EURORDIS
Rare Diseases Europe

Fundació Doctor Robert
UAB

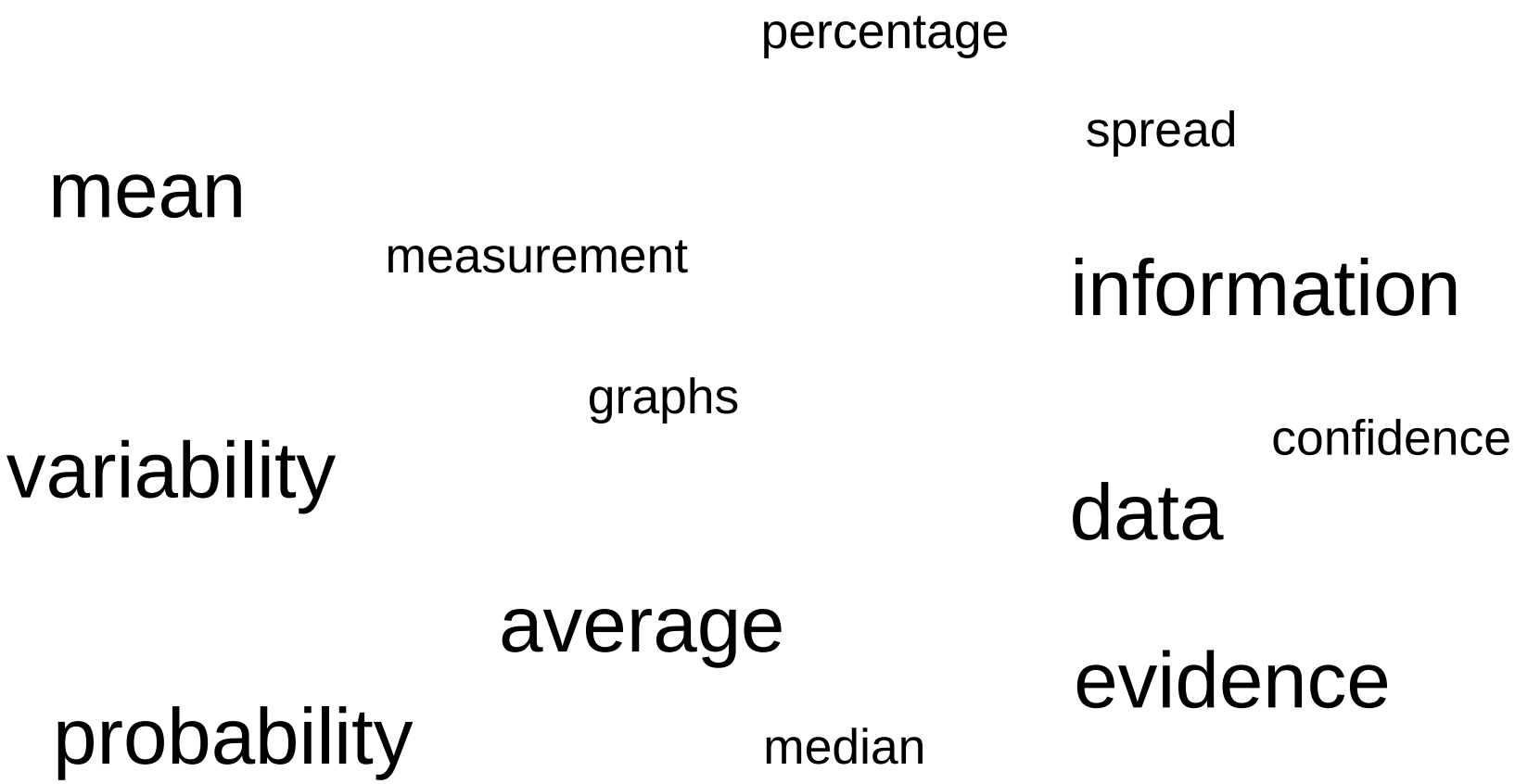
Introductory statistics for medical research

Executive Agency for Health and Consumers

Julia Saperia
European Medicines Agency

AFM
Association Française contre les Myopathies

What does *statistics* mean to you?



