HELP Statistics lecture series

How many patients do you need?

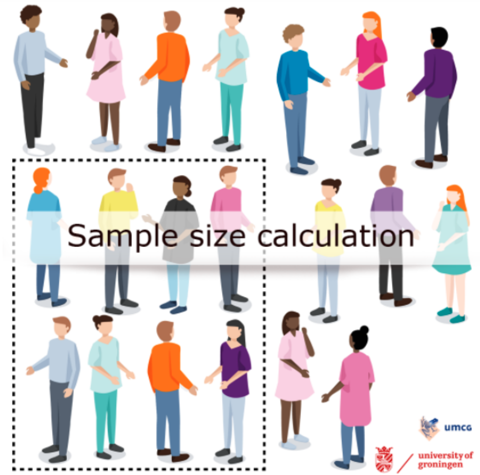
Katalin Tamási, PhD

Departments of Epidemiology and Neurosurgery

Unit of Medical Statistics and Decision Making

March 15, 2022

Introduction



https://edubox.nl/Instructie2016Html5.aspx#section=leereenheid&itemnr=1&itemid=&leereenheidid=3860

Why calculate sample size?

Imagine you could design your dream study about Wonderdrug, a new treatment. What would it look like?

How effective is Wonderdrug compared to standard treatment?

Measure of effectiveness: Proportion of people feeling better after treatment

1. Testing the whole population to obtain proportion of people feeling better after taking Wonderdrug: $\pi_1 = 0.6$
2. Erase, start over
3. Testing the whole population to obtain proportion of people feeling better after taking standard treatment: $\pi_2 = 0.3$
4. Calculating population treatment effect: $\pi_1 - \pi_2 = \delta = 0.3$

Why calculate sample size?

No access to whole population (+ no time travel possible)?

Next best thing: random sampling

- Wonderdrug sample: p_1
- Standard sample: p_2
- Calculating *observed* treatment effect: $p_1 - p_2 = d$
- + Inferential statistics:
  - What does $d = 0.3$ $(95\%\ CI: 0.1, 0.5; p = 0.003)$ mean?

How many patients do you need to:

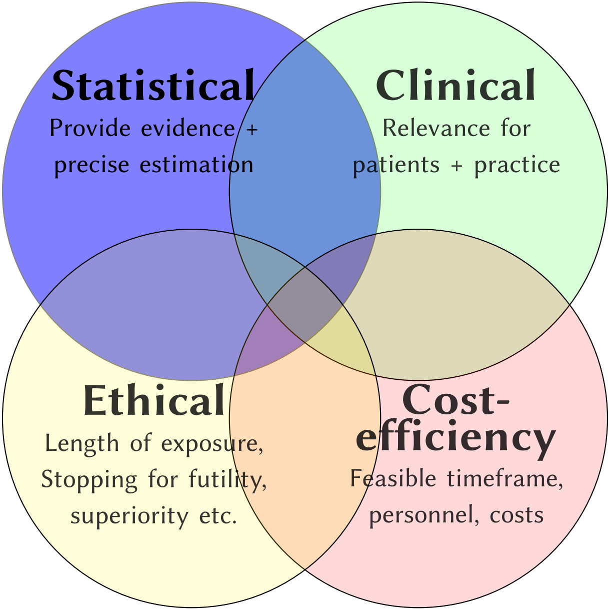
- Obtain reliable evidence of a treatment effect δ if it exists?
- Estimate the treatment effect δ precisely?

→ Sample size calculation:

Most popular request in my consultation practice

- Motivates to formulate assumptions, hypotheses in advance
- Research integrity, reproducibility: Evidence-based medicine
- Requirement for ethical reviews, grants, publications, etc.

Considerations



Statistical

Provide evidence + precise estimation

Clinical

Relevance for patients + practice

Ethical

Length of exposure, Stopping for futility, superiority etc.

Cost-efficiency

Feasible timeframe, personnel, costs

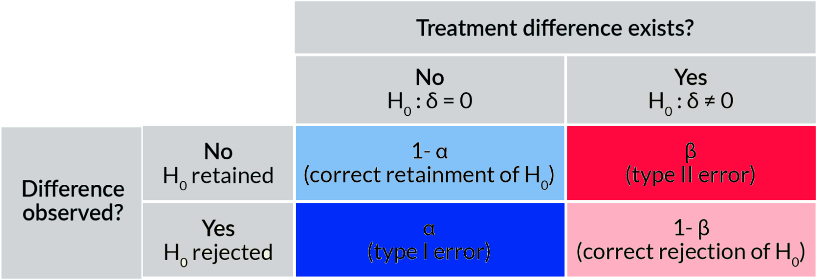
Recap on null hypothesis significance testing

