Power and Sample Size in Scientific Investigastions

with Special Focus on Veterinary Science

Platform Biostatistics

VetMedUni Vieanna

05/2014

Plattform Bioinformatik und Biostatistik

- Plattform Biostatistik
  - Univ.-Prof. Dr. Andreas Futschik
  - Dr. Alexander Tichy
  - Dr. Marlies Dolezal
  - Ass.-Prof. Dr. Margarete Hofmann-Parisot
- Plattform Bioinformatik
  - Univ.-Prof. Dr. Christian Schlötterer

Motivation

- Freiman et al. (1978) investigated 71 published randomized controlled clinical trials that did not find significant differences between groups:
  - 67 of trials had risk $>$ 10% to miss 25% therapeutic improvement
  - 50 of trials had risk $>$ 10% to miss 50% therapeutic improvement
- Many interesting/beneficial treatments might have been missed due to low power!

Freiman J.A., Chalmers T.C., Smith H., Kuebler R.R. (1978) The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 *negative* trials. N. Engl. J. Med. 299: 690-694.

Guidelines

> *The number of animals in a clinical trial should always be large enough to provide reliable answers to the questions addressed.*

European **ME**dicines **A**gency: Guideline on statistical principles for clinical trials for veterinary medicinal products (2012).

http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2012/01/WC500120834.pdf

> *. . . the number should be sufficient to achieve worthwhile results, but should not be so high as to involve unnecessary recruitment and burdens for participants.*

Guidance on sample size by the Central Office for Research Ethics Committees (COREC) (2007)

Under-powered and over-powered studies

- Under-powered studies: work with too small sample sizes, risk of not finding interesting effects (more common)
- Over-powered studies: work with too large sample sizes, smaller sample would have been sufficient

Both cases lead to waste of time and resources and may lead to unnecessary harm of animals!

Before determining sample size

we need

- research hypothesis
- determine type of data to be collected for answering the research hypothesis
- statistical method to test research hypothesis, i.e. which statistical test will be used?

Example

Renal function in older cats

- goal is to measure effect of a new treatment to improve renal function of old cats
- plan to compare two groups (standard treatment/new treatment)
- how to measure renal function?
- plasma urea and/or creatinine concentration?
- which of them should be primary outcome?
- distribution of outcome? (previous studies, or pilot study)
- if approx. normal plan to use two-sample t-test

One or several outcomes?

- investigators often look at several outcomes
- if there is a primary outcome: power/sample size computation usually done for primary outcome
- if no primary outcome can be chosen:
  - sample size required should ensure sufficient power for all outcomes!
  - also multiple testing problem needs to be addressed

Basics of hypothesis testing

- hypothesis tests used for decision between research hypothesis (H_1) and a null hypothesis (H_0, no effect hypothesis)
- hypothesis tests use a test statistic to measure the evidence against null hypothesis
- p-values ($\in [0, 1]$) are commonly used to summarize evidence against H_0
- small p-values (usually: $p < 0.05$) provide evidence against null hypothesis

True and wrong decision

	correct hypothesis	
decision for	H_0	H_1
H_0	true	type II error
H_1	type I error	true

Errors and power

- type I error: detect an effect when there is none
- hypothesis tests constructed such that

$$P(\text{type I error}) \leq \alpha = 0.05 \quad \text{(most common)}$$

- type II error: not detecting an effect that is present
- power of a test: probability of correctly detecting an effect

$$\text{power} = 1 - P(\text{type II error})$$

Power

For a given test, power depends on ...

- sample size
- permitted probability of type I error α
- actual size of effect
- power can depend also on nuisance parameters (most common: standard deviation of observations/measurements)
- one-sided vs. two-sided test

One-sided and two sided test

- Two-sided tests (standard with statistical software): look at deviations from null in both directions
- One-sided tests: look only at deviations from null in one direction (e.g. improvement)
- One-sided tests appropriate, if known in advance that only deviations from null in one direction either possible or of interest
- If observed effect goes in assumed direction then:

$$p_1 = \frac{1}{2} p_2$$

Example: power of z-test I

- consider sample $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ with known variance σ^2.
- consider one-sided test: $H_0 : \mu = 25$, $H_1 : \mu > 25$
- z-Test at level α rejects H_0, if

$$T = \sqrt{n}\frac{\bar{X} - 25}{\sigma} > Q^{(N)}(1-\alpha)$$

- under H_0 distribution of $T \sim N(0,1)$

z-test used to test for mean of normal sample with known variance, or whether population proportion equals pre-specified value

