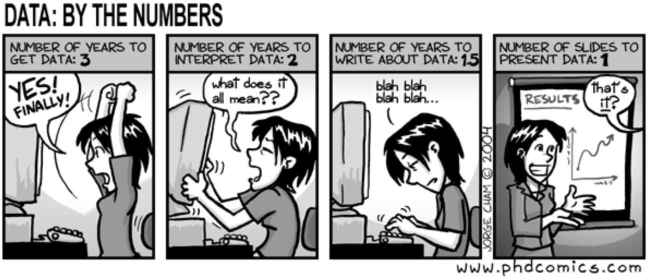
Introduction to statistical thinking

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

April 4th, 2017

Why statistics?



Why statistics?

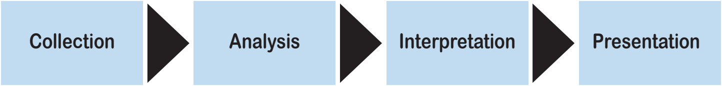
Statistics is at the core of modern research

What is statistics?

Why statistics?

Statistics is at the core of modern research

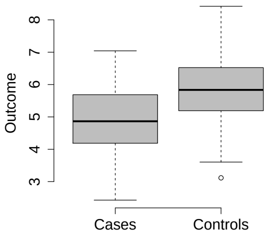
What is statistics?



$\rightarrow$ everything from experimental design to figure preparation!

What can we do with statistics?

Comparisons between two groups



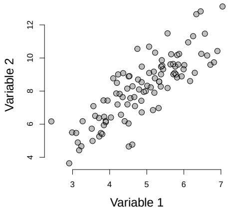
For example,

- Case/control studies
- Differential expression studies

Is this difference *statistically significant*?

What can we do with statistics?

Finding associations between variables



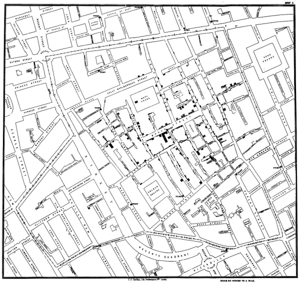
For example,

- Dose level vs survival time
- Association studies (e.g. GWAS)
- Co-expression networks

Is there an association between X and Y?

Which associations are *statistically significant*?

What can we do with statistics?



Public domain image

During the Soho cholera outbreak in 1854, John Snow used statistics to find an **association** between the quality of the water source and cholera cases

John Snow is considered one of the fathers of modern epidemiology

Outline for the course

- **Introduction to statistics**
  - Types of data
  - Descriptive statistics
  - Practical: descriptive statistics in Rmarkdown
  - Statistical inference
- **Hypothesis testing (in R)**
- **Regression analysis (in R)**
- **Multiple testing (in R)**

Introduction to statistics

Types of data

Prior to any analysis, it is important to know what type of data is available

Types of data

Prior to any analysis, it is important to know what type of data is available

Data can be classified into different **types of variables**:

- Categorical (nominal)
- Categorical (ordinal)
- Discrete
- Continuous

Types of data: categorical (nominal)

To describe categories

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

- No logical order
- Mutually exclusive fixed categories

Examples: gender, yes/no, cancer type, eye colour, ethnicity, etc.

Types of data: categorical (ordinal)

To describe categories

- There is a logical order
- Mutually exclusive fixed categories
- There is no quantification of the distance between categories

Examples: stress level ($1 = \text{low}, \ldots, 7 = \text{high}$), pain level (low/medium/high), education level (primary, secondary, ...), etc.

Types of data: discrete

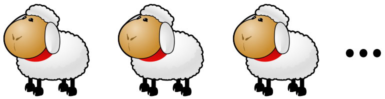


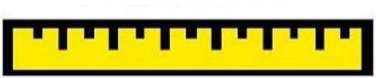
Image modified from Free Stock Photos

- Fixed (countable) possible set of values
- Distance between categories can be quantified
- ... basically, anything counted

Examples: number of tumours, hospital admissions, etc.

Sometimes treated as continuous if range is large!

Types of data: continuous



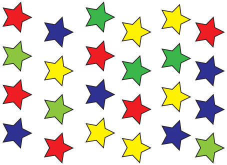
Requirements:

- Infinite number of possible values
- Given any two values, one fits between
- May have finite or infinite range

Examples: height, weight, blood presure, temperature etc.

