Sample size and power

Valeria Vitelli
Oslo Centre for Biostatistics and Epidemiology
Department of Biostatistics, UiO
valeria.vitelli@medisin.uio.no

MF9130E – Introductory Course in Statistics
08.05.2023

Outline

Aalen chapter 9.6, Kirkwood and Sterne chapter 35

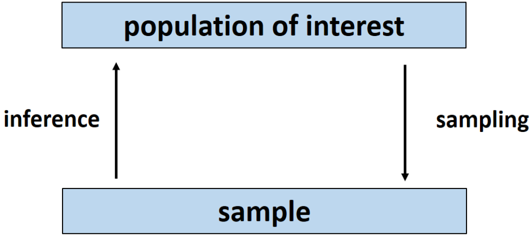
- Sample size and **random variation**
- Sample size for **precision**
- Sample size in hypothesis testing, **power**

Sample size

Planning a scientific study

- **How big your sample should be** is a crucial question when planning a study
- **Small samples** reduce the chances of getting significant findings, and thus the generalizability of the study $\rightarrow$ Non-significant results not conclusive / informative
- The **required sample size** can be calculated mathematically for different statistical methods, once we make some (reasonable) assumptions on the measured variables
  - For **basic methods**, like t-tests, chi-squared etc, we have **formulas** and simple **sample size calculators**
  - For **advanced methods** formulas become more complex, and one often do rough approximations or computer simulations
- Such calculations should be done **prior to the start of the study** (even though some referees will ask you to do them post-hoc)

Sampling and inference



- **Goal**: Make statement(s) regarding an unknown parameter value in a population, based on sample data

Two types of statistical inference

- **Confidence intervals**
  - The uncertainty of point estimates such as the mean, proportion or median
- **Hypothesis testing**
  - Assessing the strength of the evidence needed to reject the null hypothesis – the p-value

→ Two approaches to sample size calculations

- **Precision based**
  - What is the sample size required to get the confidence interval for my point estimate (e.g. mean, proportion) down to a specific width?
- **Power based**
  - What is the sample size required to detect the minimum clinically relevant difference at a given degree of certainty in a hypothesis test?

Sources of error

Two main sources:

Observed data = truth + ***systematic errors*** + ***random errors***

- **Systematic errors** (bias)
  - Faulty design; lack of randomization, blinding etc
  - Instruments not calibrated etc
- **Random errors** (chance)
  - Due to random variation in the population

The latter can be reduced by increasing the sample size!

Random errors

- The effect of random **error decreases by the square root of *n*** as the sample increases – remember from week 1 of the course:
- **Standard error of the mean**

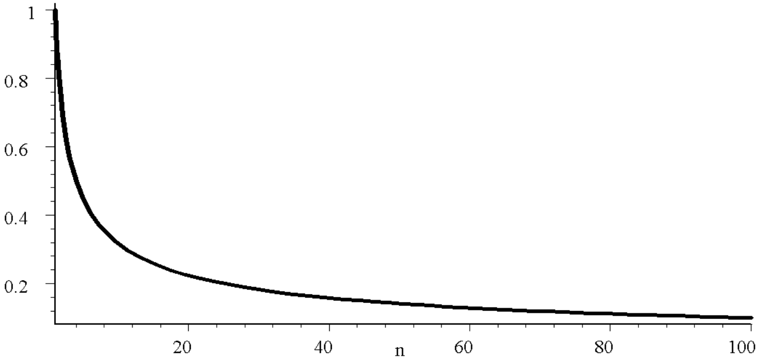
$$se = \frac{\sigma}{\sqrt{n}}$$

- **Standard error of a proportion**

$$se = \sqrt{\frac{p(1-p)}{n}}$$

Illustration: Uncertainty decreases with increasing n

- Standard error of the mean for a **variable with standard deviation $\sigma = 1$**



Input needed for sample size calculations

You **always** need to specify:

- Expected variation in the data (e.g. standard deviation)
- Significance level, e.g. 5%

When calculating a CI you need to specify:

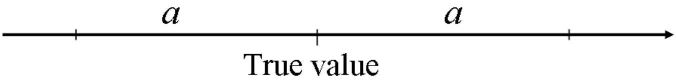
- Wanted *precision*

When doing hypothesis testing you need to specify:

- Clinically relevant difference between groups
  (OR measurements, if paired)
- Wanted *power* when computing sample size
  (or available *sample size* when computing power)

Sample size for point estimation

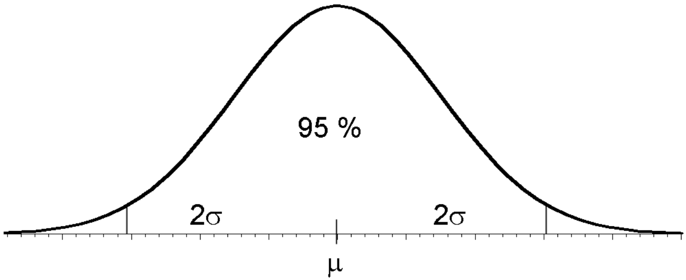
A question of precision



- Want to construct **a 95% confidence interval with a given precision** $\boldsymbol{a}$
- **Remember**: wish to be 95% certain that the estimate is within $\pm 1.96 \cdot se$ of the true value
- If the estimate is normally distributed, sample mean $\pm 1.96 \cdot se$ covers approximately 95% – this means that $a = 1.96 \cdot se$

Remember the normal distribution

- μ = expected value, σ = standard deviation
- $\mu \pm 1.96 \cdot \sigma$ covers 95% of the distribution



- We also remember that the mean itself is normally distributed with standard deviation equal to the standard error, $\sigma_{\bar{X}} = se$, so that $\boldsymbol{a \approx 2 \cdot se}$

Sample size for estimating a mean with required accuracy a

- To estimate a mean, the following number of observations are needed:

$$\boldsymbol{n = \frac{4\sigma^2}{a^2}}$$

(because $a = 2 \cdot se = 2 \cdot \frac{\sigma}{\sqrt{n}}$)

- **Example**: Number of observations required to estimate the cholesterol level (mmol/dl) in the population with a precision of 0.5. Suppose $\sigma = 1$. Number of observations required:

$$n = \frac{4 \cdot 1^2}{0.5^2} = 16$$

Sample size for estimating proportion with required accuracy a

- To estimate a proportion with a given precision, the following number of observations are needed:

$$\boldsymbol{n = \frac{4p(1-p)}{a^2}}$$

(because $a = 2 \cdot se = 2 \cdot \sqrt{\frac{p(1-p)}{n}}$)

Sample size in hypothesis testing

A question of power

- Power $= \boldsymbol{P}(\text{reject } \boldsymbol{H_0} | \boldsymbol{H_a} \text{ true})$
- **With words**: the probability of rejecting the null hypothesis if the alternative hypothesis is true
- **With other words**: the probability of finding a difference between the groups if there really *is* a difference
- More **subjects** $\rightarrow$ **power** is increasing

How large power should you have?

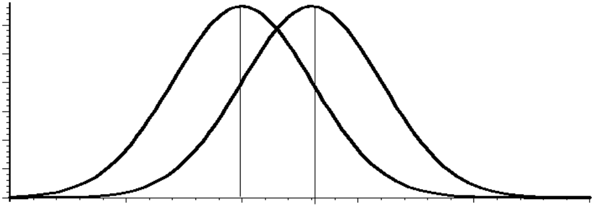
- As **large as possible**; but the question is more about what is possible with regards to sample size
- Also, even if possible, a too large sample size is costly in various respects – **sample size calculations are important**!
- Want the study sample to be so big that **an interesting effect should turn out significant** with a large probability
  - The usual minimum for a study with **"good" power** is 80% ("Industry minimum")
  - Many large **definitive studies** aim at a power of 99.9%
- If the power is large, **negative results will also be interesting** and worthy of publication

