サンプルサイズの問題
Sample size issue

Minato Nakazawa, Ph.D.
<minato-nakazawa@people.kobe-u.ac.jp>

24 April 2019

※この資料は英語なので，日本語がわかる方は，pdf のテキストの第４章を読みながら話を聞いてください

Referenes, web sites

- Reference in English
  - "Chapter 14. Sample size issues." in Machin D, Campbell MJ, Walters SJ (2007) *Medical Statistics, 4th ed.*, Wiley, pp. 261-275.
  - "Chapter 4. Comparing groups with p values: Reporting hyothesis tests." in Lang TA, Secic M (2006) How to report statistics in medicine: Annotated guidelines for authors, editors, and reviewers. 2nd ed., American College of Physicians., pp.45-60.
  - "Chapter 1. Research design" in Peacock JL, Peacock PJ (2011) Oxford handbook of medical statistics. Oxford Univ. Press, pp.1-73 (especially 56-73).
- Textbook in Japanese
  - 永田靖 (2003) サンプルサイズの決め方．朝倉書店
  - 新谷歩 (2011) 今日から使える医療統計学講座【Lesson 3】サンプルサイズとパワー計算．週刊医学界新聞，2937 号 http://www.igaku-shoin.co.jp/paperDetail.do?id=PA02937_06

Fictious example

- Due to the limited number of patients within the study period as graduate student, a student could obtain data from only 10 patients (with 10-30 healthy controls), to test the null hypothesis X, that is a factor Y has no relation with a result Z (usually suffering from that disease).
- And, as the result of statistical test, the null hypothesis X was not rejected at a significance level of 5%.
- Such non-significant results should be submitted to the journal, entitled as "Lack of association …", to avoid publication bias.
- However, the reviewer of the journal may criticize the non-significance due to too low statistical power by insufficient sample size, then the paper will be rejected.
- The graduate student may ask help for statistician, but it's too late (at best, tell the way of excuse).
- Such tragedy is frequently seen in the graduate students.

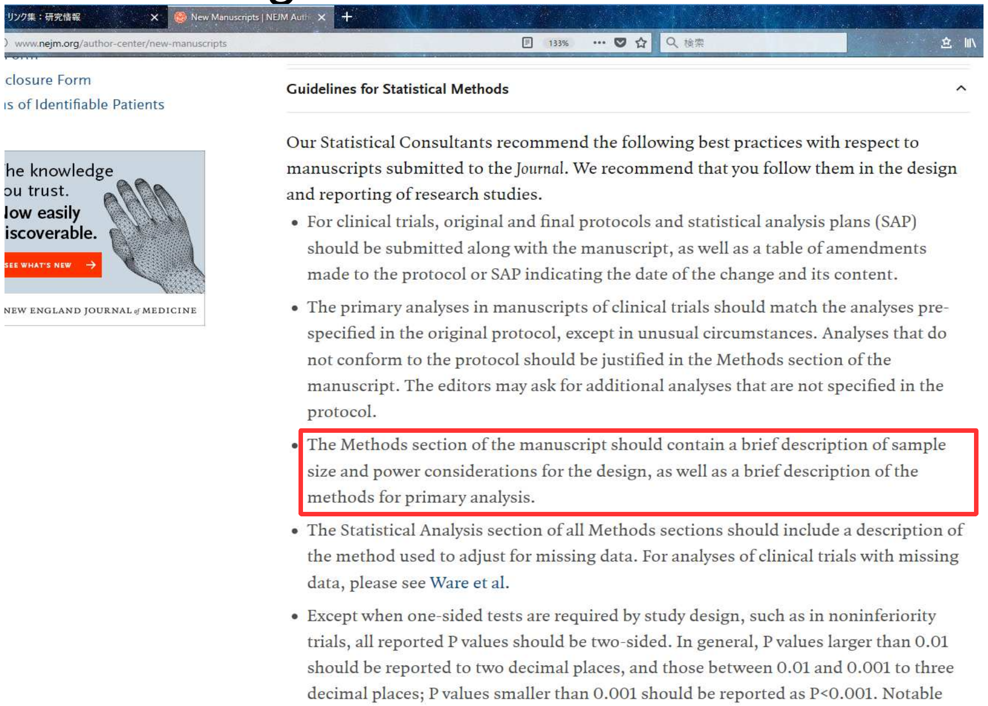
How should this student do?

