サンプルサイズの問題

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

24 April 2013

参考文献，web サイト

- 英文の参考書
  - "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

仮想的な例

- 大学院在籍期間中に研究に応じてくれた患者の数が限られていたため，要因 Y が結果 Z（通常，その病気であること）と関連がないという特定の帰無仮説 X を検定するために，10 人の患者（とそれに対応する1～3倍程度の対照群）しか調査できなかったとする。検定の結果，有意水準 5% で帰無仮説 X は棄却されなかったとする。
- 公表バイアスを避けるために，「有意でない」結果も，"Lack of association ..." のようなタイトルを付けて投稿すべき。
- しかし，おそらく査読者は，この研究結果が「有意でなかった」のは，サンプルサイズが小さかった（言い換えると，検出力が足りなかった）からと判定する。これは研究デザインの致命的な欠点なので，論文はリジェクトされる。
- 大学院生としては学位が欲しいので，ここで統計学者に助けを求めることが多いのだが，この段階ではもはや手遅れ（せいぜい言い訳の仕方を教えるか，……）。
- これは大学院生に良くある悲劇である。

この院生はどうすべきだったのか？

- この研究は，仮説検定という枠組みで行われた。
- サンプルサイズを大きくすれば統計的検出力が大きくなることは既知なので，*研究開始前に*，十分な統計的検出力を得るために必要なサンプルサイズを決定することが可能なはず。
- サンプルサイズの検討をせずに研究を開始し，データを得てしまった後でできるのは，言い訳を書くことだけ。
  - 稀な疾患のため適当な研究期間内に同意を得られた患者が少なかった
  - 資金面での制約からサンプルサイズを大きくできなかった
  - その研究分野では伝統的にこの程度のサンプルサイズで研究が行われてきた，等。
- こういう研究であっても，結果は将来のメタアナリシスに貢献するので，査読を通ることは時々ある。しかし，本当は，事前の計算から必要なサンプルサイズはこれくらいと予想されるけれども，稀な疾患なので期間内に研究対象とできる患者数はこれくらいと予想されるため，これだけの患者が集まったら分析する，と事前にデザインしておき，研究ノートあるいは短報の形で投稿すべき（きわめて重要なテーマなら原著にも）

医学統計のテキストにはどう書かれているか

- 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.'**

サンプルサイズの計算をしない理由付け

- *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.

サンプルサイズの計算が不要な場合

- 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.

探索的研究では……

- 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.

探索的研究の例

- 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.

探索的研究の例（続き）

- 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

仮説検定の原理

- 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.

仮説検定におけるサンプルサイズ計算の例

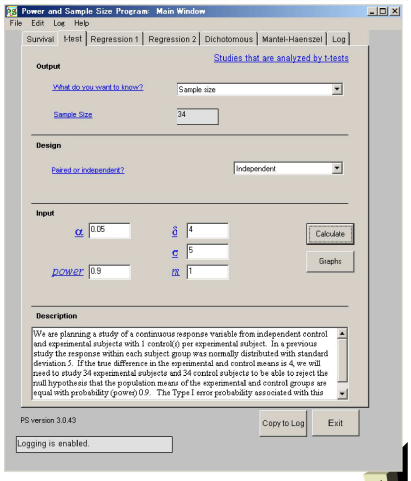
- Suppose we wish to compare the mean increased degree of elbow flexion between stimulated and control patients.
  - 4 degree difference has clinical importance.
  - Let alpha error 0.05 and statistical power 90%.
  - SD of increase of elbow flexion is assumed as 5 degree.
- Calculation by the previous equation
  - $n= 2*(z_{\alpha}-z_{1-\beta})^2*SD^2/d^2 + z_{\alpha}^2/4$
    $= 2*(-1.64-1.28)^2*5^2/4^2+(-1.64)^2/4$
    $= 27.3174 \sim 27$
- Based on this, 26 treated patients and 25 control patients were measured. They showed increase in elbow flexion 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:

典型的な記述の仕方の例

- 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. 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).
  (Lang and Secic, 2006, pp.47)
- Here the expected standard deviation 5 nor applied equation is not clearly written (both seems implicit).

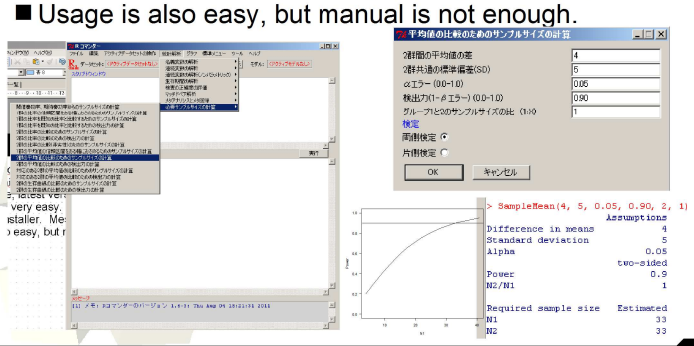
PS というソフトの使い方

- 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



EZR の使い方

- 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 1.01 on 1 April 2012.
- Installation is very easy. Just click the downloaded executable installer. Messages are given in Japanese.
- Usage is also easy, but manual is not enough.



R コンソールでは

- > 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)