Index of Lecture 13: Sample size and power

Practical Information

Today's lecture: sample size and power for experimental studies,

- partly well-known material from VHM 801, plus:
  - two-way ANOVA with/without interaction,
  - more details on equivalence/non-inferiority studies,
- software demonstrations (Minitab and Stata) using interactive menus,
- also logistics for projects (presentations, reports).

Textbook reading:

- GO Chapter 7 and Section 10.3 (brief),[1]
- supplementary "Notes on sample size calculations" (not part of course curriculum because mostly covered in textbook).

Other news / schedule:

- next, and last, lecture (April 10): project presentations, and course wrap-up,
- course syllabus posted at course homepage,
- do you want to do course evaluations? (if yes, it will be on April 10),
- deadlines: last home assignment (April 7), project report (April 9).

[1] Discussion about non-central F-distributions and power curves may be skipped because our focus is on using software for power calculations.

Project Presentations

- scheduled for April 10, 1-4pm, in the computer lab (218S),
- approx. 20 min. overview of problem, data, statistical analysis and conclusions,
  - statistical models/methods must be explained!
  - conclusions must be presented, including estimated effects,
  - reduce biological introduction and discussion to the essentials...,
- approx. 5 minutes informal discussion, involving
  - all course participants,
  - both biological and statistical issues,
- use Word, Powerpoint and Minitab/Stata (use desktop in room or your own laptop, with HDMI connection), as you prefer,
- any priorities on order? (a randomly determined order is already posted),
- marking scheme:
  - no marks for presentation alone (only combined with report),
  - my main emphasis is on your understanding of what you did...,
  - format and layout of presentation are of minor importance.

Project Reports

- recommended to aim for manuscript-like layout:
  introduction, material and methods (in particular, statistical methods), results, discussion/conclusion,
- remember, statistical methods must be described in more detail than you would do in an applied paper,
  - you need to document your analyses by suitable software listings or program files (e.g. a Stata do-file),
  - please attach a data set prepared for analysis,
- the statistical analysis often comprises several parts/methods (contrary to statistics reported in papers that are usually restricted to a single method),
- avoid your report being essentially a pile of annotated software listings,
- listings may be put in an appendix (and could be numbered),
- probably 5-10 pages of text,
- marked (30% of course mark),
  - emphasis will be on: problem and data description, statistical models and their validation, statistical inference, conclusions and presentation of results.
- due date listed at course homepage & Moodle account.

Intro Sample Size Issues

Considerations for the size of an experimental study:

- # factors and for each factor, the number of levels → # treatments (incl. control),
- experimental design, e.g. block sizes in a block design,
- cost per experimental unit, and management of experiment (e.g., size limitations).

Plain common sense considerations for choosing a "good" sample size:

- size should be sufficient to detect (statistical significance of) treatment differences of interest,
- avoid "waste" of experimental units,[2]
- having replications will reduce the sensitivity to errors.

Textbook example (GO Sections **7.1 − 7.5**, page 150 ff.):[3]

- patients with 3 different types of neurological diseases,
- "VOR" measurements (log scale) as indicators of disease status,
- preliminary data: group means 2.82, 3.89, 3.04; within-group variance $s_p^2 = 0.075$.

[2] Involves keeping cost as low as possible, and not using individuals (animals, humans) unnecessarily − in the sense of no longer contributing any important information, because the conclusion is already clear from a smaller sample.

[3] Also (VHM 801) example from the notes: within-subject differences (e.g., computed from measurements before and after an intervention) in blood pressure with an assumed standard deviation of 10 *mm* Hg.

Statistical Methods to Choose Sample Size

Fact: all formal procedures require pre-decided statistical model involving choices of:

- targeted outcome and scale for its analysis,
- targeted parameters and/or hypotheses of interest,
- assumptions involved in model/analysis,[4]

as well as detailed prior knowledge (estimates or guesses) about the outcomes:

- size of effect of interest, or desired precision for targeted estimate,
- standard deviation of observations (for normal distribution models),

— usually, all this information is not readily available.[5]

Three general statistical approaches for determining sample size:[6] [7]

(1) from desired precision (standard error, size of 95% CI) on selected estimate,

(2) from effect of interest, using "Cox's rule" (practical rule),

(3) from desired power of test for effect of interest, using always statistical software (e.g., Minitab/Stata/R), web applications or simulation (some references in the notes), avoiding hand calculations[8].