- This study was conducted as hypothesis testing.
- It's well known that the large sample increases statistical power, before conducting study, determining the enough sample size is possible.
- After getting the data, only possible way is writing excuse.
  - Too rare disease.
  - Limitation of funding.
  - As a habit in the specific research theme, small sample size has been commonly accepted
  - etc.
- If the student is lucky and the study theme is quite important, the paper may be accepted as research note or short report, instead of original article.

In the textbook of medical statistics...

- A study that is too small may be unethical, since it is not powerful enough to demonstrate a worthwhile correlation or difference.
- Similarly, a study that is too large may also be unethical since one may be giving people a treatment that could already have been proven to be inferior.
- Many journals now have checklists that include a question on whether the process of determining sample size is included in the method section (and to be reassured that it was carried out before the study and not in retrospect).
- The statistical guidelines for the *British Medical Journal* in Altman et al. (2000) state that: **`Authors should include information on … the number of subjects studied and why that number of subjects was used.'**

New England Journal of Medicine



リンク集：研究情報
New Manuscripts | NEJM Auth
www.nejm.org/author-center/new-manuscripts
133%
検索
closure Form
ıs of Identifiable Patients
he knowledge
ou trust.
low easily
iscoverable.
SEE WHAT'S NEW
NEW ENGLAND JOURNAL of MEDICINE

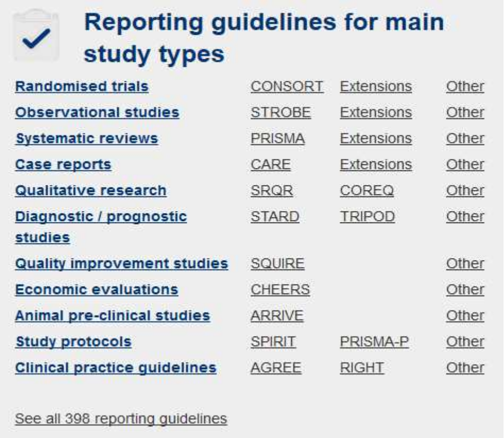
Guidelines for Statistical Methods

Our Statistical Consultants recommend the following best practices with respect to manuscripts submitted to the *Journal*. We recommend that you follow them in the design and reporting of research studies.

- For clinical trials, original and final protocols and statistical analysis plans (SAP) should be submitted along with the manuscript, as well as a table of amendments made to the protocol or SAP indicating the date of the change and its content.
- The primary analyses in manuscripts of clinical trials should match the analyses pre-specified in the original protocol, except in unusual circumstances. Analyses that do not conform to the protocol should be justified in the Methods section of the manuscript. The editors may ask for additional analyses that are not specified in the protocol.
- The Methods section of the manuscript should contain a brief description of sample size and power considerations for the design, as well as a brief description of the methods for primary analysis.
- The Statistical Analysis section of all Methods sections should include a description of the method used to adjust for missing data. For analyses of clinical trials with missing data, please see Ware et al.
- Except when one-sided tests are required by study design, such as in noninferiority trials, all reported P values should be two-sided. In general, P values larger than 0.01 should be reported to two decimal places, and those between 0.01 and 0.001 to three decimal places; P values smaller than 0.001 should be reported as P<0.001. Notable

In commonly adopted guidelines...

- STROBE Statement for Observational Study
  https://strobe-statement.org/fileadmin/Strobe/uploads/checklists/STROBE_checklist_v4_combined.pdf
  "10 Explain how the study size arrive at"
- CONSORT Statement for Clinical Trials
  http://www.consort-statement.org/
  "7a How sample size was determined"
  "7b When applicable, explanation of any interim analyses and stopping guidelines"
- See, EQUATOR
  http://www.equator-network.org/
  which gives summary information of many guidelines



Reporting guidelines for main study types

Randomised trials	CONSORT	Extensions	Other
Observational studies	STROBE	Extensions	Other
Systematic reviews	PRISMA	Extensions	Other
Case reports	CARE	Extensions	Other
Qualitative research	SRQR	COREQ	Other
Diagnostic / prognostic studies	STARD	TRIPOD	Other
Quality improvement studies	SQUIRE		Other
Economic evaluations	CHEERS		Other
Animal pre-clinical studies	ARRIVE		Other
Study protocols	SPIRIT	PRISMA-P	Other
Clinical practice guidelines	AGREE	RIGHT	Other

