomly drawn from a population

Examples where a paired test should be used:

- Before / after measures
- Twin studies

Questions + practical

Non-parametric tests

Non-parametric tests

The tests for continues variables described earlier formulate hypothesis in terms of **parameters** (e.g. the mean μ)

Typically, **parametric** tests assume the data follows a particular distribution (e.g. normal) or rely on assymtotic results (e.g. $\bar{x}$ is normally distributed for large n)

Non-parametric tests

The tests for continues variables described earlier formulate hypothesis in terms of **parameters** (e.g. the mean μ)

Typically, **parametric** tests assume the data follows a particular distribution (e.g. normal) or rely on assymtotic results (e.g. $\bar{x}$ is normally distributed for large n)

Non-parametric tests provide an alternative for situations where these assumptions do not hold

One-sample sign test

The **one-sample sign test** can be interpreted as a non-parametric version of the **one-sample t-test**

It defines hypothesis in terms of the **median** of a distribution

H_0: median is equal to θ_0

H_1: median is not equal to θ_0

As in the one-sample t-test, this assumes **independent** observations

One-sample sign test

The test statistic is calculated by comparing the observations to θ_0:

- $+$: if bigger
- $-$: if smaller
- $=$: if equal

Count the number of $+$'s and $-$'s, and calculate:

- $S^+ =$ the number of $+$'s
- $S^- =$ the number of $-$'s
- $n =$ the number of non-ties

If H_0 is true, $S^+, S^- \sim \text{Binomial}(n, 0.5)$

One-sample sign test: one example

4
1
19
8
12
15
26
4
23
19
12
1
9
25
3

15 people were asked how many times they went to the cinema during the last year

A survey from two years ago indicated that a typical person goes to the cinema 5 times per year. Is this still true?

One-sample sign test: one example

4
1
19
8
12
15
26
4
23
19
12
1
9
25
3

15 people were asked how many times they went to the cinema during the last year

A survey from two years ago indicated that a typical person goes to the cinema 5 times per year. Is this still true?

If perform a one-sample t-test for

$$H_0 : \mu = 5 \text{ versus } H_1 : \mu \neq 5,$$

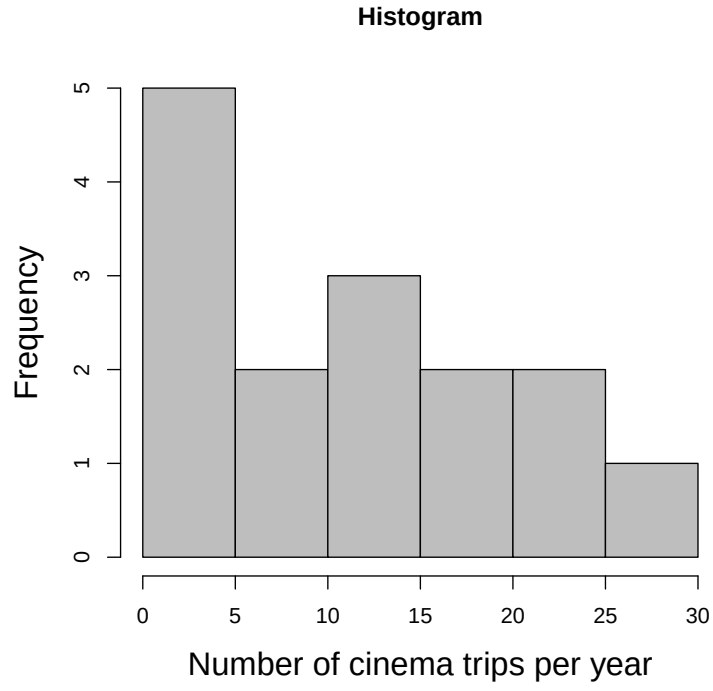
the p-value is equal to 0.007.

Therefore, the data suggests that this information is no longer true.