Population versus sample

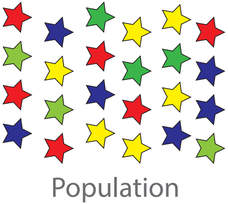
Census records information from every subject within a population



Population

Population versus sample

Census records information from every subject within a population

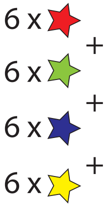


Scientific studies typically rely on a subset of a population

Assumption: the sample is a good representative of the population!

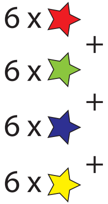
Descriptive statistics versus statistical inference

Descriptive statistics
summarizes the information
available in the data

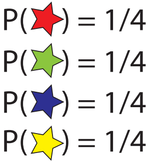


Descriptive statistics versus statistical inference

Descriptive statistics
summarizes the information available in the data



Statistical inference
extrapolates sample information to population level



Assumption: the sample is a good representative of the population!

Descriptive analyses

Descriptive analyses

There are two basic forms of data summary

- Numerical: frequency tables, summary measures, etc
- Graphical: histograms, scatter-plots, boxplots, etc

A good summary depends on the data type!

Descriptive analyses: categorical variables

Numerical summary

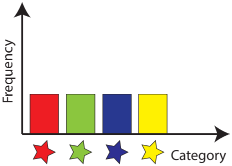
Category	Frequency	Cumulative frequency	Relative frequency	Cum. rel. frequency
	6	6	0.25	0.25
	6	12	0.25	0.50
	6	18	0.25	0.75
	6	24	0.25	1.00

Descriptive analyses: categorical variables

Numerical summary

Category	Frequency	Cumulative frequency	Relative frequency	Cum. rel. frequency
	6	6	0.25	0.25
	6	12	0.25	0.50
	6	18	0.25	0.75
	6	24	0.25	1.00

Graphical summary



Descriptive analysis: discrete/continuous variables

Frequency tables are not useful
unless we *categorize* the data

Descriptive analysis: discrete/continuous variables

Frequency tables are not useful unless we *categorize* the data

Example:

Birth weight	Rel. freq
Very low (<1.5 kg.)	0.05
Low (1.5-2.5 kg.)	0.10
Normal ($\geq$2.5 kg.)	0.85

Descriptive analysis: discrete/continuous variables

Frequency tables are not useful unless we *categorize* the data

Example:

Birth weight	Rel. freq
Very low (<1.5 kg.)	0.05
Low (1.5-2.5 kg.)	0.10
Normal (≥2.5 kg.)	0.85

Instead, we typically prefer *summary measures*

- Mean
- Median
- Mode
- Variance
- Quantiles

Which one?

Descriptive analysis: discrete/continuous variables

Firstly, we can use measures to describe the **location** of the data

Descriptive analysis: discrete/continuous variables

Firstly, we can use measures to describe the **location** of the data

Mean ($\bar{x}$)

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Median ($x_{50\%}$)

50% of observations lie below $x_{50\%}$

Descriptive analysis: discrete/continuous variables

Firstly, we can use measures to describe the **location** of the data

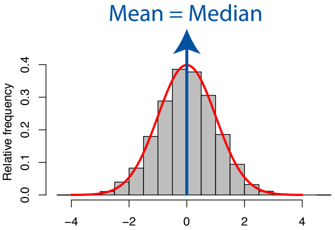
Mean ($\bar{x}$)

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Median ($x_{50\%}$)

50% of observations lie below $x_{50\%}$

When the data is normal ...



... but that is not always true

Descriptive analysis: example

Suposse we record the number of Facebook friends for 7 colleagues:

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

Mean ($\bar{x}$)

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_7}{7} = 970$$

Median ($x_{50\%}$)

$$11, 243, 270, {\color{red}310}, 311, 345, 5300 \Rightarrow x_{50\%} = 310$$

Which one provides a better description for the location of the data?

Descriptive analysis: example

Now suppose the data is slighty different:

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

Mean ($\bar{x}$)

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_7}{7} = 289$$

Median ($x_{50\%}$)

$$11, 243, 270, 310, 311, 345, 530 \Rightarrow x_{50\%} = 310$$

What happened?