[4] Most standard sample size calculations for quantitative outcomes are based on normal distribution assumptions.

[5] Some possible sources of information: *i)* published articles on same research question, *ii)* a pilot study conducted prior to the main study, and *iii)* expert opinion solicited from subject-matter experts.

[6] Approach (3) is far more common in practice, arguably too common; Bland (2009), *BMJ* 339, 1133-1135.

[7] Additionally, specific methods exist for many specialized settings and study types.

[8] Simple plug-in formulas from IPS and VER, and also Lehr's and Slovin's formula, are not recommended.

Sample Size based on Estimation Precision

General approach for normal distribution models:

- assume estimated/guessed/known standard deviation σ,
- assume (mean) parameter of interest ψ and estimate $\hat{\psi}$, with standard error $\text{SE}(\hat{\psi}) = \sigma \times c(n)$, where $c(n)$ is a known constant depending on the number of obs. (n),
- approximate 95% CI[9]: $\hat{\psi} \pm 2\,\sigma\,c(n)$,

Compute n to achieve desired margin of error (M) by solving for n in the equation:[10]

$$M(\text{desired value}) \geq 2\,\sigma\,c(n).$$

Blood pressure example: sample size for $M = 3$ mm Hg, see VHM 801: 12L−10,

VOR example: desired CI of width 0.5 ($\Rightarrow M = 0.25$) for mean differences between (any) two groups,

- model: 3 independent samples of size n from normal distributions $\text{N}(\mu_i, \sigma^2)$; use $\sigma = s_p = \sqrt{0.075} = 0.2739$,
- estimation: $\psi = \mu_1 - \mu_2$, $\hat{\psi} = \bar{y}_{1\cdot} - \bar{y}_{2\cdot}$, $c(n) = \sqrt{2/n}$,
- solve: $0.25 \geq 2 \times 0.2739\sqrt{2/n} \Rightarrow n \geq 2(2 \times 0.2739/0.25)^2 = 9.6$, or $n = 10$; rerun with 2 replaced by $t(.975, 27) = 2.052$; $n \geq 10.1$; choose $n = 11$ patients per group.

[9] Approximation requires either σ known, or σ unknown and n so large that $t^* = t_{.975}(\text{df}) \approx z^* = z_{.975} \approx 2$, say $n \geq 40$.
[10] Some software (e.g. Stata) allows to also include variability in the estimate s for σ, further complicating the procedure.

Errors of Type I–II and Power (Recap)

Errors of type I and II:

- type I error: to reject H_0, when H_0 in reality true,[11]
- type II error: to not reject H_0, when H_0 in reality false,
- power of statistical tests involves type II errors (below),
- schematic:

Conclusion	Truth about population	
from sample	H_0 true	H_a true (H_0 false)
reject H_0	type I error	no error
not reject H_0	no error	type II error

Power of a statistical test:

- involves a specific alternative, e.g. in the blood pressure example a difference of 3 mm Hg (i.e., $H_a: \mu_D = 3$),
- definition: power = probability that the statistical test will reject H_0, when the specific alternative H_a is true = 1− type II error,
- important for planning of experiments: what chance of a significant result?
- difficult to calculate in complex models (lots of formulas and software exist).

[11] The definition of statistical tests involves only type I errors, which are controlled by the significance level (α).

Sample Size based on Power or Effect Size

Cox's rule for "informal" sample size determination:[12]

- assume "effect size" $\delta = |m_0 - m_a|$, where m_0 and m_a are (true, population) values of ψ, the parameter of interest, under H_0 and H_a, respectively,
- rule: choose n by solving: $|m_0 - m_a| = 3\,\sigma\,c(n)$,
- note: does not involve power or significance level.