One-sample sign test: one example

4
1
19
8
12
15
26
4
23
19
12
1
9
25
3

Let's look at the data again



The data does not look normal!

One-sample sign test: one example

4
1
19
8
12
15
26
4
23
19
12
1
9
25
3

Instead, we can use a one-sample sign test

One-sample sign test: one example

4	-
1	-
19	+
8	+
12	+
15	+
26	+
4	-
23	+
19	+
12	+
1	-
9	+
25	+
3	-

Instead, we can use a one-sample sign test

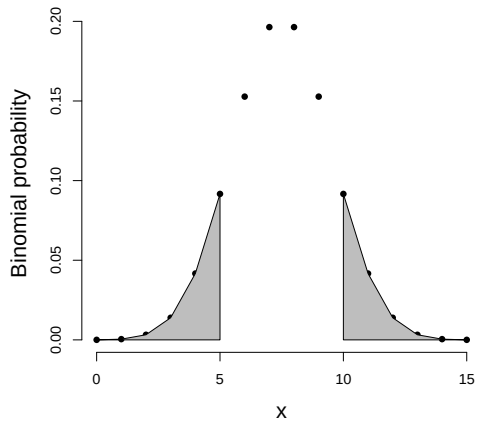
- $S^+ = 10$ (number of observations > 5)
- $S^- = 5$ (number of observations < 5)
- $n = 15$ (number of observations $\neq 5$)

One-sample sign test: one example

4	-
1	-
19	+
8	+
12	+
15	+
26	+
4	-
23	+
19	+
12	+
1	-
9	+
25	+
3	-

Instead, we can use a one-sample sign test

- $S^+ = 10$ (number of observations > 5)
- $S^- = 5$ (number of observations < 5)
- $n = 15$ (number of observations $\neq 5$)



The p-value is equal to 0.301

Recall: t-test p-value $= 0.007$

There is no evidence to suggest the information is no longer true.

Two-sample sign test

The **two-sample sign test** can be interpreted as a non-parametric version of the **two-sample paired t-test**

It defines hypothesis in terms of the **median** of the distributions

H_0: median is equal in both groups

H_1: median is not equal in both groups

As in the two-sample paired t-test, this assumes **independent pairs** of observations

Two-sample sign test

The test statistic is calculated by comparing observed values between groups:

- $+$: if bigger in the first group
- $-$: if smaller in the second group
- $=$: if equal

Count the number of $+$'s and $-$'s, and calculate:

- $S^+ =$ the number of $+$'s
- $S^- =$ the number of $-$'s
- $n =$ the number of non-ties

If H_0 is true, $S^+, S^- \sim \text{Binomial}(n, 0.5)$

Two-sample sign test: an example

General health section of SF-36 was collected in a breast cancer study. Data includes information for two time points (same patients)

Is there a difference between the time points?

H_0 : medians of the two time points are the same

H_1 : medians of the two time points are not the same

Two-sample sign test: an example

GH Value 1	GH Value 2
60	70
55	65
75	100
100	50
55	70
60	95
50	95
60	65
72	85
40	55
90	95
75	45
70	75
75	65
55	60

Two-sample sign test: an example

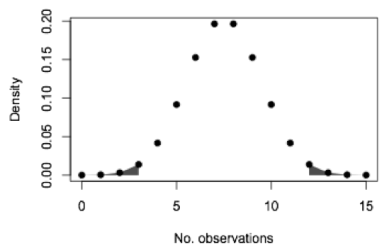
GH Value 1	GH Value 2	Difference
60	70	-10
55	65	-10
75	100	-25
100	50	50
55	70	-15
60	95	-35
50	95	-45
60	65	-5
72	85	-13
40	55	-15
90	95	-5
75	45	30
70	75	-5
75	65	10
55	60	-5