Descriptive analysis: discrete/continuous variables

It is also important to summarize the **spread** of the data

Standard deviation (sd(x))

$$\mathrm{sd}(x) = \sqrt{\frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}}$$

Interquartile range (IQR(x))

$$\mathrm{IQR}(x) = x_{75\%} - x_{25\%}$$

Descriptive analysis: example

In the first example, we have

Standard deviation ($\mathrm{sd}(x)$)

$$\mathrm{sd}(x) = \sqrt{\frac{(x_1 - \bar{x})^2 + \cdots + (x_7 - \bar{x})^2}{7}} = 1912.57$$

Interquartile range ($\mathrm{IQR}(x)$)

$$11, 243, 270, 310, 311, 345, 5300 \Rightarrow \mathrm{IQR}(x) = 345 - 243$$

Descriptive analysis: example

In the second example, we have

Standard deviation (sd(x))

$$\mathrm{sd}(x) = \sqrt{\frac{(x_1 - \bar{x})^2 + \cdots + (x_7 - \bar{x})^2}{7}} = 153.79$$

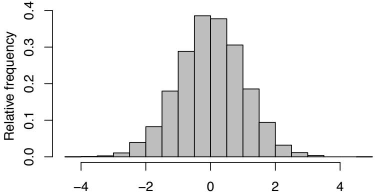
Interquartile range (IQR(x))

$$11, \color{red}{243}, 270, \color{red}{310}, 311, \color{red}{345}, 530 \Rightarrow \mathrm{IQR}(x) = 345 - 243$$

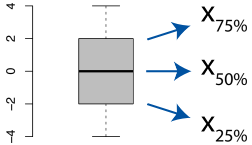
Descriptive analysis: discrete/continuous variables

We can also summarize the distribution of the data **graphically**

Histogram



Boxplot



Before we continue: the normal distribution

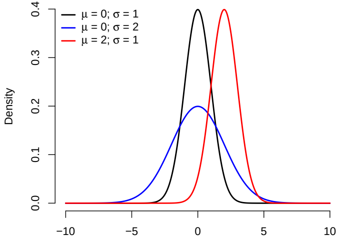
The normal (or Gaussian) distribution is very popular!

Indexed by two parameters:

- $\mu \rightarrow$ location (mean)
- σ (or σ^2) $\rightarrow$ spread (standard deviation)

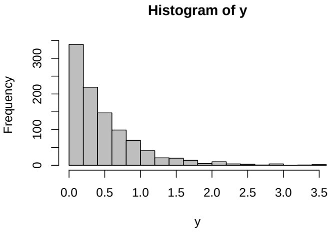
Rule of thumb:

- $\sim$ 95% of the data lies between $(\mu - 2\sigma, \mu + 2\sigma)$
- $\sim$ 99% of the data lies between $(\mu - 3\sigma, \mu + 3\sigma)$

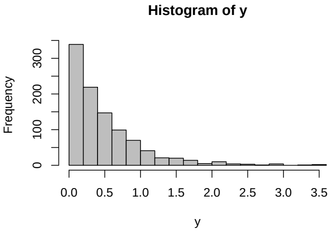


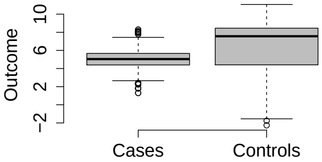
The distribution is symmetric around μ!

Descriptive analysis: is the data normally distributed?

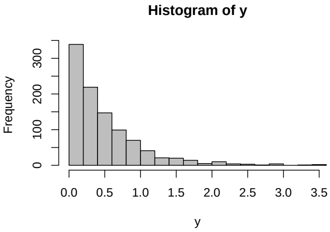


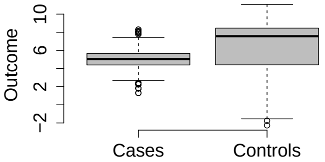
Descriptive analysis: is the data normally distributed?

