

Sample size calculations

Project: Discovering associations

Martin Otava

I-Biostat, Universiteit Hasselt
&
Janssen Pharmaceutica



Interuniversity Institute for Biostatistics
and statistical Bioinformatics

Why to bother?

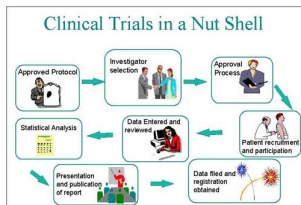
Two frameworks

- Moneyball (sabermetrics): all stats available



mindforsports.com

- Clinical trials: experiment to be done



padisysmobileclinicaltrials.wordpress.com

Two frameworks: Ebola, Zika

- First infection disease modelling:
 - What is the biological mechanism of the disease?
 - How is the disease spread?
 - What predictors are related to survival?
 - Use whatever data that are currently available
- Then vaccine development:
 - Plan the experiments to determine:
 - Is the vaccine safe?
 - Is the vaccine working?

Experimental design

- What is the research question?
- How the experiment will be organized?
 - What covariates do we want to control for?
 - How will we access randomization?
 - Technical possibilities included
- **What is the sample size?**

Importance of sample size

- Determines the power of the test
 - ⇒ determines our certainty, precision
 - ⇒ determines how big effect we will be able to detect
- How many subjects do we need to see meaningful effect?
- Using available sample size, how big effect can we detect?
- Does it make sense to perform experiment with current budget?

Points of view

- **Frequentist approach**
(length of confidence intervals or power)
- Bayesian approach
(optimizing utility function)

Power

Example: Systolic Blood Pressure

Description of the experiment:

- Two-sample experiment
- Compare systolic blood pressure (SBP) in a treatment group and control group
- SBP measured using a standard sphygmomanometer
- Treatment: we expect SBP to be reduced

Hypothesis testing

- H_0 against H_1
- Decision about H_0 : rejection or non-rejection
- Decision based on data available at hand
- Rejecting H_0 in favour of H_1 means that there is an evidence against H_0 in the data
- Not rejecting H_0 means that there is lack of evidence against H_0 in the data

Lack of evidence

- A. The H_0 is true
- B. The H_0 is not true, but we do not have enough **power**
 - **B** translates: the difference between H_0 and true underlying effect is smaller than we can see using our data set
 - We need more data to see it



Lack of evidence options

Hypothesis: treatment reduces SBP?



Is this relevant?

Possible decisions

	H_0 true	H_0 false
Reject H_0	Type I error	
Non-rejection H_0		Type II error

Errors example

In SBP example:

- **Type I error:**



When there is no treatment effect, the data might suggest a treatment effect

- **Type II error:**

When there is a treatment effect, the data might suggest no treatment effect

Trade off

- Cannot be minimized simultaneously
- I will always reject: Type II = 0, but Type I = 1
- I will never reject: Type I = 0, but Type II = 1
- Strategy: we control one and then we try to minimize the other

	H_0 true	H_0 false
Reject H_0	Type I error	
Non-rejection H_0		Type II error

Type I error

- Probability of false rejection H_0
- Translate: falsely finding effect, where there is none
- Typically fixed as **significance level**

⇒ Type I error can be controlled

Type II error

- Probability of missing rejection of H_0
- Translate: I do not reject, when there is really effect
- It is a probability \Rightarrow it depends on true effect size!
- We usually use power instead of Type II error

Power

- 1 - Type II error
- Probability of rightful rejection of H_0
- Translate: we find the true effect
- We can increase power with increasing sample size
- As Type II error, it depends on true effect size!

Dependence on true effect size

Hypothesis: treatment reduces SBP?



Is this relevant?

Power: summary

Power of the experiment depends on

- Test and hypotheses: fixed
- Desired significance level: fixed
- True underlying effect: cannot be influenced
- True underlying variability: cannot be influenced
- **Sample size**: only thing that we can improve

Summary

Why to do sample size calculation?

- We seek for biologically relevant effect size
- **Sample size** influences **power**
- **Power** determines the ability to **rightfully reject** null hypothesis H_0 for biologically relevant effect size
- **Rejection** of H_0 validates the observed effect size as **statistically significant**
- **Statistically significant effect size that is biologically relevant** is what we seek for

Wrong sample size impacts

Too small sample size?

- We will miss biologically relevant differences
- Experiment may end up as wasting of time/resources/harming patients

Too big sample size?

- We will be able to see effects that are too tiny to be important
- We will certainly waste resources/harm patients

That is why we need sample size calculation!