Two-sample sign test: an example

GH Value 1	GH Value 2	Difference	Sign
60	70	-10	-
55	65	-10	-
75	100	-25	-
100	50	50	+
55	70	-15	-
60	95	-35	-
50	95	-45	-
60	65	-5	-
72	85	-13	-
40	55	-15	-
90	95	-5	-
75	45	30	+
70	75	-5	-
75	65	10	+
55	60	-5	-

- $S^+ = 3$ (no. pairs where diff > 0)
- $S^- = 12$ (no. pairs where diff < 0)
- $n = 15$ (no. pairs where diff $\neq 0$)

p-value $= 0.035$



The data suggests medians differ, i.e. there is a difference in general health between the two time points

Advantages and limitations of sign tests

- Simple
  - Few assumptions thus widely applicable
- Less powerful than other tests
  - Does not consider magnitude of differences
  - May fail to reject H_0 when other tests would
- Can be used for quick assessment of direction

Wilcoxon signed rank test

The **Wilcoxon signed rank test** is an alternative to the **two-sample sign test** (uses the magnitude of the differences rather than just the sign)

In addition to the assumptions of the two-sample sign test, it also assumes:

symmetry of differences around true median difference

It's designed to contrast the hypothesis:

H_0: distribution of paired differences is symmetric around zero

H_1: distribution of paired differences is not symmetric around zero

Wilcoxon signed rank test

Method:

- Calculate differences for each pair
- Rank the paired differences by (absolute) magnitude
- Split the ranks into two groups: positive and negative signed
  - Calculate sum of positive ranks: W^+
  - Calculate sum of negative ranks: W^-
- Compare smaller of and W^+ and W^- to the critical value from the tables

Wilcoxon signed rank test: an example

Revisiting the previous example:

GH Value 1	GH Value 2	Difference	Sign
60	70	-10	-
55	65	-10	-
75	100	-25	-
100	50	50	+
55	70	-15	-
60	95	-35	-
50	95	-45	-
60	65	-5	-
72	85	-13	-
40	55	-15	-
90	95	-5	-
75	45	30	+
70	75	-5	-
75	65	10	+
55	60	-5	-

1. Record sign of the differences

Wilcoxon signed rank test: an example

Revisiting the previous example:

GH Value 1	GH Value 2	Difference	Sign	Abs.Diff.
60	70	-10	-	10
55	65	-10	-	10
75	100	-25	-	25
100	50	50	+	50
55	70	-15	-	15
60	95	-35	-	35
50	95	-45	-	45
60	65	-5	-	5
72	85	-13	-	13
40	55	-15	-	15
90	95	-5	-	5
75	45	30	+	30
70	75	-5	-	5
75	65	10	+	10
55	60	-5	-	5

1. Record sign of the differences
2. Record absolute differences

Wilcoxon signed rank test: an example

Revisiting the previous example:

GH Value 1	GH Value 2	Difference	Sign	Abs.Diff.
60	65	-5	-	5
90	95	-5	-	5
70	75	-5	-	5
55	60	-5	-	5
60	70	-10	-	10
55	65	-10	-	10
75	65	10	+	10
72	85	-13	-	13
55	70	-15	-	15
40	55	-15	-	15
75	100	-25	-	25
75	45	30	+	30
60	95	-35	-	35
50	95	-45	-	45
100	50	50	+	50

1. Record sign of the differences
2. Record absolute differences
3. Order observations according to absolute difference

Wilcoxon signed rank test: an example

Revisiting the previous example:

GH Value 1	GH Value 2	Difference	Sign	Abs.Diff.	Rank
60	65	-5	-	5	2.5
90	95	-5	-	5	2.5
70	75	-5	-	5	2.5
55	60	-5	-	5	2.5
60	70	-10	-	10	6
55	65	-10	-	10	6
75	65	10	+	10	6
72	85	-13	-	13	8
55	70	-15	-	15	9.5
40	55	-15	-	15	9.5
75	100	-25	-	25	11
75	45	30	+	30	12
60	95	-35	-	35	13
50	95	-45	-	45	14
100	50	50	+	50	15

