The role of chance:
P-values and confidence intervals

Gustaf Edgren, PhD
Karolinska Institutet

Disclaimer ;-)

- The following slides are meant to give you a pragmatist's introduction to probability and the measures of probability most commonly used in epidemiology/medicine
- I have no ambition to give you a thorough introduction to probability, nor statistics, but to try to give you a working knowledge of why and how we use statistics/probability

Outline

- Background and rationale
- Sampling variability and random error
- P-values
- Confidence intervals
- Power

Summary of Lind's work

- James Lind conducted a 6-armed controlled trial
  - None of ten patients receiving elixir vitriol, sea water, cider, vinegar, or spice mix showed any improvement
  - Both patients who received lemons and oranges were cured

Scurvy and vitamin C – conclusion?

- Still, Lind concluded that the study was too small and needed to be repeated by other scientists…
- Was he right?

Lind's results

- Were the differences in treatment success Lind observed true?
  - Bias: A systematic error in the conduct of a study which offsets the results in some non-random matter
  - Confounding: Mixing of the effect of the exposure and the effect of some other, discrete variable
  - Chance: Random distortion of the results due to sampling variability

Random errors

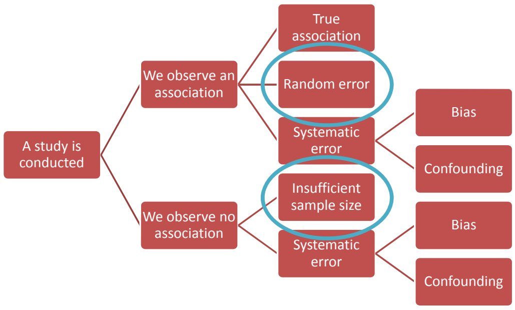
- Two types of random errors:
  - Type 1 errors – false positives

    ”We conclude that **there is a difference** when there really is none”

  - Type 2 errors – false negatives

    ”We conclude that **there is no difference** when there really is a difference”

Possible outcomes in a study



A study is conducted

We observe an association

True association

Random error

Systematic error

Bias

Confounding

We observe no association

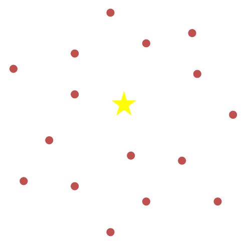
Insufficient sample size

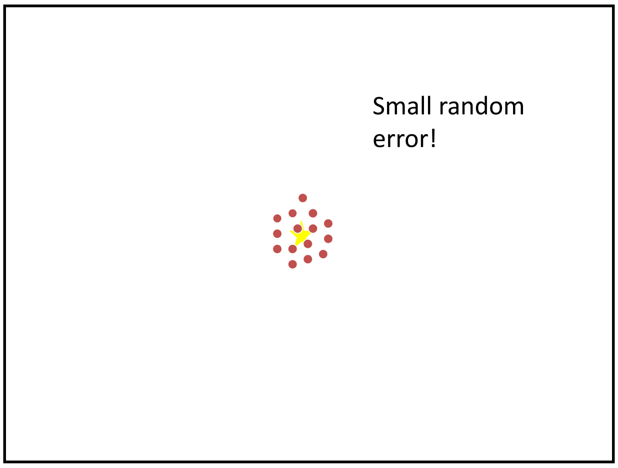
Systematic error

Bias

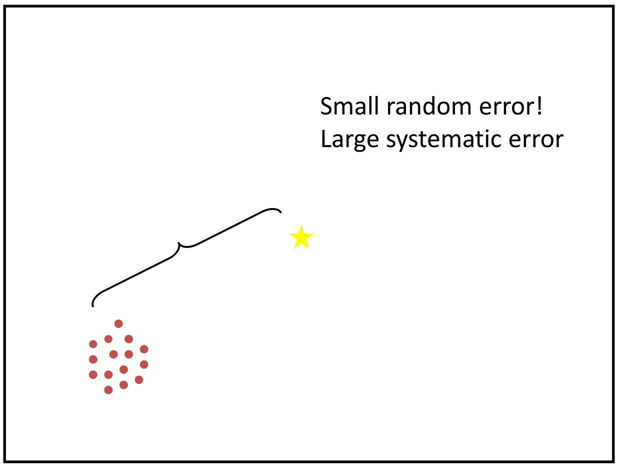
Confounding

Large random error!

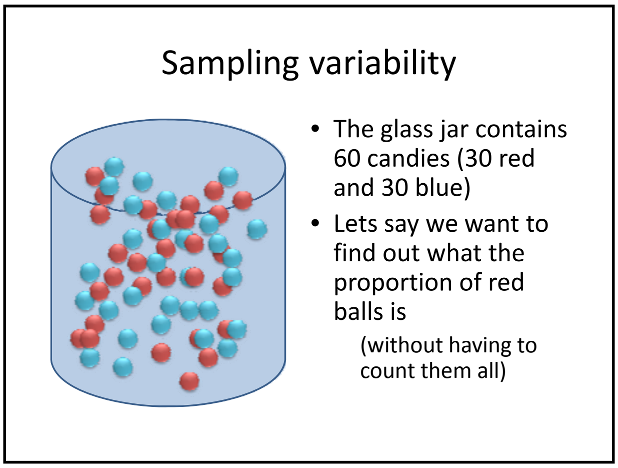




Small random error!