See all 398 reporting guidelines

Excuses for no calculation of sample sizes

- *A cynic* once said that sample size calculations are a guess masquerading as mathematics. To perform such a calculation we often need information on factors such as the standard deviation of the outcome which may not be available. Moreover the calculations are quite sensitive to some of these assumptions.
- Any study, whatever the size, contributes information, and therefore could be worthwhile and several small studies, pooled together in a meta-analysis are more generalizable than one big study.
- Often, the size of studies is determined by practicalities, such as the number of available patients, resources, time and the level of finance available.
- Studies, including clinical trials, often have several outcomes, such as benefit and adverse events, each of which will require a different sample size.

Where the sample size calculation is unnecessary

- Qualitative studies / Case report
- Small survey / pilot study
  - In descriptive study, usually previous information about the measures is unavailable, so that the sample size calculation is impossible.
  - Rules of thumb: at least 12 individuals in each group
    - List the main cross tabulations that will be needed to ensure that total numbers will give adequate numbers in the individual tables cells.

In exploratory studies ...

- Two kinds of study
  - Testing the null-hypothesis always requires the sample size calculation before the study (already explained).
  - Exploring the hidden hypothesis or describing estimates with 95% confidence intervals may not always require the sample size calculation, but power analysis (to evaluate sampling adequacy) after the study is possible.
- In the exploratory or descriptive studies
  - Prevalence estimates from small samples will be imprecise and may be misleading. For example, when we wish to get the prevalence of a condition for which studies in other settings have reported a prevalence of 10%. A small sample of, say, 20 people, would be insufficient to produce a reliable estimate since only 2 would be expected to have the condition and ±1 would change the estimate by 5%.
  - Sample size calculation determine the number of subjects needed to give a sufficiently narrow confidence intervals.

Example of exploratory study

- Values are obtained from previous studies in advance.
  - When we would like to estimate a mean, the following 3 values are needed.
    - The standard deviation (SD) of the measure being estimated
    - The desired width of the confidence interval (d)
    - The confidence level (usually 90, 95, or 99 %; 1-alpha)
  - Necessary number of samples (n) is obtained by:
n = qnorm(1-alpha/2)^2 * 4 * SD^2 / d^2
  - (e.g.) Suppose we wish to estimate mean systolic blood pressure in a patient group with a 10mmHg-wide (or 5mmHg-wide) 95% confidence interval. Previous work suggested using a standard deviation of 11.4.
    n = 1.96^2 * 4 * 11.4^2 / 10^2 = 19.97... = 20
    n = 1.96^2 * 4 * 11.4^2 / 5^2 = 79.88... = 80
  - Doubling the precision needs quadrupling the sample size.
- Estimating proportions will be given in the next slide.

(cont'd)

- Required information from previous studies and study purpose to estimate proportion
  - Expected population proportion (p)
  - Desired width of confidence interval (d)
  - Confidence level (1-alpha)
- Approximate equation to estimate the number of subjects needed [qnorm(1-alpha/2) means $z_{1-\alpha/2}$; ^2 = squared]:
  n = qnorm(1-alpha/2)^2*4*p*(1-p) / d^2
- (e.g.) Suppose we wish to estimate the prevalence of asthma in an adult population with the width of the 95% confidence interval 0.10, an accuaray of ±0.05. An estimate of the population prevalence of asthma is 10%.
  - p = 0.10, d = 0.10, alpha = 0.05
  - n = qnorm(0.975)^2*4*0.1*0.9 / 0.1^2 = 1.96^2*36 = 138

Principles of hypothesis testing

- What kind of information is needed?
  - Method of statistical test (including null-hypothesis)
  - Type I error (alpha error: probability to reject the true null-hypothesis, in other words, false positive)
  - Type II error (beta error: probability to fail to reject the false null-hypothesis / false negative) = 1 – statistical power
  - Expected values from previous studies
  - Minimum differences of clinical importance
- Equations are quite different by statistical tests (and by textbooks, because all of those are of approximation)
  - Compare means by t-test: $n= 2*(z_{\alpha}-z_{1-\beta})^2*SD^2/d^2 + z_{\alpha}{}^2/4$
  - Compare proportions by χ^2 test: $n= (z_{\alpha/2}+z_{1-\beta})^2*\{p1(1-p1)+p2(1-p2)\}/(p1-p2)^2$
- Usually special softwares (nQuery, PASS, PS) or general statistics softwares (SAS, SPSS, STATA, EZR, R, etc.) will be applied.