EURORDIS
Rare Diseases Europe

An example

- **The idea of (clinical) research is to provide answers to questions**
- **Let's ask a question…**
  - How tall are Eurordis summer school 2012 participants?

An example

- **Thinking about one (silly) question tells us some important things about data.**
  - people are different ➔variability in data
  - the reliability of an average depends on the population being studied and the sample drawn from that population ➔ valid inference

A sillier example

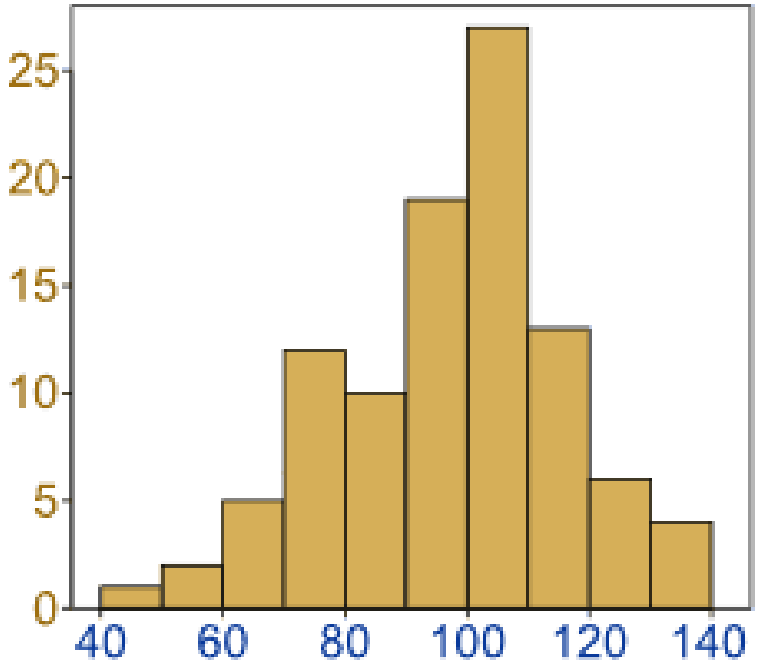
- Let's say I find out 5 heights (in cm)
  - 165, 166, 167, 168, 169
    - mean = 167.0 cm
    - standard deviation = 1.6 cm
- Let's say I find out 5 different heights (in cm)
  - 154, 165, 168, 172, 183
    - mean = 168.4 cm
    - standard deviation = 10.5 cm
- So the means are the quite similar…but the two groups look different

A sillier example

- **what we learn from the second example is that we get more information from data if we know *how* variable they are**

➔ with these pieces of information, we can draw a picture of the data (provided some assumptions are true)