Steps in a sample size calculation

Example: Systolic Blood Pressure

Description of the experiment:

- Two-sample experiment
- Compare systolic blood pressure (SBP) in a treatment group and control group
- SBP measured using a standard sphygmomanometer
- Treatment: we expect SBP to be reduced

Question:

How many patients should I include in the treatment and control group?

Steps in a sample size calculation

Checklist to determine sample size

- 1 Specify parameter, hypothesis and test
- 2 Specify significance level
- 3 Specify effect size
- 4 Obtain values or estimates of other parameters needed
- 5 Specify a value for the power

Step 1: Specify parameter and hypothesis

- **First and most important step!**
- **What do you want to test**, what is the goal of the study?
- SBP: Treatment is supposed to reduce blood pressure, thus:

$H_0 : \mu_T \geq \mu_C$, ie mean SBP same for treatment and control

$H_1 : \mu_T < \mu_C$, ie mean SBP smaller for treatment

- Parameter that is tested: $\delta = \mu_T - \mu_C$:

$$H_0 : \delta \geq 0$$

$$H_1 : \delta < 0$$

Step 1: Specify test on parameter

- Two samples: mean comparison
- Let us assume normality and homoscedasticity
- **Test:** two sample t-test with one sided hypothesis

Step 1: Specify parameter, hypothesis and test

- Designed experiment **different** than modelling!
- Computing sample size for one particular hypothesis, **one statistical test**
- Clear idea what to test and how to treat the results **before designing** the experiment
- Well designed experiment does not need complex modelling techniques

Step 2: Specify significance level

- **Significance level** = threshold on the Type I error probability
- Usually ranging between 0.001 to 0.1
- Most of the time: $\alpha = 0.05$

Step 2: Specify significance level

"The value for which $P=0.05$, or 1 in 20, is 1.96 or nearly 2; it is **convenient to take this point as a limit in judging whether a deviation ought to be considered significant or not**. Deviations exceeding twice the standard deviation are thus formally regarded as significant. Using this criterion we should be led to follow up a false indication only once in 22 trials, even if the statistics were the only guide available."

(R.A. Fisher, Statistical Methods for Research Workers, 1925)

Step 2: Specify significance level

Coefficients:	Estimate	Std. Error	t value	$\text{Pr}(> t)$
(Intercept)	10.699	12.251	0.873	0.390
hp	-0.047	0.014	-3.366	0.002
gear	0.580	1.293	0.448	0.658
am	4.167	2.093	1.991	0.057
qsec	0.400	0.520	0.770	0.448
drat	1.494	1.682	0.889	0.382

Step 2: Specify significance level

"This is an arbitrary, but convenient, level of significance for practical investigator, but it does not mean that he allows himself to be deceived once in every twenty experiments. **The test of significance only tells him what to ignore, namely all experiments in which significant results are not obtained.** He should only claim that a phenomenon is experimentally demonstrable when he knows how to design an experiment so that it will rarely fail to give a significant result. Consequently, isolated significant results which he does not know how to reproduce are left in suspense pending further investigation.

(R.A. Fisher, The Statistical Method in Psychical Research, 1929)

Step 3: Specify effect size

Effect size: the treatment effect you want to detect in your test

To define effect size, ask yourself questions like:

- What treatment effect is of scientific interest?
- What effect do you expect or hope to see?

Step 3: Specify effect size

SBP example:

- Relevant treatment effect might be -15 mm Hg
- Translate: compared to control, the treatment reduces blood pressure with 15 mm Hg
- If you were the patient, the benefits of reducing SBP by 15 mm Hg outweigh the cost, inconvenience, and potential side effects of this treatment

Step 3: Specify effect size

BUT keep in mind that:

- The smaller the effect size, the larger the sample size
- Be reasonable with choice of effect size
- Should be determined together with scientist
- Problem need to be clearly communicated

Step 3: Specify effect size

Example of issues:

MD Go ahead, reduce sample size, because we are pretty sure that treatment will do better than this!

St But what if not? What is the clinically relevant effect size?

MD [This is the effect we expect to see. That is our effect size].

MD I think power 70% is really enough for us. I see we can reduce the number of patients by decreasing the power.

St But what is the relevance of such a result? Is 70% worth even conducting experiment?

MD This book says that medium effect size is five.

St Effect size is always context dependent, we are not interested in experiment of anyone else, we need an effect size relevant for you.