Example: power of z-test II

distribution of T for general μ:

1. Here: $\bar{X} \sim N(\mu, \sigma^2/n)$
2. Thus $T \sim N(\gamma, 1)$, where $\gamma = \gamma(\mu) = \sqrt{n}\frac{\mu - 25}{\sigma}$, with 25 value μ under H_0.
3. so probability of rejecting H_0:

$$\begin{aligned} P[T > Q^{(N)}(1-\alpha)] &= P[T - \gamma > Q^{(N)}(1-\alpha) - \gamma] \\ &= 1 - \Phi[Q^{(N)}(1-\alpha) - \gamma] \end{aligned}$$

Example: power of z-test III

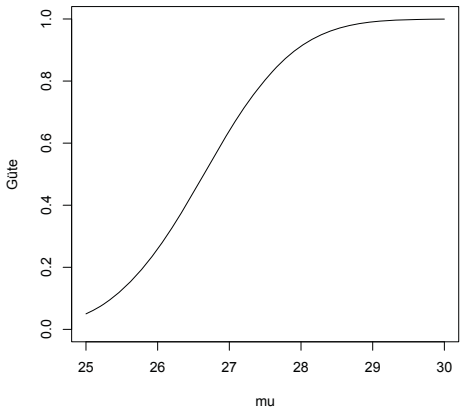
numerical example

- $\alpha = 0.05$, $n = 25$, $\sigma = 5$
- then $g(\mu)$

$$g(\mu) = 1 - \Phi[Q^{(N)}(1-\alpha) - \gamma] = 1 - \Phi[1.645 - \frac{5}{5}(\mu - 25)]$$

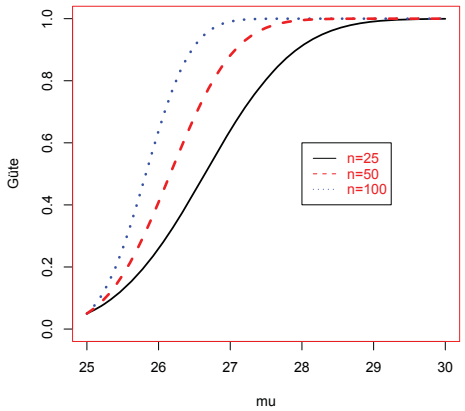
- e.g. $g(26) = 1 - \Phi(0.645) \approx 0.259$

Example: power of z-test IV



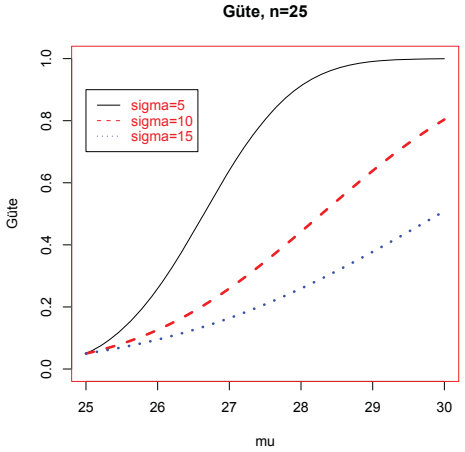
power of one-sided z-test depending on μ. Further parameters: $\alpha = 0.05$, $n = 25$, $\sigma = 5$

Power: different sample sizes



power of one-sided z-test depending on μ for different sample sizes n. ($\alpha = 0.05$, $\sigma = 5$)

Power: different magnitudes of individual variation



power of one-sided z-test depending on μ for different error SDs σ. ($\alpha = 0.05$, $n = 25$)

Power of two-sided z-Test I

- From one-sided test we know:

$$T \sim N(\gamma, 1),$$

with $\gamma = \gamma(\mu) = \sqrt{n}\frac{\mu - 25}{\sigma}$.

- reject now H_0, if $|T| > Q^{(N)}(1 - \alpha/2)$. Thus rejection probability:

$$\begin{aligned} & 1 - P[-Q^{(N)}(1-\alpha/2) \leq T \leq Q^{(N)}(1-\alpha/2)] = \\ & = \quad 1 - P[-Q^{(N)}(1-\alpha/2) - \gamma \leq T - \gamma \leq Q^{(N)}(1-\alpha/2) - \gamma] \\ & = \quad 1 - \left(\Phi[Q^{(N)}(1-\alpha/2) - \gamma] - \Phi[-Q^{(N)}(1-\alpha/2) - \gamma]\right), \end{aligned}$$

with γ depending on μ .