The histogram

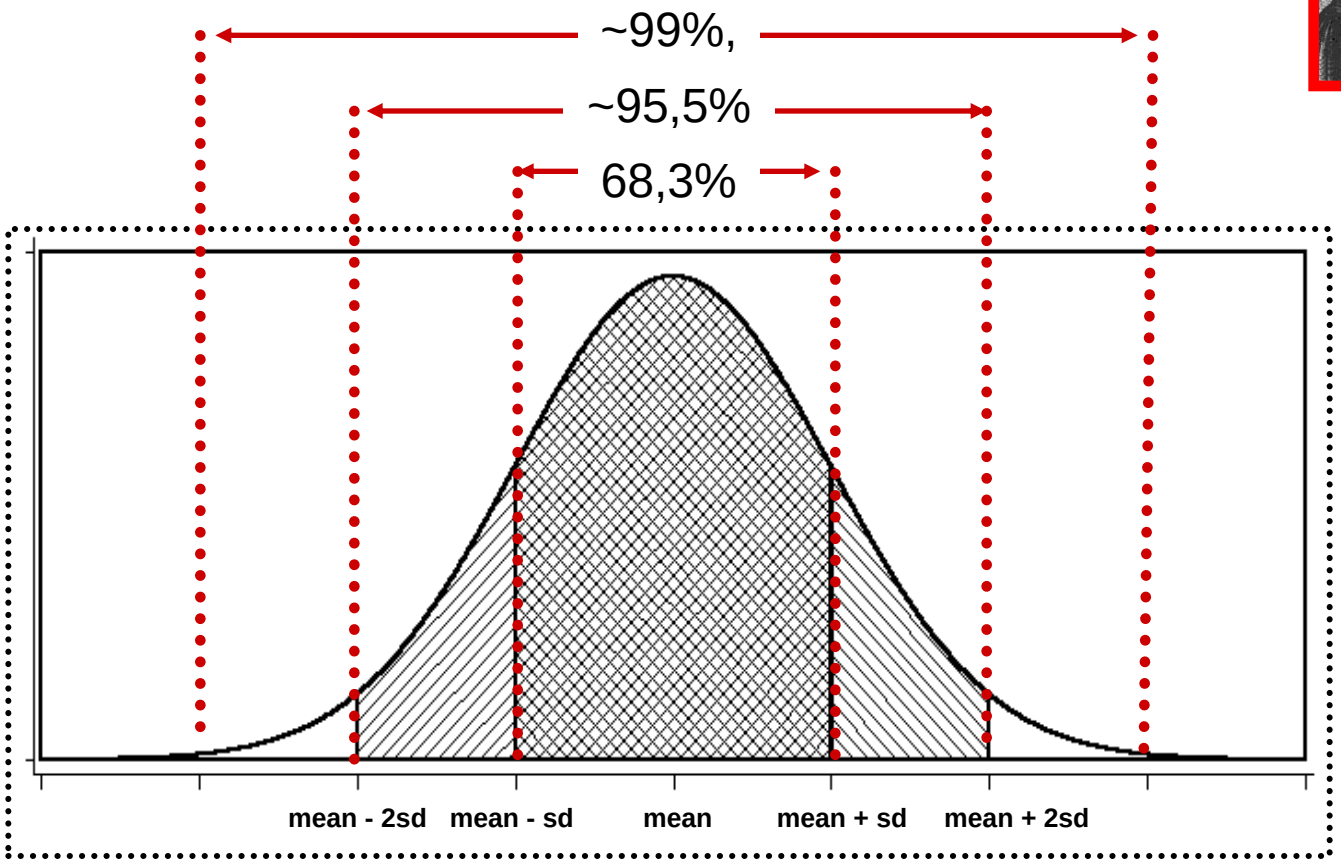


•Taken from http://www.mathsisfun.com/data/histograms.html on 28 May 2012

EURORDIS
Rare Diseases Europe

Picturing “normal” data

Johann Carl Friedrich Gauss 1777-1855



EURORDIS
Rare Diseases Europe

Biostatistics

… is the method of gaining empirical knowledge in the life sciences

… is the method for quantifying variation and uncertainty

1. How do I gain adequate data?

→ Thorough planning of studies

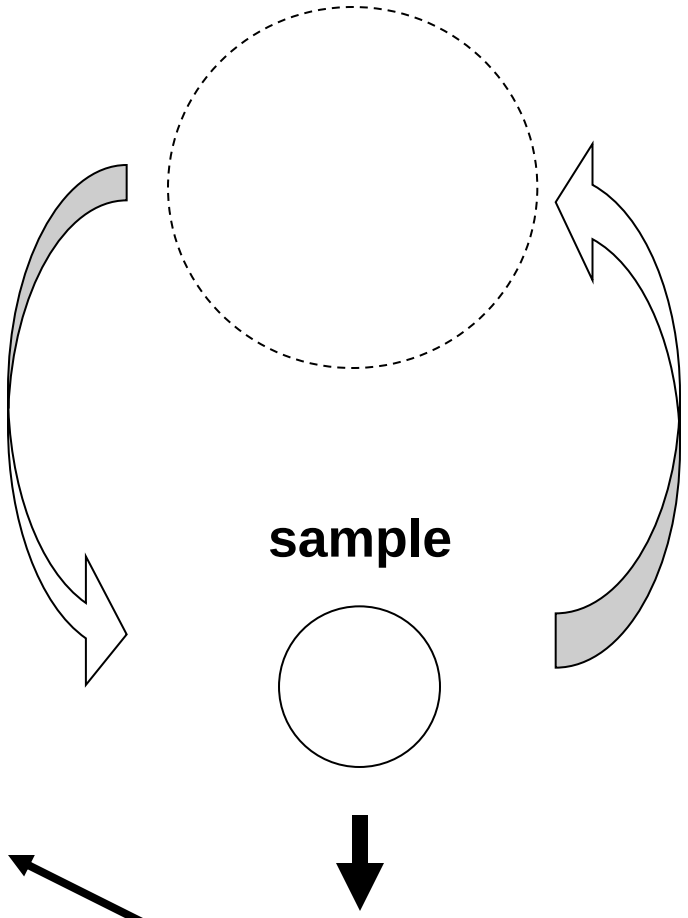
2. What can I do with my data?

→ Descriptive and inferential statistics

If (1) is not satisfactorily resolved, (2) will never be adequate.

Population

(a population is a set of all conceivable observations of a certain phenomenon)



Inferential statistics

Draw conclusions for the whole population based on information gained from a sample

Target of study design
representative sample
= ~ all "typical" representatives are included

EURORDIS
Rare Diseases Europe

Descriptive statistics (1)

- **For continuous variables (e.g. height, blood pressure, body mass index, blood glucose)**
- **Histograms**
- **Measures of location**
  - mean, median…
- **Measures of spread**
  - variance, standard deviation, range, interquartile range

Descriptive statistics (2)

- **For categorical variables (e.g. sex, dead/alive, smoker/non-smoker, blood group)**
- **Bar charts, (pie charts)…**
- **Frequencies, percentages, proportions**
  - when using a percentage, always state the number it's based on
    - **(e.g. 68% (n=279))**

Describing data

- **Going back to the heights example**
  - we didn't have time to measure the heights of the **population** (all participants), so we took a **sample**
  - we calculated the **mean** of the sample
  - we can use the mean of the sample as an **estimate** of the mean in the population
  - but seeing as it's only an estimate, we want to know how reliable it is; that is, how well it represents the **true** (population) value
  - we calculated the **standard deviation** of the sample
  - this is a fact (like the sample mean) about the **spread** of the data in the sample

Inferential statistics – confidence intervals