Another example: <https://www.youtube.com/watch?v=Hz1fyhVOjr4>

Step 3: Specify effect size

Summary:

- Effect size quantifies the effect of interest
- Should be determined based on scientific knowledge
- **Never use pilot study to obtain estimate for effect size!**

Step 4: Obtain values or estimates of other parameters

- Power can depend on **other parameters**
- These parameters can be completely unrelated to the hypotheses
- BUT they can highly influence results
- SBP example: variance of the measurements in the planned experiment
- How to find a value for these parameters?
 1. Based on scientific knowledge (experience)
 2. Use historical data
 3. Conduct pilot study

Step 4: Obtain values or estimates of other parameters

1. Based on expert scientific knowledge:

- Reliability of expert?
- Scientific consensus about the problem?
- Novelty of experiment?
- Do not overestimate nor underestimate expert opinion!

Step 4: Obtain values or estimates of other parameters

2. Use of historical data:

- In-house data
- Data from literature and public databases
- Compare designs
- Reliability of the data?
- Be careful to check that the same parameter is estimated!

Step 4: Obtain values or estimates of other parameters

Issues with historical data/scientific knowledge:

- Manufacturer of sphygmomanometers published: $sd = 2.5\text{mmHg}$, but this reflects variation in readings made on same subject
- In previous study, only women were included, but in planned study mix of gender \Rightarrow variation due to gender not included
- Historical data were collected using different measurement device or even different units

Step 4: Obtain values or estimates of other parameters

3. Conduct pilot study:

- Same experiment design with small sample size
- To estimate additional parameters (typically variance)
- It has to be sufficiently large to be reliable!
- Do not overestimate small pilot studies
- **Should not be used to look at the main effect to be studied**

Step 5: Specify a value for the power

How to choose the **power**?

- The power is the probability with which the desired effect size will be detected
- 80% is most common but also somewhat minimal
- Sometimes higher power is more appropriate (85% or 90%)
- As power increases, required sample size increases fast

Summary for SBP

Choices for SBP example:

- 1 Specify hypothesis test on parameter:

$$H_0 : \delta \geq 0$$

$$H_1 : \delta < 0$$

Tested with one sided t-test.

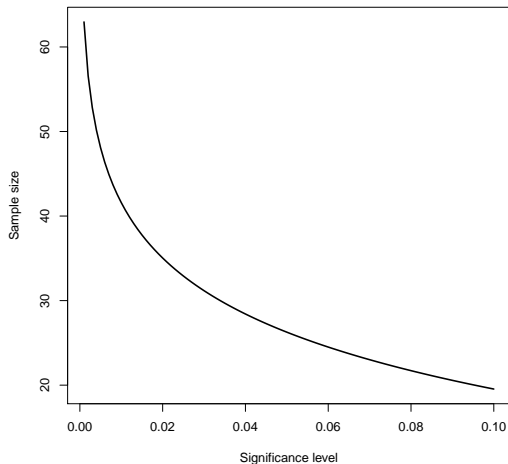
- 2 Specify significance level: $\alpha = 0.05$
- 3 Specify effect size: $\delta = -15\text{mmHg}$
- 4 Obtain values or estimates of other parameters needed:
 $sd = 20\text{mmHg}$
- 5 Specify a value for the power: $\beta = 0.85$

Resulting sample size: 27 patients/group.