Power of two-sided z-Test II

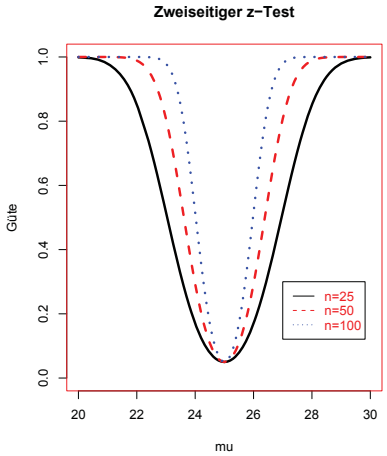
Numerical Example

- $\alpha = 0.05$, $n = 25$, $\sigma = 5$
- Then $Q^{(N)}(1 - \alpha/2) = 1.96$ and $\gamma(\mu) = \mu - 25$, with 25 again the value of μ under H_0.
- Thus power

$$g(\mu) = 1 - (\Phi[(25 - \mu) + 1.96] - \Phi[(25 - \mu) - 1.96])$$

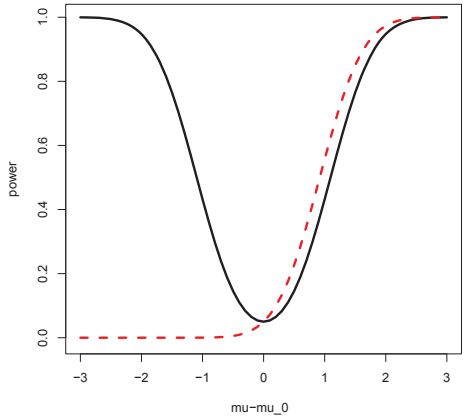
- e.g. $g(24) = 1 - (\Phi(2.96) - \Phi(-0.96)) = 1 - \Phi(2.96) + 1 - \Phi(0.96)) \approx 0.1701$

Power of two-sided z-Test III



power of one-sided z-test depending on μ and n; further parameters: $\alpha = 0.05$, $\sigma = 5$

Power one-sided vs. two-sided z-Test



power of one-sided z-test depending on μ and n; further parameters: $\alpha = 0.05$, $\sigma = 5$

Power versus sample size

- How to come up with an appropriate sample size when designing a study?
- Standard approach: choose n such that desired power is obtained for a given effect size
- For this purpose need to select:
  - α usually $= 0.05$
  - power usually ≥ 0.8
  - effect size . . . more difficult to specify

Choice of effect size

- Plausible effect size . . .
  - from previous related studies
  - from pilot studies
  - minimal clinically relevant effect
  - Cohen (1988) — not recommended

Power versus sample size

Nnumerical example for one-sided z-test

- Necessary sample size for detecting effect size of 0.5 SDs with probability 0.8? ($\alpha = 0.05$)
- Solution: $1 - \Phi[Q^{(N)}(1-\alpha) - \gamma] = 0.8$
- thus: $0.2 = \Phi[Q^{(N)}(1-\alpha) - \gamma]$
- and: $Q^{(N)}(0.95) - Q^{(N)}(0.2) = \gamma$
- $2.926 = \gamma = \sqrt{n}\frac{\sigma/2}{\sigma}$
- therefore : $n \approx 25$

How about other tests?

Do I need to carry out computations by hand?

- Similar computations possible for many other tests:
- t-test, chi-square test, nonparametric tests, ANOVA, regression,. . . ,
- distributions under alternative get often more complicated,
- but there are software packages available.

When can power be smaller than originally planned?

- wrong initial assumptions about effect size and/or random variation in measurements
- non-response
- drop out (e.g. from long term studies)
- properties of data do not match original model assumptions (e.g. normal distribution)

What if budget permits only for a certain sample size?

- Assume that you know that your resources (budget or availability of animals) only permit for a sample of size n. (e.g. $n = 100$)
- You can still compute which power to expect for this sample size (given effect size, etc.)
- If power is too low, think about whether it is possible to improve power by measures other than the sample size (more sophisticated design, more precise measurements, different response definition–that permits to see larger effects)
- If sufficient improvements in power are impossible, think whether it makes sense to carry out the proposed investigation.
- If power is too high (not so common), you can save some resources.

What if several hypotheses tested?