- the standard deviation also helps us to understand how precise our estimate of the mean is
- the precision depends on the **number** of data points in the sample
  - we can calculate the **standard error** of the mean

$$\text{standard error of the mean} = \frac{\text{standard deviation}}{\sqrt{\text{number in sample}}}$$

  - (in our more realistic example, the standard error is 4.7)
  - the **larger** the sample, the **smaller** the standard error, the more **precise** the **estimate**
- so now we have an estimate of the population mean and a measure of its precision
- so what?
- we can use the standard error of the mean and the properties of the normal distribution to gain some level of confidence about our estimate

Confidence intervals

- How confident do you need to be about the outcome of a race in order to place a bet?
- In statistics it's typical to talk about **95%** confidence
- we can use the **standard error of the mean** to calculate a **95% confidence interval** for the **mean**

$$95\%\ \text{confidence interval} \approx \text{mean} \pm 2 \times \text{standard error}$$

estimate of population mean

relates to the normal distribution

measure of precision of the sample mean

- in our example: 95%CI for the mean ≈ 168.4 2 x 4.7
  ≈ [159.0, 177.8]

EURORDIS
Rare Diseases Europe

Confidence intervals

- What does a 95% confidence interval *mean*?
- remember that we have drawn a sample from a population
- we want to know how good our estimate of the population mean (the truth) is
- the 95% CI tells us that the true mean lies in 95 out of 100 confidence intervals calculated using different samples from this population

  ➔ true mean lies between 159 cm and 177.8 cm with 95% probability
- **remember:** the larger the number in the sample, the smaller the standard error
- the smaller the standard error, the narrower the confidence interval

  ➔ the larger the number in the sample, the more precise the estimate

A real example

Lateral wedge insoles for medial knee osteoarthritis: 12 month randomised controlled trial (*BMJ 2011; 342:d2912*)

Participants 200 people aged 50 or more with clinical and radiographic diagnosis of mild to moderately severe medial knee osteoarthritis.