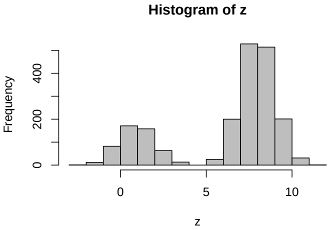




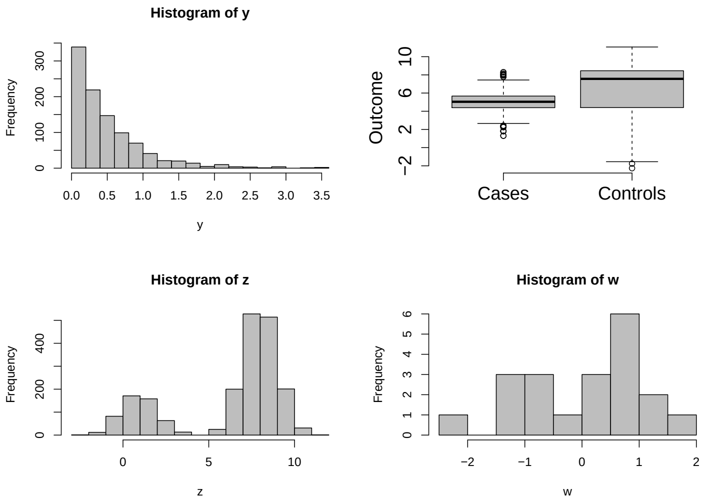
Descriptive analysis: is the data normally distributed?







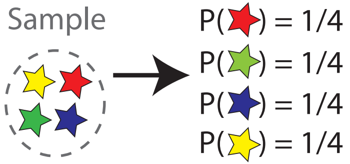
Descriptive analysis: is the data normally distributed?



Statistical inference

Statistical inference: estimation

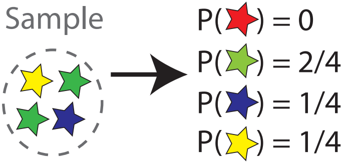
Statistical inference
extrapolates sample information
to population level



In this example, we want to **estimate** the proportion of red, green, blue and yellow stars in the population based on the observed data

Statistical inference: estimation

Statistical inference
extrapolates sample information
to population level

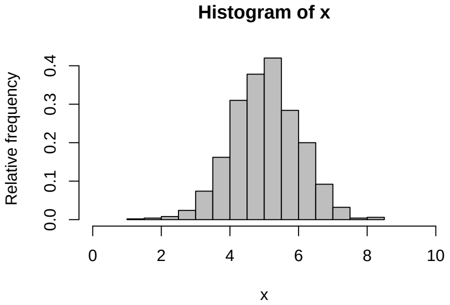


If we observe a different sample,
we obtain different results

Note: in this sample we didn't
observe red stars!

Statistical inference: estimation

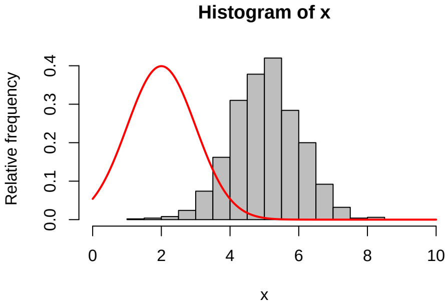
To model the data (e.g. expression of a gene across samples) as being normally distributed with mean μ and standard deviation 1.



The aim here is to estimate the value of μ!

Statistical inference: estimation

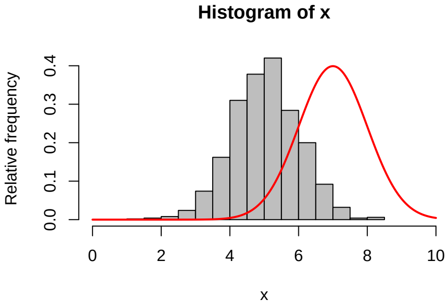
To model the data (e.g. expression of a gene across samples) as being normally distributed with mean μ and standard deviation 1.



$$\mu = 2?$$

Statistical inference: estimation

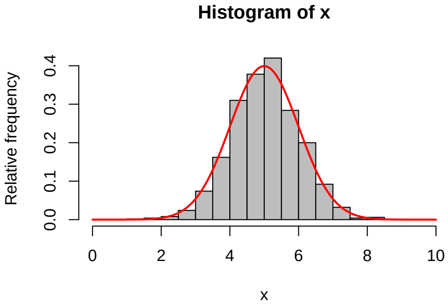
To model the data (e.g. expression of a gene across samples) as being normally distributed with mean μ and standard deviation 1.



$$\mu = 7?$$

Statistical inference: estimation

To model the data (e.g. expression of a gene across samples) as being normally distributed with mean μ and standard deviation 1.



$$\mu = 5?$$

Statistical inference: estimation

We can estimate μ using the observed data through

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n}$$

- How does $\bar{x}$ compare to μ for different sample sizes n?
- If the data is normally distributed, with standard deviation σ: how does $\bar{x}$ compare to μ for different values of σ?