Example of sample size calculation in hypothesis testing

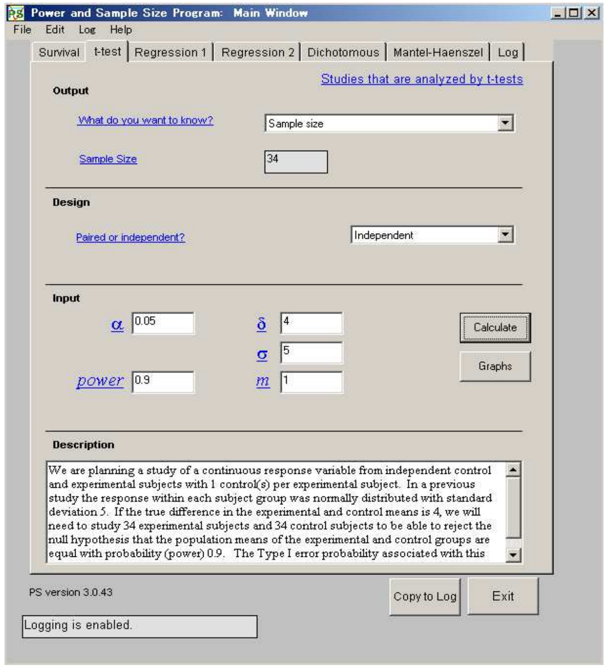
- Suppose we wish to compare the mean increased degree of elbow flexion between stimulated and control patients.
  - 4 degree difference has clinical importance.
  - Let alpha error 0.05 and statistical power 90%.
  - SD of increase of elbow flexion is assumed as 5 degree.
- Calculation by the previous equation
  - $n= 2*(z_{\alpha}-z_{1-\beta})^2*SD^2/d^2 + z_{\alpha}{}^2/4$
    = 2*(-1.64-1.28)^2*5^2/4^2+(-1.64)^2/4
    = 27.3174 ~ 27
- Based on this, 26 treated patients and 25 control patients were measured. They showed increase in elbow flexsion by 16±4.5 and 6.5±3.4, respectively. The mean difference was 9.5 (95%CI was 7.23 to 11.73), t-test resulted in t=8.43, df=49, p<0.001.
- Typical description of this design and statistical results should be written as follows:

Typical description

- We designed the study to have 90% power to detect a 4-degree difference between the groups in the increased range of elbow flexion. Alpha was set at 0.05. (Methods section)
- Patients receiving electrical stimulation (n=26) increased their range of elbow flexion by a mean of 16 degrees with a standard deviation of 4.5, whereas patients in the control group (n=25) increased their range of flexion by a mean of only 6.5 degrees with a standard deviation of 3.4. This 9.5-degree difference between means was statistically significant (95%CI = 7.23 to 11.73 degrees; two-tailed Student's t test, t=8.43; df=49; p<0.001). (Results section)
  - (Source: Lang and Secic, 2006, pp.47)
  - Here the expected standard deviation 5 nor applied equation is not clearly written (both seems implicit).

How to use “PS” software

- PS: Power and Sample Size Calculator
- http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/PowerSampleSize
- Free Software
- Survival (logrank test), t-test, Regression1, Regression2, Dichotomous (chisq-test), Mantel-Haenszel are included.
- An example of description is given as text.
- Ratio of two groups can be specified as m