Interventions Full length 5 degree lateral wedged insoles or flat control insoles worn inside the shoes daily for 12 months.

Main outcome measure Change in overall knee pain (past week) measured on an 11 point numerical rating scale (0=no pain, 10=worst pain imaginable).

Research question:

Are full length 5 degree lateral wedged insoles better than flat control insoles at reducing pain after 12 months in patients with mild to moderately severe medial knee osteoarthritis?

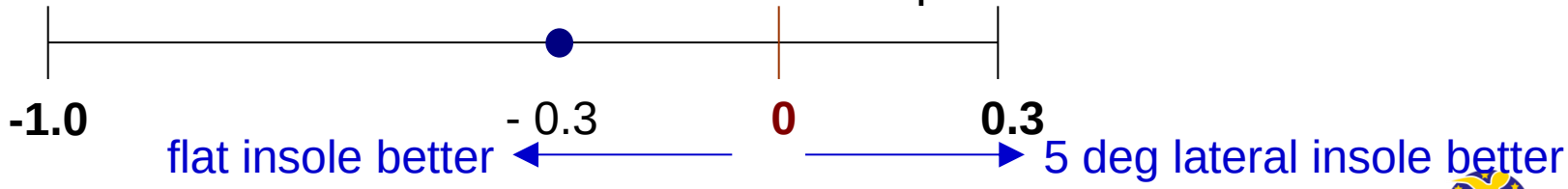
A real example

- Assume 5 degree lateral wedged insoles and flat control insoles reduce pain to the **same** extent (this is our **null hypothesis**)
- Collect data in both groups of patients and calculate:

  the mean reduction from baseline in the 5 deg lateral insole group (0.9)

  the mean reduction from baseline in the flat insole group (1.2)

  the difference between the two groups in mean change from baseline (0.9 – 1.2 = -0.3)
- We can see that there was some reduction in pain in **both** groups and that the reduction was greater in the flat insole group
- The 95% confidence interval tells us how important that difference is



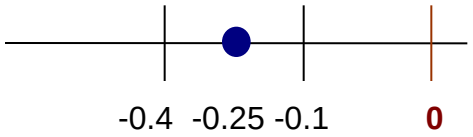
EURORDIS
Rare Diseases Europe

A real example

- **Conclusion:** with the available evidence, there is no reason to suggest that one treatment is better than the other in reducing pain scores after one year.
- **Or:** there is no **statistically significant** difference in treatments (because the 95% confidence interval crosses the null value)

Statistical significance vs clinical significance

- **Remember:** the larger the number of data points in a sample, the more precise the estimate
- Let's **imagine** that instead of enrolling 200 patients in their study, the researchers enrolled 1500.



flat insole better ← → 5 deg lateral insole better

- This time, the confidence interval does not cross the null value
- So we can conclude that the flat insole reduces pain by 0.25 points more than the 5 degree lateral insole.
- Hooray?
- Is a difference of 0.25 points **clinically relevant?**

Quantifying statistical significance – the p-value

- In the last example, we looked at the 95% confidence interval to decide whether we had a statistically significant result
- (the confidence interval is useful because it gives clinicians and patients a range of values in which the true value is likely to lie)
- another measure of statistical significance is the **p-value**
- **Remember** our **null hypothesis:** 5 degree lateral wedged insoles and flat control insoles reduce pain to the **same** extent
- We go into our experiment assuming that the null hypothesis is **true**
- After we've collected the data, the **p-value** helps us decide whether our results are due to chance alone