1. Record sign of the differences
2. Record absolute differences
3. Order observations according to absolute difference
4. Calculate ranks
   - Ties $\rightarrow$ average rank

Wilcoxon signed rank test: an example

Revisiting the previous example:

GH Value 1	GH Value 2	Difference	Sign	Abs.Diff.	Rank	Signed-Rank
60	65	-5	-	5	2.5	-2.5
90	95	-5	-	5	2.5	-2.5
70	75	-5	-	5	2.5	-2.5
55	60	-5	-	5	2.5	-2.5
60	70	-10	-	10	6	-6
55	65	-10	-	10	6	-6
75	65	10	+	10	6	6
72	85	-13	-	13	8	-8
55	70	-15	-	15	9.5	-9.5
40	55	-15	-	15	9.5	-9.5
75	100	-25	-	25	11	-11
75	45	30	+	30	12	12
60	95	-35	-	35	13	-13
50	95	-45	-	45	14	-14
100	50	50	+	50	15	15

1. Record sign of the differences
2. Record absolute differences
3. Order observations according to absolute difference
4. Calculate ranks
   - Ties $\rightarrow$ average rank
5. Calculate signed ranks

Wilcoxon signed rank test: an example

Revisiting the previous example:

GH Value 1	GH Value 2	Difference	Sign	Abs.Diff.	Rank	Signed-Rank
60	65	-5	-	5	2.5	-2.5
90	95	-5	-	5	2.5	-2.5
70	75	-5	-	5	2.5	-2.5
55	60	-5	-	5	2.5	-2.5
60	70	-10	-	10	6	-6
55	65	-10	-	10	6	-6
75	65	10	+	10	6	6
72	85	-13	-	13	8	-8
55	70	-15	-	15	9.5	-9.5
40	55	-15	-	15	9.5	-9.5
75	100	-25	-	25	11	-11
75	45	30	+	30	12	12
60	95	-35	-	35	13	-13
50	95	-45	-	45	14	-14
100	50	50	+	50	15	15

Here we obtain

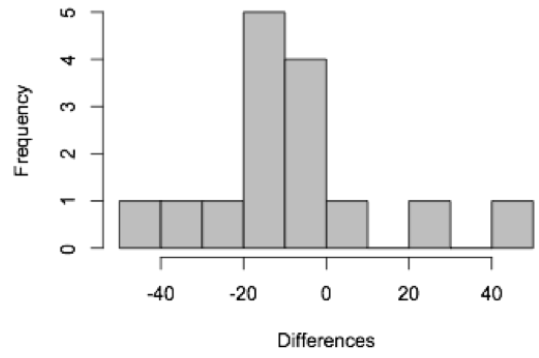
- $W^+ = 33$
- $W^- = 87$

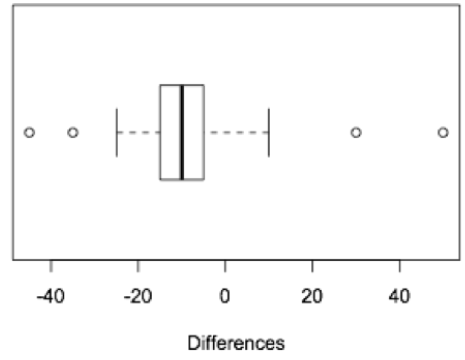
p-value $= 0.12$

No evidence to conclude that there is a difference in general health between the two time points

Wilcoxon signed rank test: an example

Validity of assumptions may affect results





Wilcoxon signed rank test

- Easy to apply
- More powerful than the sign test
- Less powerful than the paired t-test
- Relies on the symmetry assumption

Other popular tests

Here, we only explored a small subset of the tests that are available in standard statistical software. Here is a summary of other alternatives:

		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

Questions + practical

Acknowledgements

Part of the materials discussed today have been adapted from materials provided by

- Mark Dunning (CRUK-CI)