Index of Lecture 13

Practical Information

Today's lecture:

- power and sample size for *experimental studies* – partly well-known material from VHM 801, plus:
  - two-way ANOVA with/without interaction,
  - material on equivalence/non-inferiority studies,
  - software demonstrations (Minitab and Stata) using interactive menus (plus extra do-file),
- revisiting complex experimental design[1] (L10–6/9).

Reading:

- GO Chapter 7[2] and Section 10.3 (brief),
- suppl. "Notes on sample size calculations" (not part of course curriculum because mostly covered in textbook).

Schedule:

- next and last lecture (April 11): project presentations, and *course evaluation*,
- deadlines: last home assignment (April 8), project report (April 12),
- course syllabus posted at course homepage.

[1] Hasse diagrams not in course curriculum, see GO p. 296 for a summary.

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

Intro Sample Size Issues

Factors affecting the size of an experimental study:

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

Statistical considerations:

- size should be sufficient to detect (obtain statistical significance for) treatment differences of interest,
- avoid "waste" of experimental units,
- reduce sensitivity to errors (by taking replications).

Textbook example (GO 7.1-5, p. 150 ff.):

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

Additional notes example:

- within-subject differences[3] in blood pressure with an assumed standard deviation of 10 mm Hg.

[3] E.g., differences computed from measurements before and after an intervention.

Statistical Methods to Choose Sample Size

Fact: all procedures require pre-decided statistical model and detailed prior knowledge (estimates or guesses) about the outcomes:

- size of effects or precision of interest,
- standard deviation of observations (for normal data[4]).

Overview of approaches for determining sample size:

- from desired precision (standard error, size of 95% CI) on selected estimate (treatment mean, contrast),
- from effect of interest, using "Cox's rule" (practical rule),
- from desired power of test for effect of interest, preferably using statistical software:
  - Minitab/Stata (or websites) for basic designs,
  - SAS version 9 for a range of complex designs,
  - specialised software for special/advanced designs[5]:
    - (active researcher): `http://www.stat.uiowa.edu/~rlenth/Power`
    - (stats webpages): `http://statpages.org/#Power`
    - (G*Power = free power calculation software): `http://www.gpower.hhu.de/`

  or using *simulation* (some refs in the notes).

[4] More generally, for data with the variance estimated independently of the mean.

[5] Check also Stata's comprehensive *Power and Sample Size Reference Manual*, and the UCLA webpage on statistical analysis in different software programs on `https://stats.idre.ucla.edu/other/dae/`.

Sample Size based on Estimation Precision

Normal distribution models (without random effects):

- assume error std.dev. σ and estimate/guess of value,
- assume parameter of interest Par and estimate Est, with standard error $\text{SE}(Est) = \sigma A$, where $A = A(n)$ is a known constant (depending on number of obs. n),
- approximate[6] 95% CI: $Est \pm 2\,\sigma A(n)$.

Compute n to achieve desired margin of error (or CI length) by solving with respect to n in the equation:

$$\text{desired value} \;\geq 2\,\sigma A(n).$$

Blood pressure example: desired margin of error: 3 mm Hg,

- model: n i.i.d. observations (differences!) from $\text{N}(\mu,\sigma^2)$,
- Par = population mean, Est = sample mean, $A = 1/\sqrt{n}$,
- solve: $3 \geq 2 \times 10/\sqrt{n} \Rightarrow n \geq (2 \times 10/3)^2 = 44.4 \approx 45$.

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

- model: 3 independent samples of size n from normal distributions $\text{N}(\mu_i,\sigma^2)$; use $\sigma = s_p = \sqrt{0.075} = 0.2739$,
- $Par = \mu_1 - \mu_2$, $Est = \bar{y}_{1\cdot} - \bar{y}_{2\cdot}$, $A = \sqrt{2/n}$,
- 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.

[6] Valid for σ unknown and df large (say ≥ 40), so that $t_{.975}(\text{df}) \approx z_{.975} \approx 2$.