The p-value

- the p-value is a probability ➔ it takes values between 0 and 1
- the p-value is the probability of observing the data (or data more extreme) **if** the null hypothesis is true

  ➔ (if the null hypothesis is true, the two treatments are the same)

  ➔ a low p-value (a value very close to zero) means there is low probability of observing such data when the null hypothesis is true

  ➔ reject the null hypothesis (i.e. conclude that there is difference in treatments)

Significance level

- p-values are compared to a pre-specified **significance level,** alpha
- alpha is usually 5% (analogous to 95% confidence intervals)

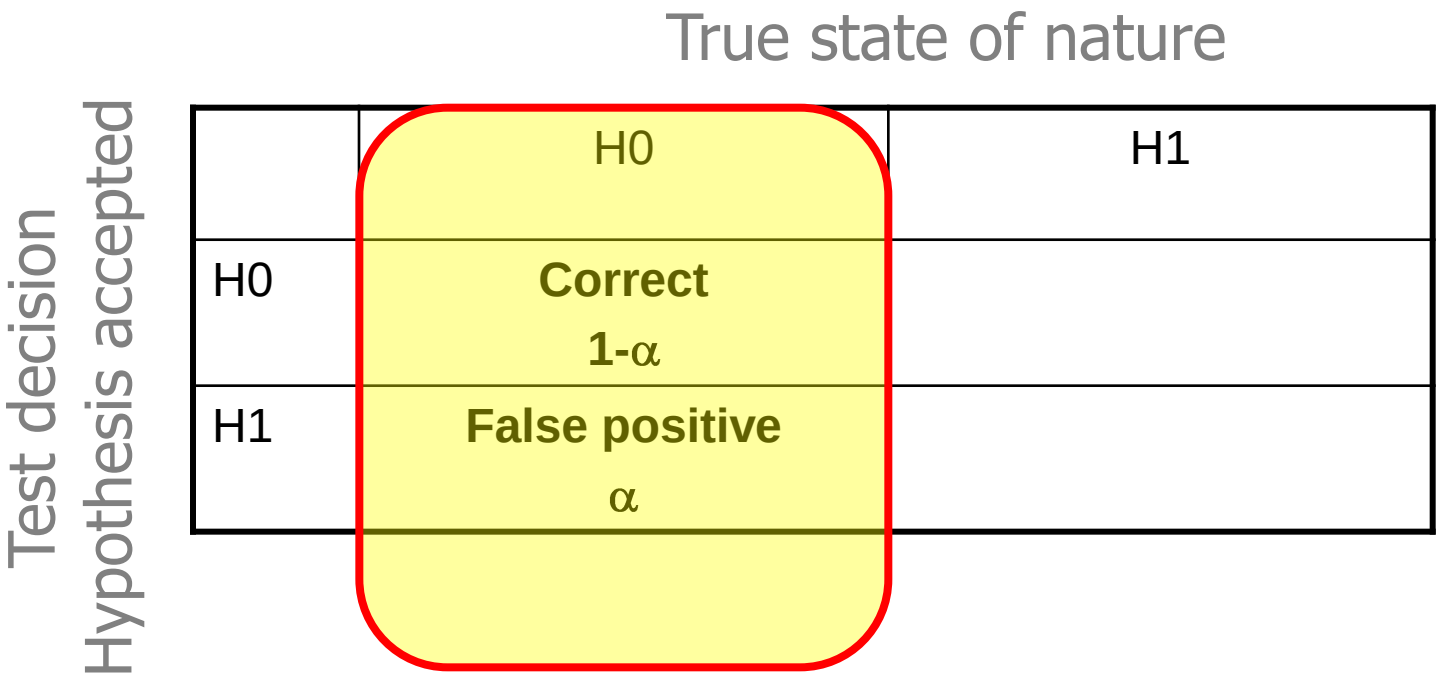
  ➔ if p ≤ 0.05 reject the null hypothesis (i.e., the result is statistically significant at the 5% level) ➔ conclude that there is a difference in treatments

  ➔ if p > 0.05 do **not** reject the null hypothesis ➔ conclude that there is not sufficient information to reject the null hypothesis

(See **Statistics notes: Absence of evidence is not evidence of absence***: BMJ 1995;311:485)*

The authors of the insole paper did not cite the *p*-value for the result we looked at but we can conclude that p>0.05 because the 95%CI crosses the null value and so we cannot reject the null hypothesis (of no difference).