Visualization

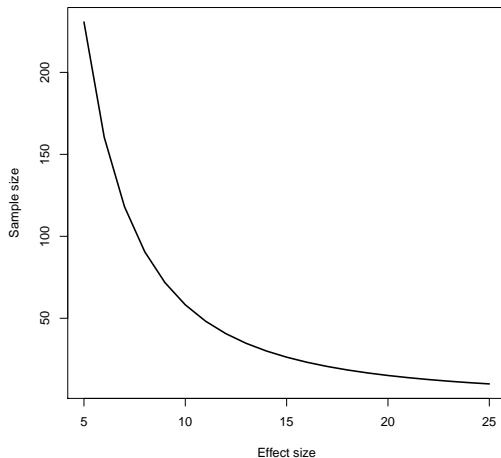
Sample size curves

SBP example: $\alpha = ?$, $\delta = 15$, $sd = 20$, $\beta = 0.85$



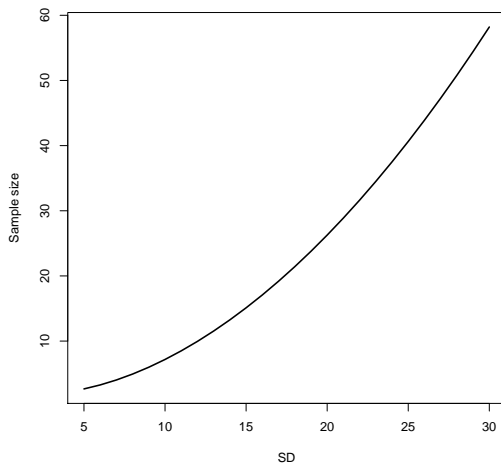
Sample size curves

SBP example: $\alpha = 0.05$, $\delta = ?$, $sd = 20$, $\beta = 0.85$



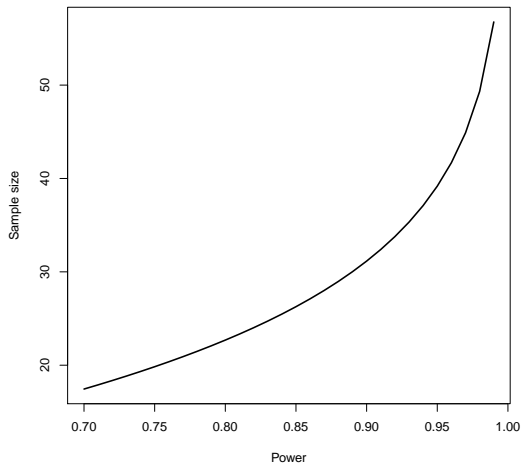
Sample size curves

SBP example: $\alpha = 0.05$, $\delta = -15$, $sd = ?$, $\beta = 0.85$



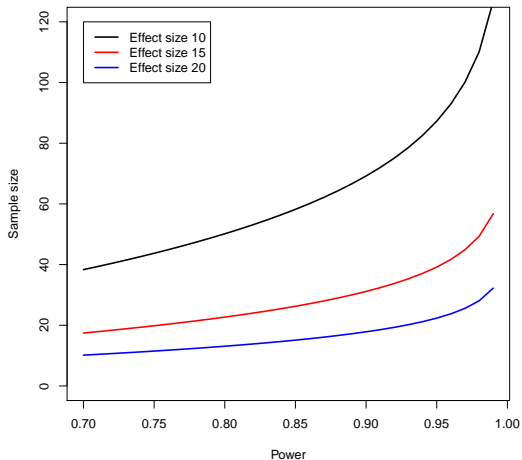
Sample size curves

SBP example: $\alpha = 0.05$, $\delta = -15$, $sd = 20$, $\beta = ?$



Sample size curves

SBP example: $\alpha = 0.05$, $\delta = ?$, $sd = 20$, $\beta = ?$



Theoretical derivation: normal case

Notation

- Two samples comparison
- SBP example
- Simplification: assume that σ is known
- $Y_{Cj} \sim N(\mu_C, \sigma^2)$ values of patients in control group
- $Y_{Tj} \sim N(\mu_T, \sigma^2)$ values of patients in treated group
- $\delta = \mu_T - \mu_C$
- $N = ?$ sample size per group

Hypothesis testing

- Test of hypothesis

$$H_0 : \delta \geq 0$$

$$H_1 : \delta < 0$$

- $\bar{Y}_C = \frac{1}{N} \sum_{j=1}^N Y_{Cj}$

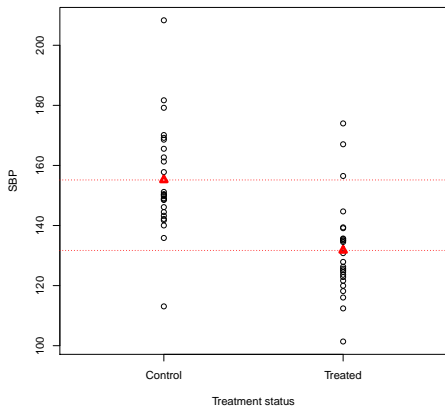
- $\bar{Y}_T = \frac{1}{N} \sum_{j=1}^N Y_{Tj}$

- $\hat{\delta} = \bar{Y}_T - \bar{Y}_C$

- That implies

$$\bar{Y}_T - \bar{Y}_C \sim N\left(\mu_T - \mu_C, \frac{2\sigma^2}{N}\right)$$

$$\hat{\delta} \sim N\left(\delta, \frac{2\sigma^2}{N}\right)$$



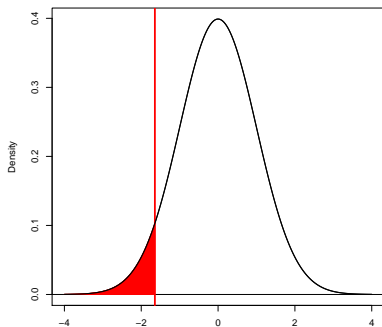
Test statistics: Z test

- We assume that σ is known
- Therefore, from $\hat{\delta} \sim N\left(\delta, \frac{2\sigma^2}{N}\right)$ it holds $\frac{\hat{\delta} - \delta}{\sqrt{2\sigma^2}} \cdot \sqrt{N} \sim N(0, 1)$
- Under H_0 : $\frac{\hat{\delta}}{\sqrt{2\sigma^2}} \cdot \sqrt{N} \sim N(0, 1)$
- Hence $S = \frac{\hat{\delta}}{\sqrt{2\sigma^2}} \cdot \sqrt{N}$ is our statistics
- Evidence against $H_0 : \delta \geq 0 \Rightarrow$ if S is high negative number