http://bioinformatics.cruk.cam.ac.uk/apps/winter-school/estimation/

Statistical inference: estimation

We can estimate μ using the observed data through

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n}$$

- How does $\bar{x}$ compare to μ for different sample sizes n?
- If the data is normally distributed, with standard deviation σ: how does $\bar{x}$ compare to μ for different values of σ?

http://bioinformatics.cruk.cam.ac.uk/apps/winter-school/estimation/

- The larger the sample size, $\bar{x}$ gets closer to μ
- The smaller σ, the faster $\bar{x}$ gets closer to μ

Statistical inference: estimation

We can estimate μ using the observed data through

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n}$$

Estimates of μ that are based on a **finite sample** are not exact:

we need to provide a measure of uncertainty

Statistical inference: estimation

We can estimate μ using the observed data through

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n}$$

Estimates of μ that are based on a **finite sample** are not exact:

we need to provide a measure of uncertainty

In this case, this can be quantified through the **standard error**:

$$\text{S.E.} = \sigma/\sqrt{n}$$

Statistical inference: estimation

We can estimate μ using the observed data through

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n}$$

Estimates of μ that are based on a **finite sample** are not exact:

we need to provide a measure of uncertainty

In this case, this can be quantified through the **standard error**:

$$\text{S.E.} = \sigma/\sqrt{n}$$

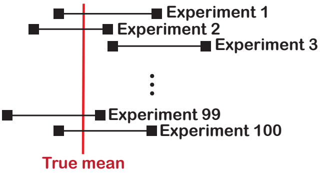
This is consistent with our previous observations

Important: standard error $\neq$ standard deviation

Statistical inference: confidence intervals

A confidence interval (CI) is a *random* interval

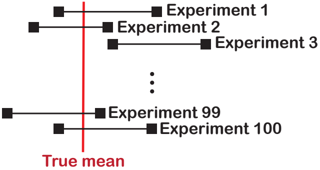
In repeated experiments ...
95% of the time CI covers the *true* mean



Statistical inference: confidence intervals

A confidence interval (CI) is a *random* interval

In repeated experiments ...
95% of the time CI covers the *true* mean



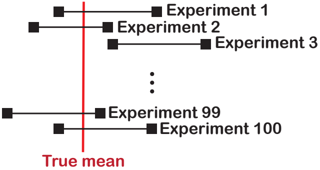
For the mean, the 95% CI is given by

$$(\bar{x} - 1.96 \text{ S.E.}, \bar{x} + 1.96 \text{ S.E.})$$

Statistical inference: confidence intervals

A confidence interval (CI) is a *random* interval

In repeated experiments ...
95% of the time CI covers the *true* mean



For the mean, the 95% CI is given by

$$(\bar{x} - 1.96 \text{ S.E.}, \bar{x} + 1.96 \text{ S.E.})$$

http://bioinformatics.cruk.cam.ac.uk/apps/winter-school/estimation/

Statistical inference: Central Limit Theorem (CLT)

This formula for the CI assumes the data is **normally distributed**.

What can we do if this is not true?

Statistical inference: Central Limit Theorem (CLT)

This formula for the CI assumes the data is **normally distributed**.

What can we do if this is not true?

If the data is not normally distributed, the **CLT** guarantees this result is still valid, provided the sample size is large:

CLT guarantees that $\bar{x}$ is normally distributed

http://bioinformatics.cruk.cam.ac.uk/apps/winter-school/clt/

Statistical inference: hypothesis testing

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

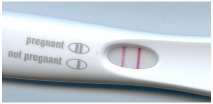
For example,

- *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.

Statistical inference: hypothesis testing

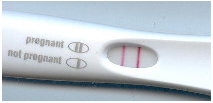


Am I pregnant?

Disclaimer: this image was downloaded from the internet and does not reflect the life of the instructors!

Statistical inference: hypothesis testing

Basic setup



Am I pregnant?

Disclaimer: this image was downloaded from the internet and does not reflect the life of the instructors!

Formulate the ‘null’ and alternative hypothesis

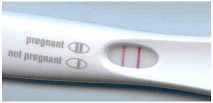
↓

Calculate a “test statistic” from the data

↓

Desicion rule

Statistical inference: hypothesis testing



Am I pregnant?

Disclaimer: this image was downloaded from the internet and does not reflect the life of the instructors!

Basic setup (example)

Formulate the 'null' and alternative hypothesis

e.g. H_o: Not pregnant vs H_1: Pregnant

Calculate a "test statistic" from the data

e.g. summary measure based on hormonal levels

Desicion rule

e.g. is the data more extreme than what is expected by chance (for a non-pregnant woman)?