Small random error!
Large systematic error

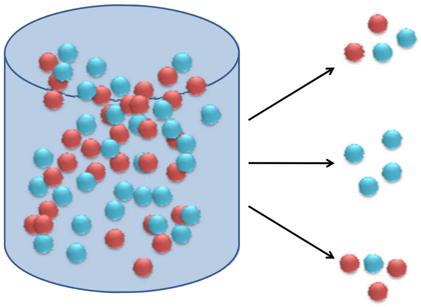


Sampling variability

- The glass jar contains 60 candies (30 red and 30 blue)
- Lets say we want to find out what the proportion of red balls is

(without having to count them all)

Sampling variability



Sample 1
50%=right!

Sample 2
0%=wrong

Sample 3
75%=wrong

Sampling variability, cont.

- Can we use statistics to guide our evaluation of the random samples?
  → Statistical inference
- Simple insights:
  - The larger the sample, the smaller the sampling variability
  - Larger samples decrease the probability of drawing unrepresentative samples

Statistical inference

- Inference from statistical samples are often based on the hypothetico-deducive model:
  1. An hypothesis is specified
  2. A sample is drawn (not necessarily of red or blue balls…)
  3. An appropriate statistical test is performed to quantify the statistical uncertainty associated with the sample to test the hypothesis

1. Specify hypothesis

- Hypotheses are often specified according to a common nomenclature:
  - – Null hypothesis (H_0):
    - No association
    - RR=1 / RD=0
  - – Alternative hypothesis (H_A/H_1):
    - There is an association
    - RR≠1 / RD≠0

2. Conduct study (i.e. draw sample)

- Conduct the study to quantify the measure of association of interest
- Even when studies are performed using the total population of a nation, the data should still be considered to be a sample

3. Perform hypothesis test

- Within the framework of the hypothetic-deductive model, hypothesis tests have a binary outcome:
  - – We can either reject our null hypothesis (H_0) – which is to say that “there is a difference”
  - – Or, we can accept our null hypothesis (H_0) – which is NOT to say that “there is no difference”
- *The choice of hypothesis test depends on the type of data and will be covered next week*

3. Perform hypothesis test, cont.

- Generally, however, hypothesis tests will help us quantify whether the observed difference could have arisen from chance alone
- Or, more correctly, could we have observed a difference of this magnitude had the null hypothesis been true?

Hypothesis tests

- A myriad of different statistical models for hypothesis testing exist:
- Basically, these tests return the probability that the results were observed even if the null hypothesis was true
  - This probability is expressed as a P-value

P-values

- The infamous P-value is the probability that the observed results could have arisen due to chance, given that the null hypothesis is true

- On the basis of the P-value we will either reject or accept the null hypothesis

Significance level

- Commonly, the significance threshold of 5% (0.05) is entertained
- At or below this arbitrary level,
  - We reject the null hypothesis
  - Consider our findings “statistically significant”
  - Conclude that “chance is an unlikely explanation for the observed difference”

Hypothesis testing, cont.

- At P-values above 0.05:
  - We accept the null hypothesis,
  - consider our findings “not statistically significant,”
  - and conclude that chance cannot be excluded as an explanation for the observed difference
- *Note: At P-values close to (but above 0.05), it is not statistically sound to refer to the findings as “borderline significant”*

Significance level, cont.

- At the common significance level of 0.05 we ensure that the type 1 error rate is only 5%:
  - This means, that for every 100 hypothesis tests performed, 5 will be “false positives”
  - Is this acceptable?
  - What about at a significance level of 0.01 or 0.001?
  - What about in a criminal case?

Example of P-value – Sally Clark

- A British mother of two who lost first on child to SIDS in 1996 and another in 1998, and who subsequently was prosecuted for manslaughter.
- At her trial, a certain Professor Meadow concluded that the probability of loosing two children two SIDS was 1 in 73 million (P=1/73,000,000)
- This calculation was based on the accepted probability for a woman of Sally Clark's age to loose a child to SIDS was 1/8543
- Comments?

Note on P-values

- It is important to stress that although P-values are powerful tools for the assessment of the role of chance:
  - Significant P-values (irrespective of at what level) DO NOT EXCLUDE THE POSSIBILITY OF BIAS and say NOTHING WHATSOEVER ABOUT CAUSALITY
  - Thus, before interpreting a P-value, we need to consider bias and confounding as explanations for our findings

Confidence intervals

- Confidence intervals represent an alternative, complementary way of performing hypothesis tests
- A confidence interval describes the sampling variability of the point estimate
- As with P-values, confidence intervals are constructed and evaluated with a type 1 error rate (significance level) in mind (typically 5%)

Confidence intervals, cont.