Significance level

Under null hypothesis:

- $P_{H_0}(S < z_\alpha) = \alpha$
- ⇒ rejecting if $S < z_{0.05}$, we have only 5% chance of wrong rejection
- ⇒ rejecting if $\hat{\delta} < z_\alpha \cdot \sqrt{\frac{2\sigma^2}{N}}$

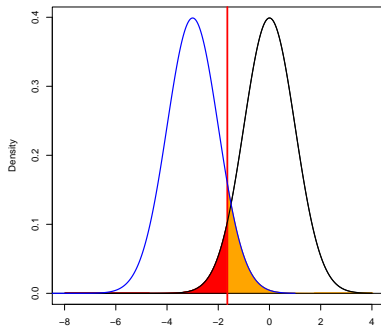


Tough part: include control of power

Under alternative hypothesis:

- $\frac{\hat{\delta} - \delta}{\sqrt{2\sigma^2}} \cdot \sqrt{N} \sim N(0, 1)$
- and we want

$$P_{H_\delta} \left(\hat{\delta} < z_\alpha \cdot \sqrt{\frac{2\sigma^2}{N}} \right) = \beta$$
- subtract δ and multiply by $\sqrt{\frac{N}{2\sigma^2}}$ to get variable with $N(0, 1)$
- $$P_{H_\delta} \left(\frac{\hat{\delta} - \delta}{\sqrt{2\sigma^2}} \sqrt{N} < z_\alpha - \delta \sqrt{\frac{N}{2\sigma^2}} \right) = \beta$$



Formula for sample size

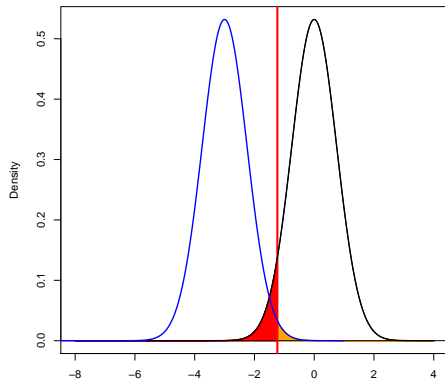
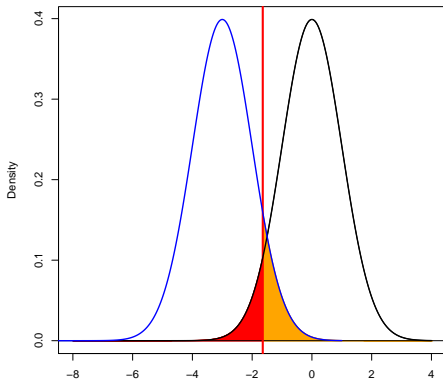
- $\Phi \left(z_{\alpha} - \delta \sqrt{\frac{N}{2\sigma^2}} \right) = \beta$

- $z_{\alpha} - \delta \sqrt{\frac{N}{2\sigma^2}} = z_{\beta}$

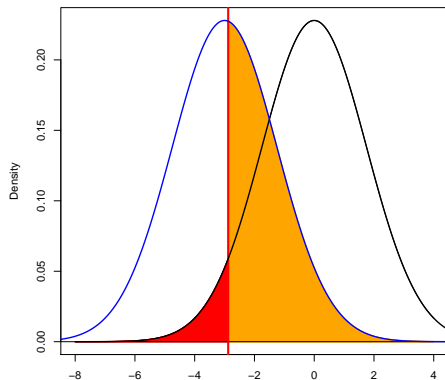
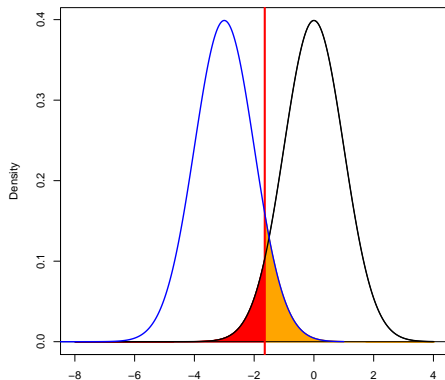
- Finally, formula for samples size calculation:

$$N = 2 \cdot (z_{\alpha} - z_{\beta})^2 \cdot \frac{\sigma^2}{\delta^2}$$

Sample size effect



Sample size effect



Notation comment

- Our final formula:

$$N = 2 \cdot (z_{\alpha} - z_{\beta})^2 \cdot \frac{\sigma^2}{\delta^2}$$

- Note: $z_{\alpha} = -z_{1-\alpha}$

$$N = 2 \cdot (z_{1-\alpha} + z_{\beta})^2 \cdot \frac{\sigma^2}{\delta^2}$$

- We denote β as power, but sometimes $1 - \beta$ denotes power. Then:

$$N = 2 \cdot (z_{\alpha} - z_{1-\beta})^2 \cdot \frac{\sigma^2}{\delta^2}$$

- **BE CAREFUL WHEN USING LITERATURE!**

Extensions of the framework

Two-sided hypothesis

- Theoretical derivation assumed one sided hypothesis
- What if hypothesis would be two-sided:

$$H_0 : \delta = 0$$

$$H_1 : \delta \neq 0$$

- Then for normal case:

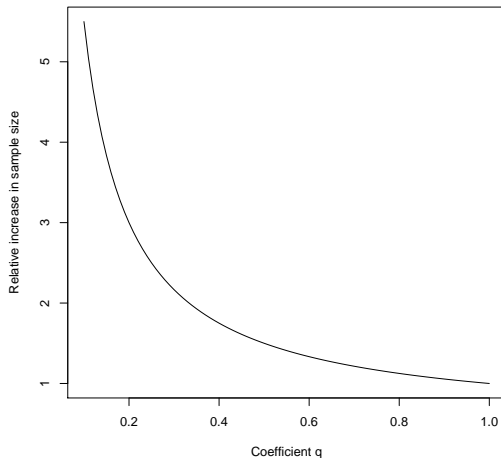
$$N = 2 \cdot \left(z_{\frac{\alpha}{2}} - z_{\beta} \right)^2 \cdot \frac{\sigma^2}{\delta^2}$$

Unbalanced design

- Theoretical derivation assumed same number of patients in both groups
- It can occur that we have to plan unbalanced design
- One group has N observations, other one qN , with $q \in (0, 1)$
- Then for normal case:

$$N = \frac{q+1}{q} \cdot (z_\alpha - z_\beta)^2 \cdot \frac{\sigma^2}{\delta^2}$$

Unbalanced design: increment in sample size



Heteroscedasticity

- Theoretical derivation assumed same variance for both groups
- What if variance is different, σ_1^2 and σ_2^2 ?
- Then for normal case:

$$N = \frac{\sigma_1^2 + \sigma_2^2}{\delta^2} \cdot (z_\alpha - z_\beta)^2$$

Case of t-test

- Variance σ^2 unknown
- In final experiment, it will be estimated from the data
- If we adjust for this, we need to use t-distribution quantiles $t_{\alpha,n}$
- **Degrees of freedom n depends on sample size**
- Iterative procedures are necessary
- Software solutions quite straightforward
- By hand: normal case as approximation (big samples)

More groups

- What if we have more groups to compare?
- One-way ANOVA setting
- Getting difficult, because degrees of freedom of F distributions depends on N
- Iterative procedures are necessary
- Software solutions quite straightforward

ANOVA mean structure

- Example: difference of interest $\delta = 5$; assume 5 groups
- Nuisance parameter: within variability
- Between variability: depends on mean structure
- Option 1: 10 - 10 - 10 - 10 - 15 $\Rightarrow BW = 2.24$
- Option 2: 10 - 15 - 10 - 15 - 10 $\Rightarrow BW = 2.74$
- Different between variability \Rightarrow which one to choose
- Solution may be test all and retain worse case
- Many software: enter directly between variability

Binary response

- Response of interest has binary value 0/1
- Parameter π represents probability of success
- Two-groups comparison setting
- Control $X \sim Be(\pi_C)$ and Treatment $Y \sim Be(\pi_T)$
- We are interested in difference between success probabilities
 $\delta = |\pi_T - \pi_C|$

Binary response

- Then one sided hypothesis:

$$N = (z_{\alpha} - z_{\beta})^2 \cdot \frac{\pi_T(1 - \pi_T) + \pi_C(1 - \pi_C)}{\delta^2}$$