Remember the types of errors in hypothesis testing

- **Type I error**: To conclude that there is an effect when in reality there is no
  - Controlled by the significance level - probability $\boldsymbol{\alpha}$
  - Usually choose significance level $\alpha = 5\%$

- **Type II error**: Not discovering a true effect
  - Controlled by the sample size - probability $\boldsymbol{\beta}$
  - Power: 1-β (probability of discovering a true effect)
  - A minimum for 'good' power was said to usually be 80%; so 80% chance of rejecting the null hypothesis if there is an effect

Example: Testing for difference in two normally distributed variables



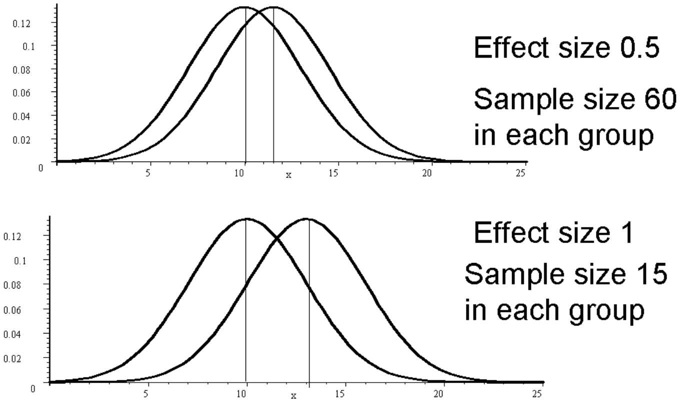
Effect size

- In sample size calculations for group comparisons you can calculate the **effect size**
- A standardized version of the **clinically relevant difference** of the phenomenon you study (effect size is sometimes also called *standardized difference*)
- Formulas for the effect size **depends on the type of analysis**
  - Two sample t-test?
  - Paired t-test?
  - Comparing proportions?

Estimating the effect size

- Effect size depends on:
  - **Clinically relevant difference** (often denoted Δ)
  - The **standard deviation** of the variable you study
- When knowledge of this variable is incomplete, **base it on existing research**:
  - Meta-analysis or reviews
  - Similar single studies that might exist
  - Perform a pilot study
- **Theoretical importance**:
  - Specifying minimum effect of interest

Different effect sizes



The smaller the effect size is, the more data one needs to sample to get high power. This is because the data distributions overlap more and more for decreasing effect sizes.

Cohen's Standard	Effect Size	Percentile Standing	Percent of Nonoverlap
	2.0	97.7	81.1%
	1.9	97.1	79.4%
	1.8	96.4	77.4%
	1.7	95.5	75.4%
	1.6	94.5	73.1%
	1.5	93.3	70.7%
	1.4	91.9	68.1%
	1.3	90	65.3%
	1.2	88	62.2%
	1.1	86	58.9%
	1.0	84	55.4%
	0.9	82	51.6%
LARGE	0.8	79	47.4%
	0.7	76	43.0%
	0.6	73	38.2%
MEDIUM	0.5	69	33.0%
	0.4	66	27.4%
	0.3	62	21.3%
SMALL	0.2	58	14.7%
	0.1	54	7.7%

Example of calculation: Comparing the means of two groups

- Clinically relevant difference: Δ
- Standard deviation in both groups: σ
- **Effect size is given by:** $\boldsymbol{\Delta/\sigma}$
- Say you compare cholesterol levels in two groups which have a clinically relevant difference $\Delta = 0.5$ and a standard deviation $\sigma = 1$:
  - Effect size: $\Delta/\sigma = 0.5$
  - A relevant difference of 0.5 means that if the average level in the two groups is for example 5.7 and 6.2, it would be important to discover
- We are going to use a **two sample t-test**
- We choose (for ex.) a **significance level** $\boldsymbol{\alpha = 0.05}$ **and power** $\boldsymbol{1 - \beta = 0.80}$

Sample size calculations **in R**: use the package pwr

```
# install package pwr
install.packages('pwr')
# load package pwr
require(pwr)
```