Statistical inference: hypothesis testing

No test is exact:

	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

Statistical inference: hypothesis testing

No test is exact:

	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

Error rates depend on e.g. sample size … what else?

Statistical inference: hypothesis testing

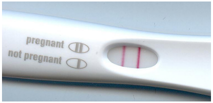
No test is exact:

	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

Error rates depend on e.g. sample size ... what else?

Beware of multiple testing correction issues!

Statistical inference: hypothesis testing



Am I pregnant?

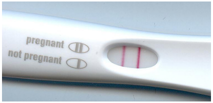
Suppose that 100 women take the test

- 70 of them were trully pregnant
- the test was positive for 75 women
- the result was wrong for 10 non-pregnant women

Complete the table:

	Null hypothesis does not hold	Null hypothesis holds
Reject null hypothesis		
Do not reject null hypothesis		

Statistical inference: hypothesis testing



Am I pregnant?

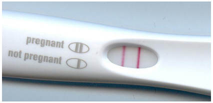
Suppose that 100 women take the test

- 70 of them were trully pregnant
- the test was positive for 75 women
- the result was wrong for 10 non-pregnant women

Complete the table:

	Null hypothesis does not hold	Null hypothesis holds
Reject null hypothesis		10/30
Do not reject null hypothesis		

Statistical inference: hypothesis testing



Am I pregnant?

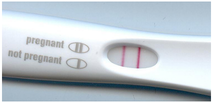
Suppose that 100 women take the test

- 70 of them were trully pregnant
- the test was positive for 75 women
- the result was wrong for 10 non-pregnant women

Complete the table:

	Null hypothesis does not hold	Null hypothesis holds
Reject null hypothesis		10/30
Do not reject null hypothesis		20/30

Statistical inference: hypothesis testing



Am I pregnant?

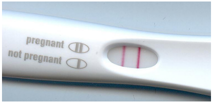
Suppose that 100 women take the test

- 70 of them were trully pregnant
- the test was positive for 75 women
- the result was wrong for 10 non-pregnant women

Complete the table:

	Null hypothesis does not hold	Null hypothesis holds
Reject null hypothesis	65/70	10/30
Do not reject null hypothesis		20/30

Statistical inference: hypothesis testing



Am I pregnant?

Suppose that 100 women take the test

- 70 of them were trully pregnant
- the test was positive for 75 women
- the result was wrong for 10 non-pregnant women

Complete the table:

	Null hypothesis does not hold	Null hypothesis holds
Reject null hypothesis	65/70	10/30
Do not reject null hypothesis	5/70	20/30

Statistical inference: hypothesis testing

Typically, two types of error are of main interest:

- Type I error

$$\alpha = p(\text{Reject } H_0 \text{ when } H_0 \text{ is true})$$

- Type II error

$$\beta = p(\text{Do not reject } H_0 \text{ when } H_0 \text{ is false})$$

	Null hypothesis does not hold	Null hypothesis holds
Reject null hypothesis	65/70	10/30
Do not reject null hypothesis	5/70	20/30

Statistical inference: hypothesis testing

Typically, two types of error are of main interest:

- Type I error

$$\alpha = p(\text{Reject } H_0 \text{ when } H_0 \text{ is true})$$

- Type II error

$$\beta = p(\text{Do not reject } H_0 \text{ when } H_0 \text{ is false})$$

	Null hypothesis does not hold	Null hypothesis holds
Reject null hypothesis	65/70	10/30
Do not reject null hypothesis	5/70	20/30

Most hypothesis tests are designed to control type I error!

Statistical inference: hypothesis testing

In terms of this table

	Null hypothesis does not hold	Null hypothesis holds
Reject null hypothesis	65/70	10/30
Do not reject null hypothesis	5/70	20/30

- What is the **significance level** of the test?
- What is the **power** of the test?

Statistical inference: hypothesis testing

In terms of this table

	Null hypothesis does not hold	Null hypothesis holds
Reject null hypothesis	65/70	10/30
Do not reject null hypothesis	5/70	20/30

- What is the **significance level** of the test?
  - Type I error (α), i.e. $10/30 = 0.33$
- What is the **power** of the test?