Statistical mistakes/errors



Type I error:

Error of rejecting a null hypothesis when it is actually true
(α error, error of the first kind)

α represents a FALSE POSITIVE decision: A placebo is declared to be more effective than another placebo!)

EURORDIS
Rare Diseases Europe

Statistical mistakes/errors

True state of nature

Test decision
Hypothesis accepted

	H0	H1
H0	**Correct** **1-α**	**False negative** β
H1	**False positive** α	**Correct** **1-β**

Type II error:

The alternative hypothesis is true, but the null hypothesis is erroneously not rejected. (β error, error of the second kind)

FALSE NEGATIVE decision: an effective treatment could not be significantly distinguished from placebo

Statistical mistakes/errors

True state of nature

Test decision
Hypothesis accepted

	H0	H1
H0	**Correct** **1-α**	**False negative** β
H1	**False positive** α	**Correct** **1-β**

Significance level α is usually predetermined in the study protocol, e.g., **α=0.05 (5%, 1 in 20), α= 0.01 (1%, 1 in 100), α= 0.001 (0.1%, 1 in 1000)**

POWER 1- β : The power of a statistical test is the probability that the test will reject a false null hypothesis.

Statistical mistakes/errors

True state of nature

Test decision
Hypothesis accepted

	H0	H1
H0	**Correct** **1-α**	**False negative** β
H1	**False positive** α	**Correct** **1-β**

⇨**Consumer's risk?**

⇨**Producer's risk?**

Statistical tests vs court of law

True state of nature

Test decision

Decision of the court

	H0 Innocent	H1 Not Innocent
H0 Judged "innocent"	**Correct** **1-α**	**False negative** **(i.e. guilty but not caught)** β
H1 Judged "Not innocent"	**False positive** **(i.e. innocent but convicted)** α	**Correct** **1-β**

False positive: A court finds a person guilty of a crime that they ***did not*** actually commit.
False negative: A court finds a person not guilty of a crime that they ***did*** commit.

Minimising statistical errors

Remember:

How do I gain adequate data?

→ Thorough planning of studies

- Defining acceptable levels of statistical error is key to the planning of studies
- **alpha** (in clinical trials) is pre-defined by regulatory guidance (usually)
- **beta** is not, but deciding on the power (1-beta) of the study is crucial to enrolling sufficient patients
- the power of a study is usually chosen to be 80% or 90%
- conducting an “underpowered” study is not ethically acceptable because you know in advance that your results will be inconclusive

Deciding how many patients to enrol

- the sample size calculation depends on:
  - the **clinically relevant effect** that is expected (1)
  - the amount of **variability** in the data that is expected (2)
  - the **significance level** at which you plan to test (3)
  - the **power** that you hope to achieve in your study (4)
- if you knew (1) and (2), you wouldn't need to conduct a study

  ➔ **picking your sample size is a gamble**
- the smaller the treatment effect, the more patients you need
- the more variable the treatment effect, the more patients you need
- the smaller the risks (or statistical errors) you're prepared to take, the more patients you need

Conclusion

- amount of variability in data defines conclusions from statistical tests
- clinical trial methodology is underpinned by statistics
- valid statistical inference, and therefore decision-making, depends on robust planning
- understanding statistics isn't as hard as it looks (I hope)

Further reading

- www.consort-statement.org
  - standards for reporting clinical trials in the literature
- **Statistical Principles for Clinical Trials ICH E9**
  - useful glossary
- http://openwetware.org/wiki/BMJ_Statistics_Notes_series
  - coverage of a number of topics related to statistics in clinical research, mostly by Douglas Altman and Martin Bland