H_0: There is no difference between the Wonderdrug and standard populations: $\delta = 0$

H_1: There is a difference in either direction between the Wonderdrug and standard populations: $\delta \neq 0$

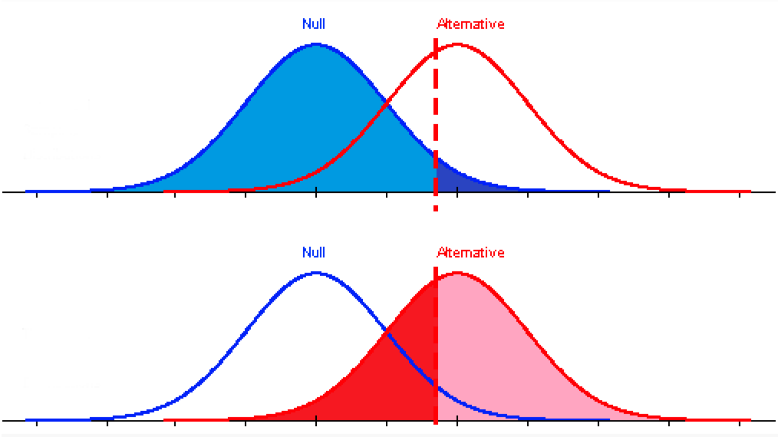
p value

The probability of observing a sample treatment difference of d or larger in either direction *assuming that the null hypothesis is true*: $P(d|H_0)$.



		Treatment difference exists?	
		No $H_0 : \delta = 0$	Yes $H_0 : \delta \neq 0$
Difference observed?	No H_0 retained	1- α (correct retainment of H_0)	β (type II error)
	Yes H_0 rejected	α (type I error)	1- β (correct rejection of H_0)

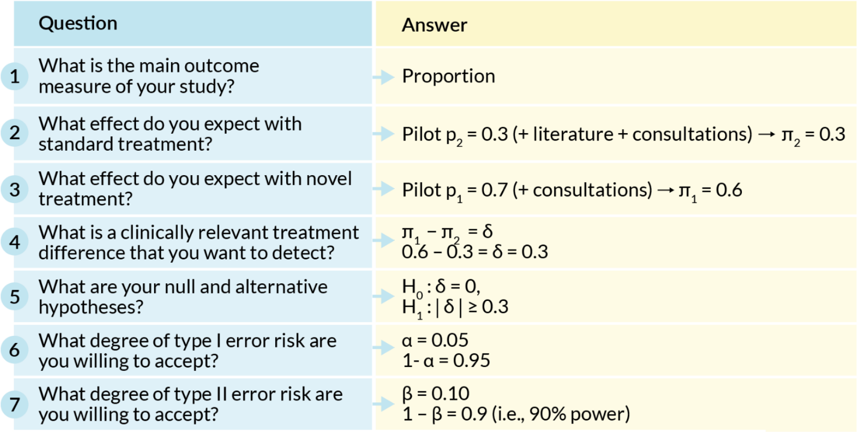
Recap



- Population parameters, e.g., treatment effect ($\pi_1 - \pi_2 = \delta$)
- Spread (for proportions: $\sqrt{\frac{\pi(1-\pi)}{n}}$, for means: σ)
- Risk of Type I error (α)
- Risk of Type II error (β) / power (1-β)

Designing your study

Main outcome measure → Sample size calculation → Main analysis

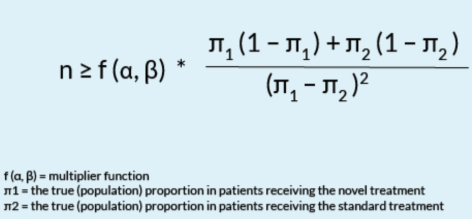


	Question	Answer
1	What is the main outcome measure of your study?	Proportion
2	What effect do you expect with standard treatment?	Pilot $p_2 = 0.3$ (+ literature + consultations) → $\pi_2 = 0.3$
3	What effect do you expect with novel treatment?	Pilot $p_1 = 0.7$ (+ consultations) → $\pi_1 = 0.6$
4	What is a clinically relevant treatment difference that you want to detect?	$\pi_1 - \pi_2 = \delta$ $0.6 - 0.3 = \delta = 0.3$
5	What are your null and alternative hypotheses?	$H_0 : \delta = 0$, $H_1 : \lvert \delta \rvert \geq 0.3$
6	What degree of type I error risk are you willing to accept?	$\alpha = 0.05$ $1 - \alpha = 0.95$
7	What degree of type II error risk are you willing to accept?	$\beta = 0.10$ $1 - \beta = 0.9$ (i.e., 90% power)