Statistical inference: hypothesis testing

In terms of this table

	Null hypothesis does not hold	Null hypothesis holds
Reject null hypothesis	65/70	10/30
Do not reject null hypothesis	5/70	20/30

- What is the **significance level** of the test?
  - Type I error (α), i.e. $10/30 = 0.33$
- What is the **power** of the test?
  - 1 - type II error ($1 - \beta$), i.e. $65/70 = 0.93$

Statistical inference: hypothesis testing

Recall: basic setup

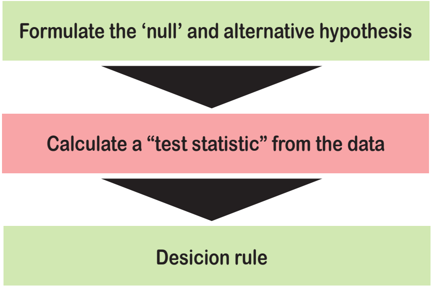
Formulate the ‘null’ and alternative hypothesis

Calculate a “test statistic” from the data

Desicion rule

Statistical inference: hypothesis testing

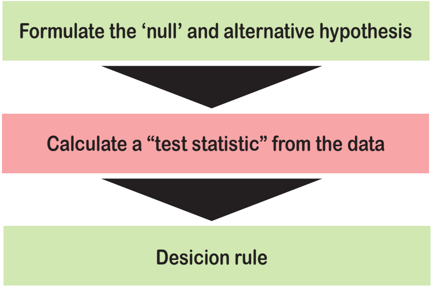
Recall: basic setup



A **test statistic** is a summary calculated from the data whose distribution is known (if H_0 is true)

Statistical inference: hypothesis testing

Recall: basic setup

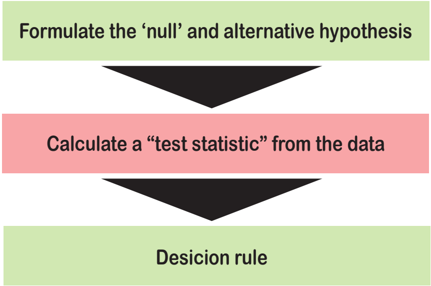


A **test statistic** is a summary calculated from the data whose distribution is known (if H_0 is true)

We will explore some example test statistics later today

Statistical inference: hypothesis testing

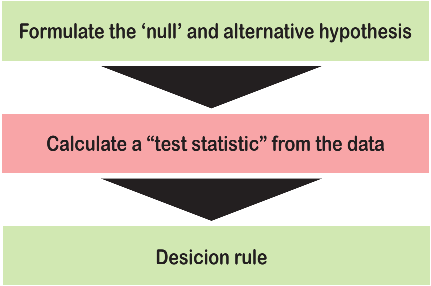
Recall: basic setup



We can also use a **p-value**

Statistical inference: hypothesis testing

Recall: basic setup



We can also use a **p-value**

What is a p-value?

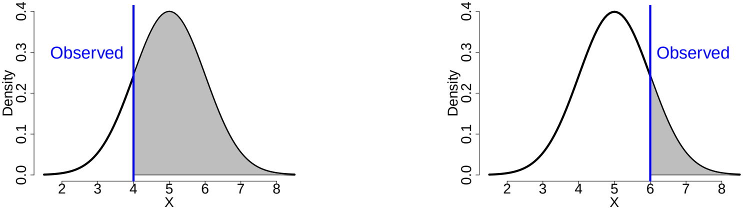
Statistical inference: what is a p-value?

A **p-value** is the probability of observing the current data (or more extreme) given than H_0 is true.

Statistical inference: what is a p-value?

A **p-value** is the probability of observing the current data (or more extreme) given than H_0 is true. For example

$$H_0 : \mu = 5 \text{ vs } H_1 : \mu > 5$$

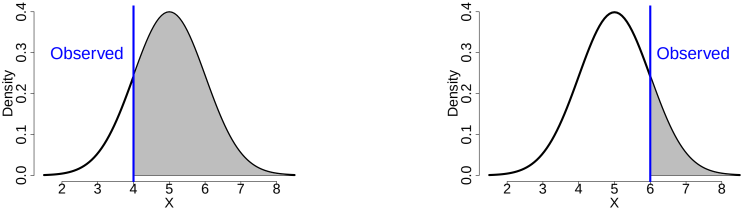


The p-value is equal to the gray area

Statistical inference: what is a p-value?