Two sample t-test

For the two sample t-test we use the R function `pwr.t.test`

```
pwr.t.test(n = NULL,
           d = 0.5,
           sig.level = 0.05,
           type = 'two.sample',
           alternative = 'two.sided',
           power = 0.8)
```

R output

```
Two-sample t test power calculation

              n = 63.76561
              d = 0.5
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

Conclusion in the cholesterol example

- An effect size of 0.5 gives a sample size of $n = 64$, i.e. need to measure cholesterol level in 64 patients in each group in order to get a significant effect
- This means a **total sample size of 128**
- **Remark: R** always states sample size **within-group**

Sample size for paired data

- E.g. a cross-over study
- **Effect size:** $\boldsymbol{\Delta/\sigma_d}$, where σ_d is the standard deviation of the difference between the two measurements
- New cholesterol reducing drug, what is the effect of using it for a month?
- A difference of 0.5 is important to discover, assume $\sigma_d = 1$
- Same function `pwr.t.test`, now with the argument `type = 'paired'`
- Need to sample 34 patients (next slide)

Using R

```
pwr.t.test(n = NULL,
           d = 0.5,
           sig.level = 0.05,
           type = 'paired',
           alternative = 'two.sided',
           power = 0.8)
```

Output:

```
Paired t test power calculation

              n = 33.36713
              d = 0.5
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number of *pairs*
```

Sample size for proportions

- Compare proportions in two groups
  - Initial guess on the proportions: p_1 and p_2
  - Relevant difference: $p_1 - p_2$
  - Average proportion: $\bar{p} = (p_1 + p_2)/2$
  - **Effect size:** $\frac{p_1 - p_2}{\sqrt{\bar{p} \times (1 - \bar{p})}}$
- **Example**: Compare the prevalence of depression in two populations
  - Guess on the proportions: 0.10 and 0.20
  - Average proportion: $\bar{p} = 0.15$
  - Effect size: $\frac{0.20 - 0.10}{\sqrt{0.15 \times (1 - 0.15)}} = 0.28$

Example: Compare the prevalence of depression in two populations

- Choose:
  $\alpha = 0.05$ (5%)
  1 - *beta* $= 0.80$ (80%)
- An effect size of 0.28 gives a sample size of 390, i.e. **need 195 patients in each group** to get a significant difference (see next slide)
- **Using R:**
  1. first compute the effect size with function `ES.h`
  2. then use it to compute the sample size (or power) with the function `pwr.2p.test`

Using R

```
effect.size <- ES.h(p1 = 0.1, p2 = 0.2)
pwr.2p.test(h = effect.size,
            n = NULL,
            sig.level = 0.05,
            power = 0.8,
            alternative = 'two.sided')
#-------------------------------------------------#
Difference of proportion power calculation
for binomial distribution (arcsine transformation)

              h = 0.2837941
              n = 194.9081
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: same sample sizes
```

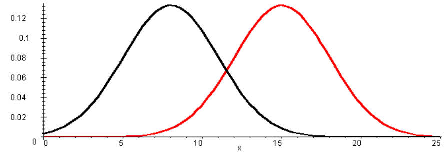
Sample size calculations the other way around: Find power given a fixed sample size

- **Example**: Want to test the effect of nicotine gum. 15% of quitting smokers are still not smoking after 6 months. If gum increases this proportion to 30%, we want a significant test
- Average proportion is 0.225, effect size is 0.36
- For financial reasons, can only afford **200 patients in total** (Remark: **R** wants sample size **per group**)
- **Find power** of 82.3%, i.e. 82.3% chance of discovering this effect with a one-sided test (see next slide)

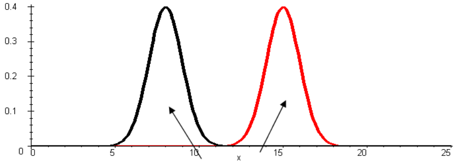
Using R