Requirements for sample size determination based on power:[13]

- statistical model / design, a hypothesis to test and the desired "effect size" δ,
- standard deviation of model (for normal distribution data/models),
- desired value of power (0.8, or 80%, commonly used),
- significance level (α) and direction (one/two-sided) of test used (usually $\alpha = 0.05$).

Blood pressure example: "effect size" (true difference) of 3 mm Hg,

- Cox's rule: $3 = 3 \cdot 10\sqrt{1/n} \Rightarrow n = 100$,
- power of 0.8 (at $\alpha = 0.05$, and with two-sided H_a): $n = 90$ (Minitab/Stata).

VOR example: power for $n = 4$ and $\alpha = 0.01$ with means from preliminary data:

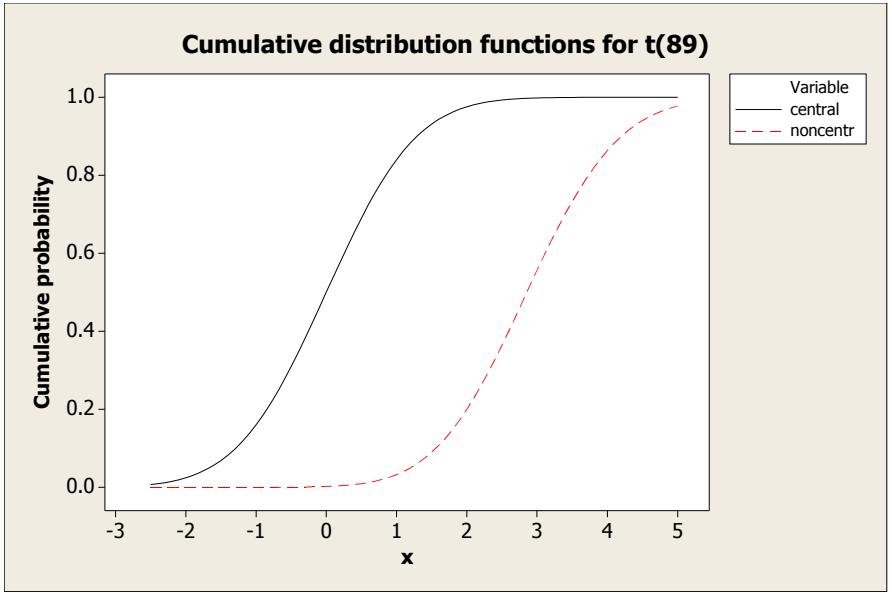
- Textbook / Stata menu: 0.930 (all means), Minitab (maximal difference): 0.896.

[12] Discussed in Christensen (1996): *Analysis of Variance, Design, and Regression.*
[13] Post-hoc power calculation (after study has been carried out) is controversial and *not* recommended, see 13L−13.

Non-Central Distributions

Several of the common reference distributions for tests ($t-$, χ^2-, $F-$ distributions) have a non-centrality parameter (ζ):

- $\zeta = 0 \sim$ usual reference distribbution under null hypothesis H_0,
- $\zeta \neq 0 \sim$ distribution of test statistic under specific alternative hypothesis H_a, where ζ reflects the magnitude of deviation from H_0 (or effect size),
- power is then computed from tail areas in distributions with $\zeta \neq 0$,
- example: non-central F-distribution in ANOVA table
  – see GO Figure 7.1, where $\zeta = \sum_i n_i \alpha_i^2 / \sigma^2$,
- blood pressure example:
  non-central t-distribution:
  - $t = \hat{\mu}/\mathrm{SE}(\hat{\mu}) \sim$ non-central t,
    where $\hat{\mu} \sim \mathrm{N}(\delta, \sigma^2_{\hat{\mu}})$ and $\zeta = \delta/\sigma_{\hat{\mu}}$,
  - with $n = 90$, we get:
    $\zeta = 3/(10/\sqrt{90}) = 2.846$
    and $t^* = t_{.975,89} = 1.987$.