Errors of Type I–II and Power

Errors of type I and II:

- type I error: to reject H_0, when H_0 in reality true,
- type II error: to not reject H_0, when H_0 in reality false,
- definition of statistical tests involves only type I errors, which are controlled by the significance level (α),
- power of statistical tests involves type II errors (below),
- overview:

Conclusion from sample	Truth about population	
	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., $H_0: \mu_D = 3$),
- 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 all models (use software!),
- typical values for planning of a study: 0.8, 0.9, 0.95.

Sample Size based on Power or Effect Size

Cox's rule for "informal" sample size determination:[7]

- assume effect size $|m_0 - m_a|$, where m_0 and m_a are (true, population) values of Par under H_0 and H_a, respect.,
- rule: choose n by solving: $|m_0 - m_a| = 3\,\sigma A(n)$,
- note: does not involve power or significance level.

Requirements for sample size calc. based on power:[8]

- statistical model / design and corresponding software,
- size of effect desired to be detected,
- standard deviation of model (normal data),
- desired value of power (0.8, or 80%, commonly used),
- signif. level and direction (one/two-sided) of test used.

Blood pressure example: true difference of 3 mm Hg,

- Cox's rule: $3 = 3 \cdot 10\sqrt{1/n} \Rightarrow n = 100$,
- Minitab / Stata, power 0.8 & sign. level 0.05: $n = 90$.

VOR example: power for $n = 4$ and sign. level 0.01 with means as observed in preliminary data:

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

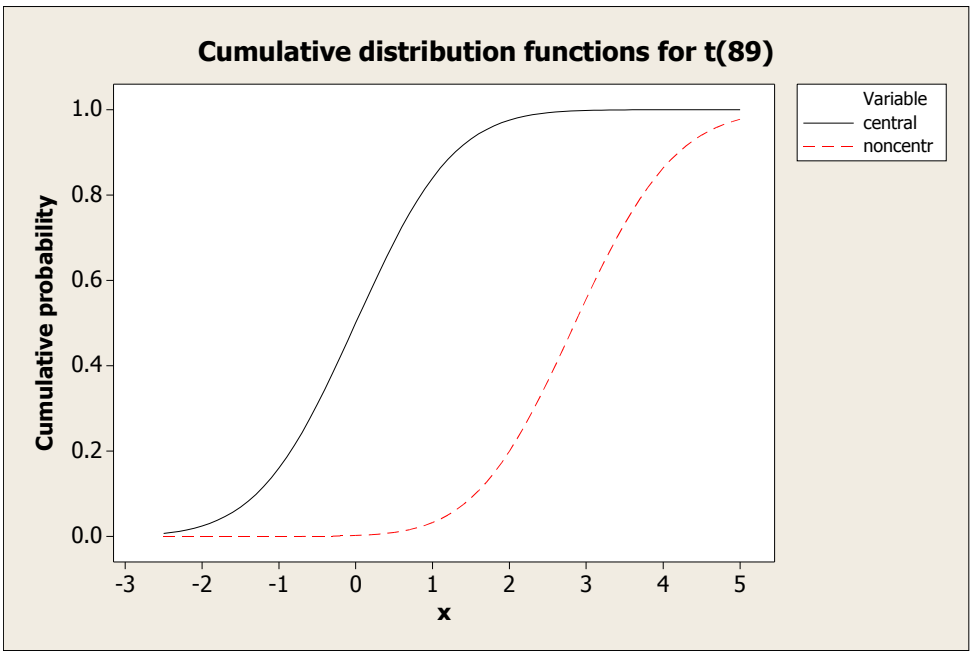
[7] Discussed in Christensen (1996): *Analysis of Variance, Design, and Regression.*

[8] Post-hoc power calculation (after study has been carried out) is controversial and *not* recommended, see page L13–11.

Non-Central Distributions

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

- $\zeta = 0 \sim$ usual ref. distrib. under null hypothesis H_0,
- $\zeta \neq 0 \sim$ distrib. of test statistic under specific alternative hypothesis H_a, where $\zeta \sim$ deviation from H_0,
- power is computed from tail areas in distrib. 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,

- factor: calcium (placebo, 800 mg/day),
- factor: vitamin D (placebo, 300 IU/day)[9],
- outcome: change in bone mineral density (BMD),
- assume n subjects per calcium $\times$ vitamin D group,[10]
- assume calcium effect size of interest: $\delta = 5$ BMD units,
- assume within-group standard deviation, *for the change in BMD*: $\sigma = 10$ units.