```
effect.size <- ES.h(p1 = 0.15, p2 = 0.3)
pwr.2p.test(h = effect.size,
            n = 200 / 2, # sample size PER GROUP!
            sig.level = 0.05,
            alternative = 'less')
#------------------------------------------------#
Difference of proportion power calculation for
binomial distribution (arcsine transformation)

              h = -0.3638807
              n = 100
      sig.level = 0.05
          power = 0.8233406
    alternative = less

NOTE: same sample sizes
```

The variance can be reduced by averaging over multiple measurements



- Normal distributions with expectations 8 and 15 and standard deviations 3



- Distributions of means based on samples of size 9 from each normal distribution

Becomes easier to see differences, since the distributions do not overlap as much anymore!

Different group sizes

- If you need 400 patients in total, do you need exactly 200 in each group?
- Broadly speaking: **It does not matter that much**
- If you need 400 patients in total, you can put 250 in one group and 150 in the other
- If you want an exact answer, **there are formulas** for this – use literature

Minimizing required sample size

- **Continuous measurements** instead of categories: actual measurements yields more power
- **Paired measurements**: each measurement is matched with its own control – less variance
- Allow for **unequal group sizes** – might be feasible to recruit additional individuals in one group; e.g. the control
- **Expand clinically relevant difference**: Perhaps Δ is unnecessarily small
- **Increase measurement precision**

Sample size calculations are uncertain

- **If the difference is smaller than expected, the power will decrease**
- Numbers for **clinically relevant difference and variance has a lot of impact** on the calculations – are they correct?
- If very uncertain, do **sensitivity analysis** – calculate for different scenarios

You are happy with your sample size calculation: What can still go wrong?

- Have **been too optimistic** on how fast patients are included in the study
  - Too strict inclusion criteria
  - Too time demanding
  - Drop outs

Wise to include more patients than estimated from the sample size calculations to allow for drop outs etc

Multiple or changing hypotheses

- Important to come up with **a few main hypotheses** that you wish to test
  - Choosing significance level 5% means that you will reject a null hypothesis that is true in reality 5% of the time!

- If you find that other research hypotheses are more interesting after you have collected your data, the **initial sample size calculations may be worthless**

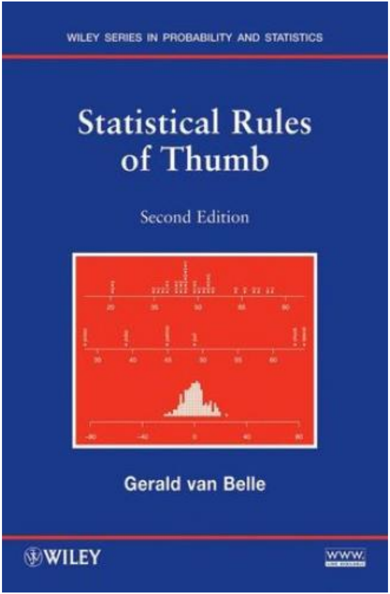
Sample size for other tests and methods

- **Non-parametric tests**:
  - Rule of thumb: Calculate for corresponding parametric test and add 15%
- **Regression analyses** with many variables:
  - Sample size calculation quickly becomes uncertain and more difficult
  - Generally **need larger samples to control for more variables**
  - Software, formulas and rules of thumb exist in different extent also for multiple regression and more advance methods
- **Be pragmatic!**

Software for sample size calculations

- **Licensed software:**
  - **STATA**: very nice interface
  - **SamplePower**: `www-03.ibm.com/software/products/en/spss-samplepower`
  - **nQuery**: `www.statsols.com/products/nquery`

- **Free programs: (other than R!)**
  - **Statpages**: `statpages.org`
  - **G*Power**: `http://www.gpower.hhu.de`


WILEY SERIES IN PROBABILITY AND STATISTICS
Statistical Rules of Thumb
Second Edition
Gerald van Belle
WILEY
WWW.

Summary

- Random and systematic error
- **Sample size** for given **precision**
- **Sample size** for **testing hypotheses** at a given power
- **Power** when testing hypotheses with a given sample size
- Use of **R** and rules of thumb