- One hypothesis selected in advance as being of primary interest
  - Sample size computed for this hypothesis
- Several hypotheses equally interesting (e.g. questionnaire)
  - Sample sizes should be sufficient for all these hypotheses
  - *Multiple testing should be taken into account!*

Multiple testing I

- If many hypotheses are tested, chance rises to obtain some falsely rejected null hypotheses.
- Example 100 tests at level $\alpha = 0.05$, assume H_0 to be true in all cases.
- How many falsely rejected null hypotheses would we expect?
- *Five, and with independent test statistics probability* 0.994 *of at least one false rejection.*

Bonferroni correction for multiple testing

If total of k hypotheses:

- test each hypothesis at level α/k,
- equivalent: multiply al p-values by k.
- Ensures that “familywise error” (prob. of one or more false rejections) stays below α
- There are more complex multiple testing procedures that exploit special problem structures (e.g. Tukey test for all pairwise comparisons, Dunnett test for multiple comparisons with a control)

Examples

- Three hypothesis tests in an investigation gave following-values:
- Test I: 0.03, Test II: 0.12, Test III: 0.004
- Bonferroni–corrected p-values: Test I: 0.09, Test II: 0.36, Test III: 0.012
- After Bonferroni correction (for $\alpha = 0.05$) H_0 rejected only got test III

Consequence for power-sample size computations

If total of k hypotheses:

- sample size computations required at level α/k
- depending on k, considerably larger sample sizes needed!

Example

- One-sided z-test for $k = 1$ and $k = 10$, $\alpha = 0.05$
- Desired power 0.8
- Effect size 0.5σ
- Then $n_1 = 25$ ($k = 1$), and $n_{10} = 47$ ($k = 10$).

Comparing multiple treatments with control

- Dunnett test
- Optimum Allocation of sample size between k treatments and control:
- if sample size for each treatment is n, control group should get $\sqrt{k}n$ observations to minimize total sample variance.

Parametric versus Nonparametric Tests

- Parametric tests have a slightly higher power than non-parametric tests, if data follow the more stringent distributional assumptions (often normal distribution).
- If distributional assumptions for parametric test are not satisfied, nonparametric tests often have a higher power.

Sample size computations in equivalence and non-inferiority trials

- Sometimes sufficient to show that new treatment is equivalent to established intervention
- Reasonable when new treatment has fewer side effects, is less expensive, or has other benefits
- Exact equivalence cannot be shown with finite sample sizes–need to show *essential* equivalence
- In practice: specify difference δ, and call the treatments *essentially* equivalent, if they differ by no more than δ.
- Here alternative hypothesis is essential equivalence, null hypothesis is difference by more than given margin
- Power/sample size computation based on margin δ.

Repeated measurements and experimental units

- Notice that repeated measurements on the same animal do not lead to independent observations. Dependencies need to be taken into account in power/sample size computations!
- Further challenge: need to know about type and amount of dependence at planning stage
- Power computations should be done in terms of experimental units–the smallest unit to which treatment can be randomly assigned.
- Depending on type of investigation, experimental unit can be animal, part of animal (eye, leg), or collection of animals (herd, pen, aquarium)
- Dependencies sometimes exploited by clever design to reduce random variation—thus not necessarily bad

Further strategies to increase power

- Reduce Variation:
  - improve accuracy of measurements (if measurement error is an issue)
  - exploit dependencies-comparison done on otherwise similar subjects (if large interspecific variation)
    - homogeneous group(s), cross-over trials, matched pairs design

Group sequential and adaptive designs

- In group sequential designs, data are collected in batches
- between batches interim analyses, are conducted
- early stopping possible under previously defined circumstances (such as: negative treatment effect, negative side effects, futility, positive effect already clear at earlier stages)
- Adaptive Design: Some changes (e.g. sample size) in plan for later stages possible–needs careful specification of action plan in advance

Bias in (published) studies with low power

- Studies with low power tend to report upward biased effect size estimates
- If power is low, significant results are only obtained when data lead to large effect size estimates.
- If no significant results are obtained (cases when estimated effect size estimates are smaller), results are often not published or reported.
- Overall, low power per se does not lead to biased estimates–but leads to a biased picture in connection with common reporting & publication practices.

High power does not protect against other sources of bias

- bad randomization
- lack of (double-)blinding
- biased sampling in situations involving inference from sample to population
- non-response bias
- recall bias in retrospective studies . . .