Sample size calculations

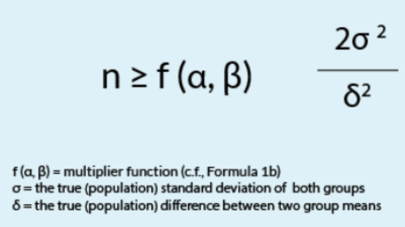
Comparing
two independent proportions

Comparing
two independent means



$$n \geq f(\alpha, \beta) * \frac{\pi_1(1-\pi_1)+\pi_2(1-\pi_2)}{(\pi_1-\pi_2)^2}$$

$f(\alpha, \beta)$ = multiplier function
$\pi 1$ = the true (population) proportion in patients receiving the novel treatment
$\pi 2$ = the true (population) proportion in patients receiving the standard treatment



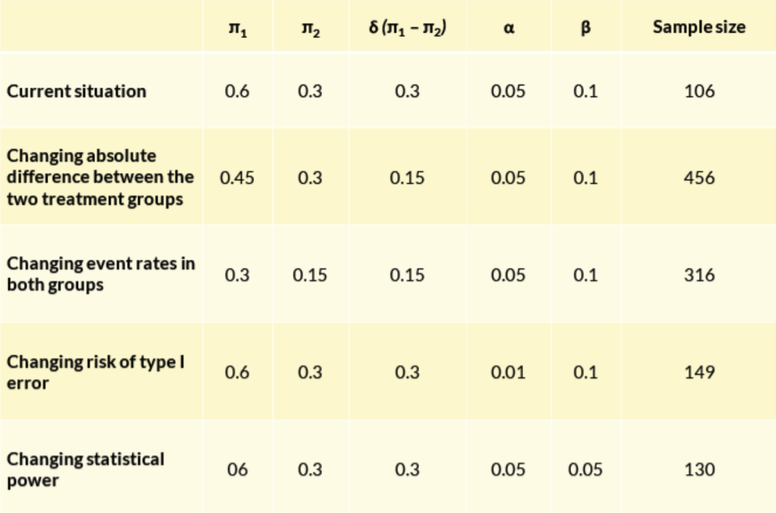
$$n \geq f(\alpha, \beta) \quad \frac{2\sigma^2}{\delta^2}$$

$f(\alpha, \beta)$ = multiplier function (c.f., Formula 1b)
σ = the true (population) standard deviation of both groups
δ = the true (population) difference between two group means

n: minimum required sample size per group
N = 2*n: total minimum required sample size

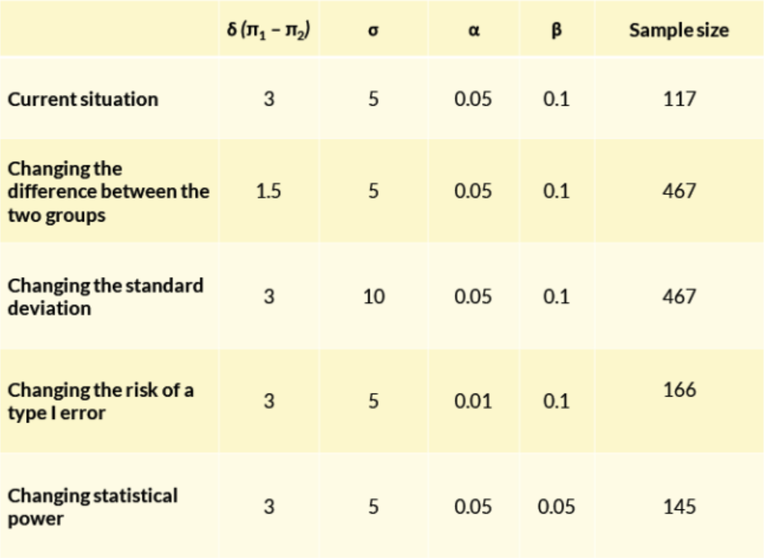
Some scenarios with proportions

Beware of fragility of your assumptions:
→ Consider a range of outcomes
→ Be realistic to give your study the best chances at succeeding



	π_1	π_2	$\delta\ (\pi_1 - \pi_2)$	α	β	Sample size
Current situation	0.6	0.3	0.3	0.05	0.1	106
Changing absolute difference between the two treatment groups	0.45	0.3	0.15	0.05	0.1	456
Changing event rates in both groups	0.3	0.15	0.15	0.05	0.1	316
Changing risk of type I error	0.6	0.3	0.3	0.01	0.1	149
Changing statistical power	06	0.3	0.3	0.05	0.05	130