The smaller p-value, the stronger the evidence against H_0

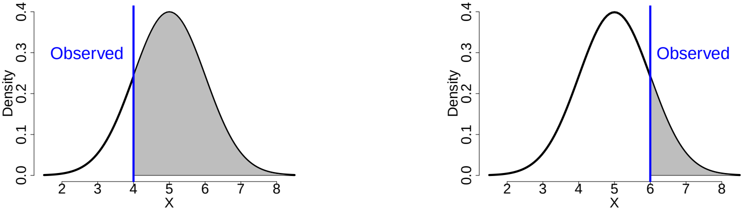
$$H_0 : \mu = 5 \text{ vs } H_1 : \mu > 5$$



Statistical inference: what is a p-value?

The smaller p-value, the stronger the evidence against H_0

$$H_0 : \mu = 5 \text{ vs } H_1 : \mu > 5$$



- How do p-values behave when the null hypothesis H_0 is true?
- How do p-values behave when the null hypothesis H_0 is false?
- How do p-values behave for different sample sizes?

http://bioinformatics.cruk.cam.ac.uk/apps/winter-school/pvalue/

Statistical inference: decision rule of an hypothesis test

Recall: basic setup

Formulate the ‘null’ and alternative hypothesis

↓

Calculate a “test statistic” from the data

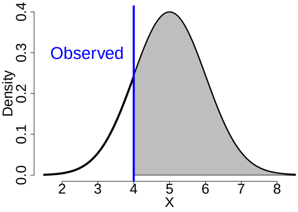
↓

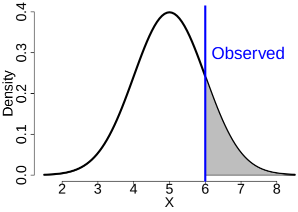
Desicion rule

Statistical inference: decision rule of an hypothesis test

The smaller p-value, the stronger the evidence against H_0

$$H_0 : \mu = 5 \ \text{ vs } \ H_1 : \mu > 5$$



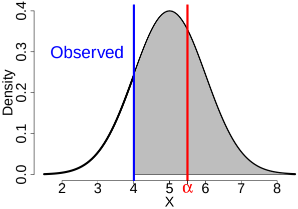


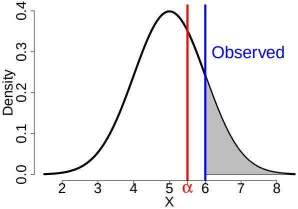
Where to place the cut-off?

Statistical inference: decision rule of an hypothesis test

The cut-off is typically set to control **type I error**

$\alpha = \mathrm{P}$ (Reject H_0 when H_0 is true) (typically $\alpha = 0.05$)

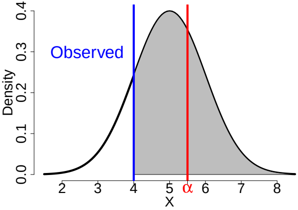


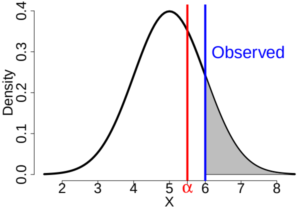


Statistical inference: decision rule of an hypothesis test

The cut-off is typically set to control **type I error**

$\alpha = \mathrm{P}$ (Reject H_0 when H_0 is true) (typically $\alpha = 0.05$)





Cut-off must be set before running the test!

Statistical inference: what test to use?

Standard statistical softwares (e.g. R) have a wide range of hypothesis tests built-in $\Rightarrow$ easy to get numbers back!

Statistical inference: what test to use?

Standard statistical softwares (e.g. R) have a wide range of hypothesis tests built-in $\Rightarrow$ easy to get numbers back!

How to choose?

Statistical inference: what test to use?

Standard statistical softwares (e.g. R) have a wide range of hypothesis tests built-in $\Rightarrow$ easy to get numbers back!

How to choose?

Before running a test it is important to recognise:

- What type of data is available?
- What is the relevant hypothesis?
- What assumptions underlie the test?

Statistical inference: what test to use?

Standard statistical softwares (e.g. R) have a wide range of hypothesis tests built-in $\Rightarrow$ easy to get numbers back!

How to choose?

Before running a test it is important to recognise:

- What type of data is available?
- What is the relevant hypothesis?
- What assumptions underlie the test?

We will explore some examples in detail this afternoon ...

Acknowledgements

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

- Mark Dunning (CRUK-CI)
- Paul Kirk (MRC-BSU)