Experiments versus observational studies

- Power/sample size computations work the same way both for experiments and observational studies
- Keep in mind, however, that it is considerably more difficult to establish causal relationships from observational studies

Unless lots of resources:

Do not try to answer all questions you may possibly have in one investigation

- Many covariates in a model –require large samples!
- When studying treatment for renal function, no point in investigating 7 breeds of cats, both male and female over a large age range unless you plan to take very large samples (why?)

Post-Hoc Power Analysis

- Sometimes power/sample size analysis done post-hoc, i.e. after collecting & analyzing the data, especially when no significant results were obtained.
- May help for sample size planning when designing further studies.
- *Be careful, what is wrong with the following argument?*
- Suppose we do not get a significant p-value, although a medium effect size was estimated from our data. A post-hoc power analysis shows a low power for this effect size.
- We then argue that the data suggest a medium size effect, and add that the only reason for not obtaining a significant p-value is that–unfortunately–the power was too low.

Literature

- Several textbooks on biostatistics and statistics in medicine contain material on power/sample size computations:
  - L.M. Friedman et al. Fundamentals of Clinical Trials, Springer 2010.
  - S.-C. Chow and J.-P. Liu. Design and Analysis of Clinical Trials: Concepts and Methodologies, Wiley 2004.
  - M. Schumacher, G. Schulgen. Methodik klinischer Studien: Methodische Grundlagen der Planung, Durchführung und Auswertung. Springer, 2008.
  - S.-C. Chow, M. Chang. Adaptive Design Methods in Clinical Trials. Chapman & Hall, 2011.

Literature

- Short discussion/reviews in papers:
  - E. McCrum-Gardner(2010): Sample size and power calculations made simple. International Journal of Therapy and Rehabilitation, 17, 1.
  - S.R: Jones et al. (2003) An introduction to power and sample size estimation. Emerg. Med. J. 20: 453-458.
  - Lenth, R. V. (2001), “Some Practical Guidelines for Effective Sample Size Determination,” The American Statistician, 55, 187-193.
  - Hoenig, John M. and Heisey, Dennis M. (2001), “The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis,” The American Statistician, 55, 19-24.

Summary of some key points

- Sample size computations require clear ideas about objective of study, type of data collected and statistical methods that will be used.
- Effect size and variability of observations are also needed but can be difficult to specify. (Pilot studies or adaptive sequential designs can help.)
- Due to uncertainties at planning stage, power computations will only be approximate (but nevertheless important)!
- If you cannot ensure sufficient power for your investigation, there is no point carrying it out. (Unless you are doing an exploratory/pilot study.)
- Good design and accurate measurement can help to improve power!

Free software for power-sample size computations

- **G*Power 3.1.9:** Windows and Mac, *http://www.gpower.hhu.de/*
- **PS:** Power and Sample Size Calculation version 3.0.43, 2011, free for for Windows, Mac, and Linux *http://biostat.mc.vanderbilt.edu/wiki/Main/PowerSampleSize*
- **R:** packages *pwr* and *samplesize* – power computations for parametric and nonparametric tests
- **POWER V3.0:** for logistic regression, Windows, free, provided by NIH *http://dceg.cancer.gov/tools/design/power*
- Web based sample size calculator (can also be downloaded): *http://www.stat.uiowa.edu/rlenth/Power/*

Power computations for many of the basic tests.

Commercial software for power-sample size computations

- **PASS 13:** power analysis and sample sizes for over 230 statistical tests and confidence intervals, Windows, commercial, free trial *http://www.ncss.com/software/pass/*
- **SPSS:** SamplePower, extra package
- **SAS:** PSS application, PROC POWER, PROC GLMPower, JMP

For statistics with means and differences in means, correlation, one-way and factorial analysis of variance (ANOVA), regression and logistical regression, survival analysis, equivalence tests and more.

If you are still unsure after this workshop. . .

Quoted from: Lenth (2006)

I receive quite a few questions that start with something like this: Ï'm not much of a stats person, but I tried [details...] – am I doing it right?" Please compare this with:
Ï don't know much about heart surgery, but my wife is suffering from ... and I plan to operate ... can you advise me?"
Folks, just because you can plug numbers into a program doesn't change the fact that if you don't know what you're doing, you're almost guaranteed to get meaningless results – if not dangerously misleading ones.

. . .

If your scientific integrity matters, and statistics is a mystery to you, then you need expert help. Find a statistician, and talk to her face-to-face if possible.