Sample Size for Two-way ANOVA

Example 13.5−6 in IPS (5th/6th/7th ed.): calcium supplementation as a treatment for Osteoporosis in elderly people,

- factors: calcium (placebo, 800 mg/day), vitamin D (placebo, 300 IU/day)[14],
- outcome: y = change in bone mineral density (BMD),
- assume: n subjects per calcium × vitamin D group[15], calcium effect size of interest: $\delta = 5$ BMD units, and within-group standard deviation: $\sigma = 10$ units.

Statistical model (anticipated): $y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$,
where α_i, β_j are main effects and $(\alpha\beta)_{ij}$ the interaction, and $\varepsilon_{ijk} \sim \mathrm{N}(0, \sigma^2)$.

Two possible scenarios:

- no interaction: sample size based on main effects,
- interaction, and different approaches for sample size:
  - based on contrast of interest for one factor within one level of other factor (≈ two-sample situation),
  - based on F-test for interaction,
  - based on contrast within interaction.[16]

[14] Vitamin D is needed for the body to efficiently utilize calcium.
[15] In the notes, the number of subjects per group is denoted by c.
[16] For a 2×2 factorial such a contrast is equivalent to the full interaction.

Calculations for 2-way ANOVA Example

No interaction model — sample size for calcium effect (δ):

- two-sample calculation (power 0.80) gives sample size 64 $\Rightarrow$ actual $n = 64/2 = 32$, because the two vitamin D groups both contribute to the sample size,[17]
- using Cox's rule: $\text{SE} = \sigma\sqrt{1/(2n)+1/(2n)} = \delta/3 \Rightarrow n = (3\sigma/\delta)^2 = 36$.

Interaction model − two approaches:

- sample size for calcium effect in high vitamin D group: above two-sample calculation applies[17]: $n = 64$ (same for low vitamin D group),
- sample size for interaction of size δ:
  - interaction contrast[18] and its SE:

$$\begin{aligned} \hat{\gamma} &= \bar{y}_{11\cdot} + \bar{y}_{22\cdot} - \bar{y}_{12\cdot} - \bar{y}_{21\cdot}, \\ \text{SE}(\hat{\gamma}) &= \sigma\sqrt{1/n + 1/n + 1/n + 1/n} = \\ &= \sigma\sqrt{4/n} = (\sqrt{2}\sigma)\sqrt{2/n}, \end{aligned}$$

  - last formula $\sim$ two-sample comparison with standard deviation $\sqrt{2}\cdot\sigma$, from which we get $n = 127$ (power 0.80).[17]

[17] Actual power will not be exactly 0.80 because of different df in the two-way ANOVA and the two-sample situation; alternatively, the calculation can be done directly for a two-way ANOVA (Stata).

[18] The contrast estimates the difference in calcium effect between the two vitamin D levels, or the difference in vitamin D effect between the two calcium levels.

Unequal Sample Sizes

Generally speaking, it is most efficient (gives best precision) to have equal sample sizes in different levels (or categories) of a factor.

Two special cases where unequal distributions across levels are useful:

- one control group and several treatment groups all to be compared to control only(!): the optimal sample size for control is then larger than for treatment groups,
  - g treatments (incl. control), $n_c =$ size of control, $n_t =$ size for other treatments $\Rightarrow$ solve equation: $(n_c/n_t)^2 = (g-1)$,
  - for example, for $g=5$ we get $n_c = 2n_t$ (control group of double size!),
- factor with quantitative (and equidistant) levels where certain polynomial contrasts are of primary interest:
  - for example, linear contrasts always give highest weights to the most extreme categories (Table D.6 in GO) $\Rightarrow$ higher precision for a linear contrast may be achieved by overrepresenting the most extreme categories.[19]

[19] The gain in precision is counterbalanced by reduced precision for other contrasts; in the most extreme case of only including the two extreme categories, the linear contrast is estimated with best precision, but no other polynomial contrasts can be estimated.