Output

What do you want to know? Sample size

Sample Size 34

Design

Paired or independent? Independent

Input

α 0.05

power 0.9

δ 4

σ 5

m 1

Description

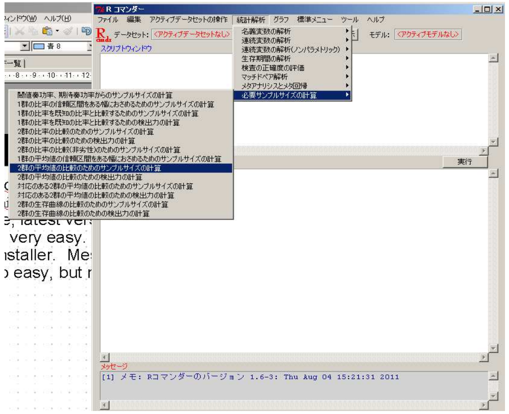
We are planning a study of a continuous response variable from independent control and experimental subjects with 1 control(s) per experimental subject. In a previous study the response within each subject group was normally distributed with standard deviation 5. If the true difference in the experimental and control means is 4, we will need to study 34 experimental subjects and 34 control subjects to be able to reject the null hypothesis that the population means of the experimental and control groups are equal with probability (power) 0.9. The Type I error probability associated with this

PS version 3.0.43

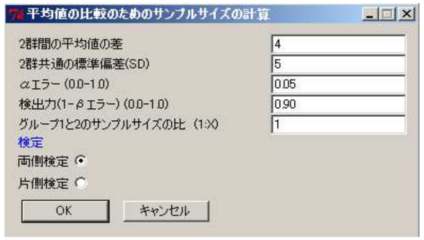
Logging is enabled.

How to use “EZR” software

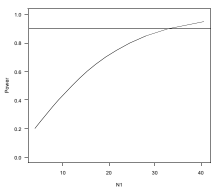
- EZR on Rcmdr is developed by Jichi Medical School
- http://www.jichi.ac.jp/saitama-sct/SaitamaHP.files/statmed.html
- Free software; latest version is 1.35 on 1 March 2017.
- Installation is very easy. Just click the downloaded executable installer. Messages are given in Japanese.
- Usage is also easy. Introductory textbooks are available in Japanese.

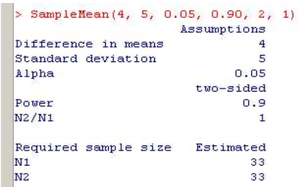

R コマンダー
ファイル
編集
アクティブデータセットの操作
統計解析
グラフ
標準メニュー
ツール
ヘルプ
データセット: <アクティブデータセットなし>
モデル: <アクティブモデルなし>
スクリプトウィンドウ
名義変数の解析
連続変数の解析
連続変数の解析(ノンパラメトリック)
生存期間の解析
検査の正確度の評価
マッチドペア解析
メタアナリシスとメタ回帰
必要サンプルサイズの計算
閾値奏功率、期待奏功率からのサンプルサイズの計算
1群の比率の信頼区間をある幅におさめるためのサンプルサイズの計算
1群の比率を既知の比率と比較するためのサンプルサイズの計算
1群の比率を既知の比率と比較するための検出力の計算
2群の比率の比較のためのサンプルサイズの計算
2群の比率の比較のための検出力の計算
2群の比率の比較(非劣性)のためのサンプルサイズの計算
1群の平均値の信頼区間をある幅におさめるためのサンプルサイズの計算
2群の平均値の比較のためのサンプルサイズの計算
2群の平均値の比較のための検出力の計算
対応のある2群の平均値の比較のためのサンプルサイズの計算
対応のある2群の平均値の比較のための検出力の計算
2群の生存曲線の比較のためのサンプルサイズの計算
2群の生存曲線の比較のための検出力の計算
実行
メッセージ

```
[1] メモ: Rコマンダーのバージョン 1.6-3: Thu Aug 04 15:21:31 2011
```



2群間の平均値の差: 4
2群共通の標準偏差(SD): 5
αエラー (0.0-1.0): 0.05
検出力(1-βエラー) (0.0-1.0): 0.90
グループ1と2のサンプルサイズの比 (1:X): 1
両側検定 ●
片側検定 ○





```
> SampleMean(4, 5, 0.05, 0.90, 2, 1)
                             Assumptions
Difference in means                    4
Standard deviation                     5
Alpha                               0.05
                               two-sided
Power                                0.9
N2/N1                                  1

Required sample size           Estimated
N1                                    33
N2                                    33
```

Using R console to calculate sample size

- > power.t.test(delta=4, sd=5, sig.level=0.05, power=0.9)

```
     Two-sample t test power calculation

              n = 33.82555
          delta = 4
             sd = 5
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

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

- The result is 34 (similar to PS)