Statistical model (anticipated):

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

where α_i, β_j, $(\alpha\beta)_{ij}$ are the main effects and interaction, respectively, and $\varepsilon_{ijk} \sim \mathrm{N}(0, \sigma^2)$.

Two possible situations:

- 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 ($\approx$ two-sample situation),
  - based on F-test for interaction,
  - based on contrast within interaction (for 2×2 factorial: equivalent to full interaction).

[9] Vitamin D is needed for the body to efficiently utilize calcium.
[10] In the notes, the number of subjects per group is denoted by c.

Calculations for 2-way ANOVA Example

Sample size for calcium effect (δ) in no interaction model:

- two-sample calc. (power 0.80, Minitab) gives $n = 64$
  $\Rightarrow$ actual $n = 64/2 = 32$, because the two vitamin D groups both contribute to the sample size,[11]
- 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 (or low) vitamin D group: above two-sample calc. applies[11]: $n = 64$,
- sample size for interaction of size δ:
  - interaction contrast[12] and 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 $\sim$ 2-sample comparison with standard deviation $\sqrt{2} \cdot \sigma$:
    $\Rightarrow n = 127$ (power 0.80, Minitab).

[11] Actual power will differ slightly from 0.80 because the df is smaller/larger in the two-way ANOVA than the two-sample situations.

[12] 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 groups of a factor.

Two special cases where unequal distributions across groups are useful:

- One control group and several treatment groups all to be compared to control only(!): the optimal sample size for control is larger than for treatment groups,
  - g treatments (incl. control), n_c=size of control, n_t= size for other treatments $\Rightarrow$
    solve equation: $(n_c/n_t)^2 = (g-1)$,
  - e.g., 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 particular interest:
  - e.g., linear contrasts always give highest weights to the most extreme categories (Table D.6 in GO) $\Rightarrow$ higher precision for linear contrast may be achieved by overrepresenting the most extreme categories.[13]

[13] The gain in precision is counterbalanced by reduced precision for other contrasts; for example, if only the two extreme categories are included, the linear contrast is estimated with best precision, but no other polynomial contrasts can be estimated.

Sample Size Misconceptions

Common misconceptions[14] in sample size calculations:

- use of standard effect sizes (general definitions of "small", "medium" and "large" effects, relative to std. dev.[15]): 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[15]:
  - 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 instead be strengthened by an equivalence test (next slides).

Valid reasons to do statistical power calculations:

- planning a new study, or getting a sense of a new 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/settings turned out to be unrealistic, e.g. for σ (but typically not effect size),
  - insufficient info existed for an analysis of interest.

[14] Largely based on Lenth (2001), *The American Statistician* **55**, 187−193.

[15] Also the Stata manual does this, with literature references; this does not make the approach more valid...

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, typically small amount.

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, typically small amount.

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
  $\Rightarrow$ weak, non-quantitative and essentially useless conclusion,
- need to set up different statistical hypotheses, so that the desired outcome corresponds to *rejection* of the null hypothesis,
- typically, the new H_0 will involve a range of parameter values (instead of the usual single value), as will H_a,
- typically, the new H_0 and H_a will require extra information from the study's context: the magnitude of a "true" difference between groups that is not considered important.

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 obs. 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.[16]

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 $\mathrm{H_0}^{(\mathrm{ne})} : |\theta| \geq \delta$ against the $\mathrm{H_a}^{(\mathrm{ne})} : |\theta| < \delta$, as follows (at a 5% significance level):

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

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

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`.

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

Project Presentations

- scheduled for April 11, 9:00-10:40am,
- approx. 15 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 min. informal discussion, involving
  - all course participants,
    - is it a good idea to assign "opponents" to each presentation?
  - both biological and statistical issues,
- use whiteboard, overhead, Powerpoint and Minitab/Stata (use my laptop or bring your own), as you like,
- any priorities on order? (otherwise random),
- 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 are of minor importance.

Project Reports

- 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),
- *not* a pile of annotated Minitab/Stata 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.