Sample Size Misconceptions (Recap)

Common misconceptions[20] in sample size calculations:

- use of standard effect sizes (general definitions of "small", "medium" and "large" effects, relative to standard deviation): effects of interest should be determined exclusively from the context of your study,
- retrospective power calculation: after a study has been carried out and using its estimated values:
  - power/sample size calculations aid in planning of new studies, not in interpreting results of data analysis,
  - confidence intervals give the best information about the unknown parameters from a study,
  - if H_0 was not rejected, the conclusion may sometimes be strengthened by an equivalence test (instead of arguing from the study's power), see next slides.

Valid reasons to do statistical power calculations:

- planning a new study, or exploring a study's feasibility (given logistical and financial constraints),
- revising a pre-study sample calculation after the study has been carried out, if:
  - some of the assumptions turned out to be unrealistic, e.g. for σ (but typically not effect size),
  - insufficient information was available for a particular analysis.[21]

[20] Largely based on Lenth (2001), *The American Statistician* 55, 187−193; also discussed in Greenland et al. (2016).
[21] Thereby necessitating the sample calculation to have been based on other parameters.

Equivalence and Non-Inferiority Testing in Two-Sample Situations

An equivalence test is for making a (statistical) statement that effects of two treatments (or other groups) are "equivalent", in the sense that their effects differ at most by some pre-set amount, typically small.

A non-inferiority test is for making a (statistical) statement that the effect of one treatment is "not inferior" to that of another treatment, in the sense that its effect is either better or possibly not worse by more than some pre-set amount, typically small.

Comparison with the typical situation comparing treatment and control groups (say, with means μ_1 and μ_2),

- still relevant to test $H_0 : \mu_1 = \mu_2$ against a one- or two-sided H_a, but the desired outcome is to not reject H_0 – a weak, non-quantitative and pretty useless conclusion,
- set up different statistical hypotheses, so that the desired outcome is to reject H_0,
- typically, the new H_0 involves a range of parameter values (instead of a single value),
- typically, the new H_0 and H_a require extra information from the study's context: the magnitude of a "true" difference between groups that is "acceptable".

Example to illustrate ideas (from Minitab, possibly made-up data):
protein content in cat food, comparing new ("Discount") and "Original" products, in two-sample study with 10 and 9 observations and equivalence defined by a difference not exceeding 0.5 grams (per 100 grams).

Methods for Equivalence Analyses

Main message: no new models needed, and typically a minor adjustment of standard methods is sufficient.[22]

General (approximate) method for equivalence testing of a difference parameter θ (e.g., $\theta = \mu_1 - \mu_2$ for a two-sample situation) up to an "importance threshold" $\delta > 0$: test the non-equivalence hypothesis $H_0^{(ne)} : |\theta| \geq \delta$ against the $H_a^{(ne)} : |\theta| < \delta$, as follows (at a 5% significance level):

* compute a 90% CI (not 95% CI) for θ,
* reject $H_0^{(ne)}$ (and favour $H_a^{(ne)}$), if the interval $(-\delta, \delta)$ entirely includes the CI.

General method for non-inferiority test of $H_0^{(ni)} : \theta \geq \delta$ against the $H_a^{(ni)} : \theta < \delta$ ($\sim \mu_2 > \mu_1 - \delta$) — use same procedure as for testing the standard $H_0 : \theta = \delta$ against the one-sided alternative.[23]

Equivalence analysis in practice:

- Minitab 18+ has built-in equivalence trial menu covering some standard designs (1-sample, 2-sample, cross-over), and additional menu for sample size calculation.
- Stata options less transparent, in `pk` (Pharmacokinetic) module, and also in add-on package `tost`.

[22] Reference text: Chow & Liu (2009), *Design and Analysis of Bioavailability and Bioequivalence Studies*. CRC Press.

[23] If $H_0^{(ni)}$ is rejected, we can claim that $\theta < \delta$, or $\mu_2 > \mu_1 - \delta$, so that group 2 is not inferior to group 1.