Some scenarios with means



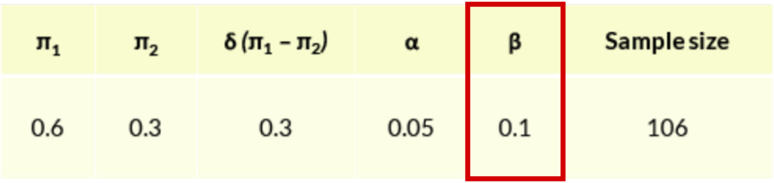
	$\delta (\pi_1 - \pi_2)$	σ	α	β	Sample size
Current situation	3	5	0.05	0.1	117
Changing the difference between the two groups	1.5	5	0.05	0.1	467
Changing the standard deviation	3	10	0.05	0.1	467
Changing the risk of a type I error	3	5	0.01	0.1	166
Changing statistical power	3	5	0.05	0.05	145

Calculations can be reproduced by R code in E-learning module

Power calculations

Given

- Population parameters, e.g., effect size (δ)
- Risk of Type I error (α)
- Feasible sample size (N)



π_1	π_2	$\delta\ (\pi_1 - \pi_2)$	α	β	Sample size
0.6	0.3	0.3	0.05	0.1	106

→ Achievable power ($1 - \beta$)

Lots of statistical packages can help (see Resources slide)

Outcome I: Power to correctly reject H_0

Suppose you recruited 106 people (53 per group) based on your sample size calculation and obtained

- d = 0.3 (95 % CI: 0.1, 0.5, p = 0.003)

Confidence intervals

Assume we take repeated random samples from the population and for each sample we calculate a 95% CI, then in the long run 95% of these CI's will include the population treatment effect δ.

- $p < 0.05$ = 95% CI does not contain 0: We found strong evidence against H_0
  $\rightarrow$ We reject H_0
- Relatively narrow CI's: δ was precisely estimated, well-powered study

Outcome II: Power to correctly retain H_0

Suppose you recruited 106 people (53 per group) based on your sample size calculation and obtained

- $d = 0.1$ (95 % CI: -0.1, 0.3, p = 0.42)
- $p \geq 0.05$ = 95% CI contains 0: We did not find enough evidence against H_0
  $\rightarrow$ We retain H_0
- Same CI length as outcome I: probably δ was overestimated, more precision needed to detect a smaller δ (if there at all)

Underpowered studies

Studies not properly powered to detect a particular effect

- Using a rule of thumb
- Relying on unrealistic assumptions (e.g., too large δ)
- Not conducting a sample size calculation

1. High risk of futility
   - Exposing patients without good reason
2. High risk of inflated effect
   - Exacerbating publication bias

Outcome III: High risk of Type II error

Suppose you recruited 30 people (15 per group) and obtained

- d = 0.3 (95 % CI: -0.1, 0.7, p = 0.14)
- $p \geq 0.05$ = 95% CI does contain 0: We found not enough evidence against H_0
  $\rightarrow$ We retain H_0
- Large CI's: δ was not precisely estimated: more precision needed to detect an effect (more subjects)
- With this sample size, only 37 % power!

Outcome IV: High risk of Type I error

Suppose you recruited 30 people (15 per group) and obtained

- $d = 0.5\ (95\% CI : 0.2, 0.9, p = 0.01)$
- $p < 0.05$ = 95% CI does not contain 0: We found strong evidence against H_0
  $\rightarrow$ We reject H_0
- Relatively large CI's: δ was not precisely estimated, too few subjects
- With this sample size, d might be biased or a chance finding
- Only 3 more people felt better in Wonderdrug sample than in Outcome III

How to increase power?

By increasing sample size:

- Multi-centre trials
- Broad inclusion, few exclusion criteria
- No more than 2 treatment groups

Additional means:

- Increasing treatment difference
- Increasing risk of Type I error
- Decreasing random variation
- Recording baseline and additional values

Reporting

For the sake of reproducibility:
List all assumptions and steps, including software info

Sample size calculation for two independent proportions
Summary statement

Group sample sizes of 53 in group one and 53 in group two achieve 90% power to detect a difference of 0.3 between the group proportions. The proportion in group one (the treatment group) is assumed to be 0.6 under the null hypothesis and 0.3 under the alternative hypothesis. The proportion in group two (the control group) is 0.3. The test statistic used is the two-sided *z* test with unpooled variance. The significance level of the test was targeted at 0.05. The significance level actually achieved by this design is 0.052.

PASS 11 example output (Hintze, J., 2011)

Resources

Besides references in E-learning module:

- https://clusterrcts.shinyapps.io/rshinyapp/
- https://monash-biostat.shinyapps.io/OpenCohort/
- https://statpages.info/#Power/
- https://martonbalazskovacs.shinyapps.io/SampleSizePlanner/
- Your friendly neighborhood statistician
  - Departmental epidemiologist / statistician
  - For students: https://www.rug.nl/gmw/methodology-shop
  - For researchers: Clinical Research Office
  - Support highly recommended for complex designs