- Problem: unknown π_T and π_C appears in formula
- You need to have an idea about their value, not only difference

Poisson distribution

- Count data
- Represents number of events during some period of time
- No restriction on maximum
- Parameter λ represents frequency of events
- If $\lambda > 20$ asymptotic starts to work
- $X \sim \text{Pois}(\lambda) \Rightarrow X \approx N(\lambda, \lambda)$
- Then, comparing two frequencies is same as comparing two means of normal distributions, while variances are different

Survival analysis

- Survival response: time to event
- Censored samples
- Only observed responses counts (not censored ones)
- Needs to be adjusted for censoring proportion (similar as missing data)

Multiple goals

A. Clear primary endpoint

- Use primary goal to determine sample size
- Secondary endpoints: sample size calculation to see if they are meaningful

B. Multiple equally important endpoints

- Sample size calculation has to be done for all the goals
- Choose maximal sample size needed

Multiple goals

C. Multiple regression: covariates to investigate

- E.g. age, gender, interaction term
- Similarly as previous case: determine what is primary interest
- Compute sample size for one or more covariates: take maximum
- Consider the interaction term necessity
- Remember your choices while interpreting results

Multiple comparisons

- Several tests are performed simultaneously
- We want to be protected against any false finding (family-wise error rate)
- We can have multiple t-test, but also combination of the tests
- How to incorporate this into the sample size calculation framework?
- Adjusting confidence level is usually sufficient solution
- Simulations can provide solution in more complex cases

Correlated data

Correlated data

- N individual observations clustered into N_{cl} clusters, each with n observations
- Hence: $N = nN_{cl}$
- Repeated measurements per patient, patients in same hospital, etc.
- Individual observations carries less information compared to case of independence
- Example: measure height of 10 men or one man for 10 consecutive days
- Effective sample size
- Adjustment for sample size calculations is needed
- **Goal: compare overall mean of several groups**

Correlated data

- Effective sample size: depends on data structure
- If correlation within cluster is $|\rho| = 1$, all observations carry same information
- Very conservative approach: calculate required N_{ind} and take put $N_{cl} = N_{ind}$, i.e. taking nN_{ind} observations in total
- BUT: if $|\rho| < 1$, this overestimates sample size needed
- There exist formulas to tackle the problem
- Dependent on structure of correlation: CS, AR(1), etc.

Correlated data

- Consult appropriate literature (e.g. Faes et al, 2009)
- Software solutions (e.g. `longpower` in R)
- Simulation to determine sample size (see later)
- For special cases simple approximations enough
- Correlation structure estimated from the pilot study

Correlated data: compound symmetry

- All observations within subject have pairwise same correlation
- Compute sample size N_{ind} as if your data are independent
- Inflate N_{ind} to adjust for correlation
- Number of clusters: $N_{cl} = \frac{N_{ind}}{n} \cdot [1 + (n - 1)\rho]$
- with n number of observations per cluster

Correlated data: compound symmetry

- Hypothetical example:
- $\rho = 0.2$, $n = 8$ observations per cluster
- Sample size computation for independence: $N = 150$
- $N_{cl} = 150/8 \cdot [1 + 7/5] = 45$
- We need 45 clusters with 8 observations each; in total 360 observations

Longitudinal data: autoregressive structure

- First order autoregressive structure
- Example: patient with n consecutive observations
- Compute sample size N_{ind} and inflate it, as above
- $$N_{cl} = \frac{1+\rho}{n-(n-2)\rho} N_{ind}$$
- with n number of observations per cluster (i.e. timepoints)

Longitudinal data: slopes comparison

- Different goal: compare the slopes of two groups
- Control: $Y_{ij} = \beta_{0C} + \beta_{1C}t_{ij} + \varepsilon_{ij}$
- Treated: $Y_{ij} = \beta_{0T} + \beta_{1T}t_{ij} + \varepsilon_{ij}$
- Hypothesis: $\beta_{1C} = \beta_{1T}$
- Assume compound symmetry again
- $N_{cl} = \frac{N_{ind}}{n} \frac{1-\rho}{s_t^2}$
- $s_t^2 = \sum_{j=1}^J (t_j - \bar{t})^2$ depends on structure of the timepoints

Longitudinal data: Note

- Group average comparison: $N_{cl} = \frac{N_{ind}}{n} \cdot [1 + (n - 1)\rho]$
- Higher correlation **increases** sample size
- Slope comparison: $N_{cl} = \frac{N_{ind}}{n} \frac{1-\rho}{s_t^2}$
- Higher correlation **decreases** sample size
- Group average relies on amount of clusters/patients
- Slope comparison is within-patient/cluster measure: higher within-correlation is better
- **Sample size always depends on the test and context!**

Longitudinal data: Extensions

- Various different contrast tests are needed
- Formulas depend on the context
- Search literature and software
- Use simulations!