- Theoretically, if 100 samples were drawn from the same population and 100 point estimates and confidence intervals were calculated, the point estimate would end up within the intervals 95 times
- Alternatively, the confidence interval gives the range within which the true point estimate occurs with a certain probability

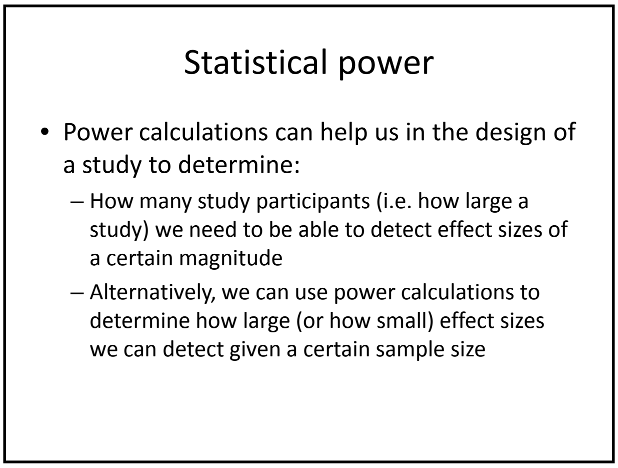
Confidence intervals, cont.

- In practical terms a confidence interval can be used for hypothesis testing by examining whether the interval excludes the null hypothesis (e.g. RR=1, RD=0)
- Furthermore, the width of the interval also informs us about the precision of the point estimate
  - → Confidence intervals are therefore more versatile and useful than P-values

Type 2 errors

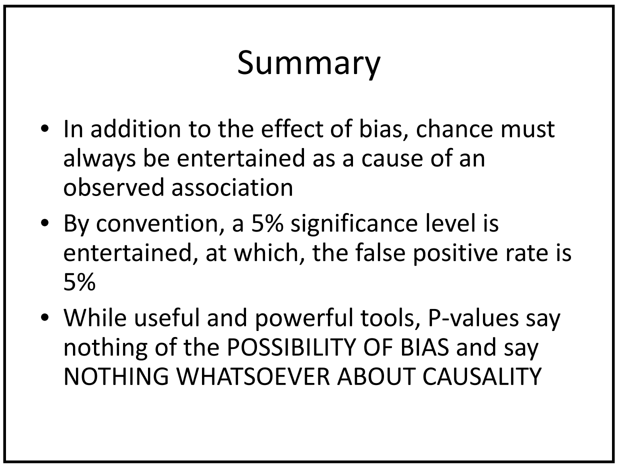
- On the other end of the statistical spectrum are type 2 errors:
  - i.e. ”that we conclude that **there is no difference** when there really is a difference”
- Intuitively, type 2 errors occur when the sample size is too small:
  - i.e. when we have insufficient power*

*Power = 1-Type 1 error rate = 1- β



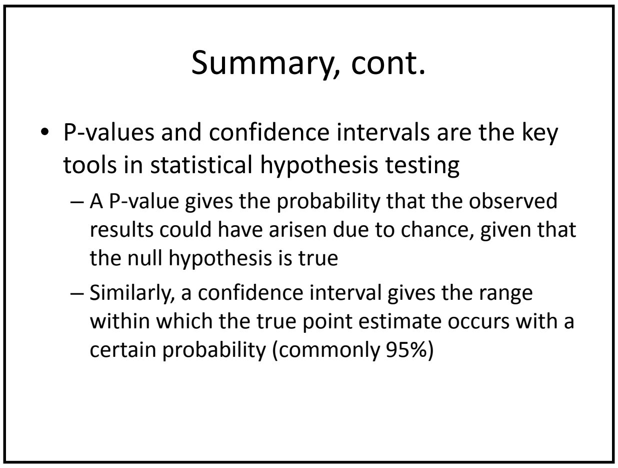
Statistical power

- Power calculations can help us in the design of a study to determine:
  - How many study participants (i.e. how large a study) we need to be able to detect effect sizes of a certain magnitude
  - Alternatively, we can use power calculations to determine how large (or how small) effect sizes we can detect given a certain sample size



Summary

- In addition to the effect of bias, chance must always be entertained as a cause of an observed association
- By convention, a 5% significance level is entertained, at which, the false positive rate is 5%
- While useful and powerful tools, P-values say nothing of the POSSIBILITY OF BIAS and say NOTHING WHATSOEVER ABOUT CAUSALITY



Summary, cont.

- P-values and confidence intervals are the key tools in statistical hypothesis testing
  - A P-value gives the probability that the observed results could have arisen due to chance, given that the null hypothesis is true
  - Similarly, a confidence interval gives the range within which the true point estimate occurs with a certain probability (commonly 95%)