Simulation: universal solution

Simulation method

- Possible in any setting
 - Especially useful in complex situations
1. Fix sample size as N
 2. Fix other parameters, δ , α
 3. Simulate huge number of experiments and test
 4. Estimate power
 5. Change N accordingly and run 1-4 until you reach desired power.
 6. Use final N as your sample size

Simulation method: initial N

- Use "good guess"
- It does not really matter
- Only purpose is to get idea of power for this N
- Then change N accordingly
- Possible to run over grid of N
- If you lack time (real or computational), often faster to do it "by hand"
- Always keep in mind precision that you need

Simulation method: other parameters

- This step is essentially same as in traditional sample size calculation
- You need to fix effect size, SD etc.

Simulation method: simulate many data sets

- Core step and often the trickiest
- Simulate experiment that has all the properties that you fixed already
- E.g. two-sample t-test: simulate two groups with difference δ in means, SD as fixed in previous step and with N fixed
- It may get more laborous or more complex situations
- Often, you can simplify it by ignoring variables that are not tested
- Perform the appropriate statistical test and retain p-value
- Repeat this step B times, with sufficient B
- Can be computationally intensive

Simulation method: get power and repeat

- Compare retained p-values with α level
- How many tests are significant?
- Proportion of significant test among B replicates equals power
- If the power is too high or too low, change initial N and repeat the exercise

Simulation method: examples

- Examples follow in provided R codes
- Possible in SAS and other softwares as well
- Be always aware of its computational intensity
- Keep in mind the precision that is needed

Software

General

- Formulas for complicated cases are complex
- Once they are programmed, their application is usually straightforward
- Be very careful with any software
- Even commercial software can be misleading

Free software

- R
- Online tools
- Java applications (Russel Lenth)
- G*Power
- Again, be **VERY CAREFUL** with all of them (except R stats procedures)

Commercial software

- SAS, Stata, Statistica, StatXact, SPSS,...
- Complex statistical softwares always have procedures to compute sample size
- Sometimes dealing only with simple cases

- Specialized software often handles more situations
- nQuery, N
- PASS
- East (Cytel)
- Pharsight Trial Simulator

Additional comments

Budget limitation

- When available sample size insufficient, focus on other aspects of study quality
- Consider improvements of the study design (blocking, stratification)
- Consider changes in hypothesis (e.g. interactions), effect size
- If no improvement possible, cancel the experiment

Missing observations

- Especially important when dealing with human patients
- Computed sample size applies for complete data set
- Actual power can be smaller when observations are missing
- Necessary to adjust our sample size calculations to possible missingness
- Simply by multiplying resulting N with constant representing amount of dropouts
- Pilot study or literature can give us an idea

Russ Lenth's advices

NOT to do

- Retrospective power
- Specify T-shirt effect sizes

GOOD to do

- Use power prospectively
- Put science before statistics
- Do pilot studies

Retrospective power

- Wrong reasoning:

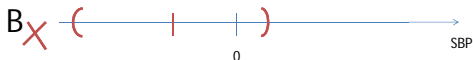
1. Conducting experiment that results as non-significant
2. Calculate "observed power", as if observed effect is true
3. Conclude about evidence in favour of H_0

- Explanation why is it wrong:

- Power is related to true effect not observed one
- Any calculations with observed values are in relationship with results of the hypothesis testing
- It says same thing as p-value, there is no extra information

Confidence intervals

- Give us additional information about the precision
- They actually tell us most likely value of true effect
- Correct thing to look at instead of "observed power"



Sample size for confidence intervals

- Base sample size on length d instead of effect δ
- Power: What is the smallest effect we want to see significant?
- CI: What is the precision of estimation we want to achieve?
- Similar theoretical background
- Different interpretation and outcome

References

- 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.
- Machin, David and Campbell, Michael Joseph (1987). *Statistical tables for the design of clinical trials*, Oxford
- Faes, Christel, Molenberghs, Geert, Aerts, Marc, Verbeke, Geert and Kenward, Michael G. (2009). "The Effective Sample Size and an Alternative Small-Sample Degrees-of-Freedom Method," *The American Statistician*, 63

Acknowledgements for materials

- Tomasz Burzykowski
- Luc Bijnen
- An Cremers
- Helena Geys
- Ziv Shkedy