Austin Biometrics and Biostatistics

Open Access

Austin Publishing Group

Research Article

Optimal Sample Size Allocation under Financial Constraint

Jiangao Luo[1]*, Yanpin Wang[2] and Jane Meza[1]
[1]Department of Biostatistics, University of Nebraska Medical Center, USA
[2]First National Bank, USA

***Corresponding author:** Jiangao Luo, Department of Biostatistics, College of Public Health, University of Nebraska Medical Center, 984375 Nebraska Medicine, Emile and 42nd St, Omaha, NE 68198-4375, USA

Received: March 09, 2015; **Accepted:** June 18, 2015; **Published:** July 07, 2015

Abstract

Optimal sample sizes for comparison of two groups with financial constraints are discussed in this paper. We study two types of optimal sample sizes under the financial constraints: (a) minimize the variance of the difference and ratio of two independent binary data under financial constraint, (b) maximize power for detecting the difference of two proportions, two survival rates and two correlations with financial constraint.

Keywords: Sample size; Power; Lagrange method; Financial constraint

Introduction

In the designing of medical studies we often face to decide the optimal sample sizes for interventions and controls. It has been decades since Cochran [1] studied the optimal sample size allocation under different sampling schemes. Allison et al. [2] have considered power, sample size and financial efficiency simultaneously. Guo et al. [3] have studied the sample size allocation ratio by minimizing the cost and maximizing the power. Guo and Luh [4] have also studied sample size allocation of comparing two trimmed means under given total cost. This is very important since nowadays investigators are facing the funding cut for their studies. Therefore it is crucial to get the optimal clinical trial results under financial cut and constraints. The main focus of this paper is to discuss how to get optimal precision for difference and ratio of two binary data and power for detecting the difference of two proportions, two survival rates and two correlations with financial constraints.

Minimal variance under financial constraints

Assume we have a clinical trial in which the sample sizes for intervention and control are n_1 and n_2, respectively. We use p_1 and p_2 to denote the proportions for binary responses for two groups, respectively. For continuous data, we use μ_1 and μ_2 for the means. Let $p = p_1 - p_2$ and $R = p_1/p_2$ and $\mu = \mu_1 - \mu_2$.

We assume that $n\hat{p}_1 \sim Bin(n_1, p_1)$, $n\hat{p}_2 \sim Bin(n_2, p_2)$ and $\hat{\mu}_1 \sim N\left(\hat{\mu}_1, \frac{\sigma_1^2}{n_1}\right)$, $\hat{\mu}_2 \sim N\left(\hat{\mu}_2, \frac{\sigma_2^2}{n_2}\right)$ respectively. Under the independent assumption

$$Var(\hat{p}) = \frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2} \quad (1)$$

and

$$Var(\hat{\mu}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \quad (2)$$

Using delta method we can get

$$Var(\hat{R}) = \left(\frac{p_1}{p_2}\right)^2\left(\frac{q_1}{n_1p_1} + \frac{q_2}{n_2p_2}\right) \quad (3)$$

Brittain and Schlesselman [5] have discussed the minimal solutions of (1) and (3) by finding the ratio of $\frac{n_1}{n_1+n_2}$. But in real situation we often face the constraint of budget. Say the costs for each subject in intervention (group 1) and control (group 2) are C_1 and C_2, respectively, and the total cost is C. Therefore, we have the following constraint

$$n_1C_1 + n_2C_2 = C \quad (4)$$

The optimal solution of (2) under (4) was given by Cochran [1] with

$$n_1 = \frac{C\sigma_1}{\sqrt{C_1}\left(\sigma_1\sqrt{C_1} + \sigma_2\sqrt{C_2}\right)}, \quad n_2 = \frac{C\sigma_2}{\sqrt{C_2}\left(\sigma_2\sqrt{C_1} + \sigma_2\sqrt{C_2}\right)} \quad (5)$$

or

$$n_1 : n_2 = \sigma_1\sqrt{C_2} : \sigma_2\sqrt{C_1} \quad (6)$$

Cochran obtained this result under the setting of optimum allocation of double sampling, C_1 and C_2 were unit sampling costs, respectively, and the structures of σ_1 and σ_2 were more complicated than here. Guo et al. [3] have proved that (6) also attains optimal power for fixed total cost.

The optimal solution of (1) under constraint (4) is

$$n_1 = \frac{C\sqrt{p_1q_1}}{\sqrt{C_1}\left(\sqrt{p_1q_1C_1} + \sqrt{p_2q_2C_2}\right)}, \quad n_2 = \frac{C\sqrt{p_2q_2}}{\sqrt{C_2}\left(\sqrt{p_1q_1C_1} + \sqrt{p_2q_2C_2}\right)} \quad (7)$$

according to Lagrange multiplier theory [6]. Therefore

$$n_1 : n_2 = \frac{\sqrt{p_1q_1}}{\sqrt{C_1}} : \frac{\sqrt{p_2q_2}}{\sqrt{C_2}} = \sqrt{p_1q_1C_2} : \sqrt{p_2q_2C_1} \quad (8)$$

but

$$n_1 : n_2 = \sqrt{p_1q_1} : \sqrt{p_2q_2}$$

if there is no financial constraint (4) according to [5].

Similarly, (3) is minimized under constraint (4) when

$$n_1 = \frac{C}{C_1}\frac{\sqrt{\frac{q_1C_1}{p_1}}}{\sqrt{\frac{q_1C_1}{p_1}} + \sqrt{\frac{q_2C_2}{p_2}}}, \quad n_2 = \frac{C}{C_2}\frac{\sqrt{\frac{q_2C_2}{p_2}}}{\sqrt{\frac{q_1C_1}{p_1}} + \sqrt{\frac{q_2C_2}{p_2}}} \quad (9)$$

which imply

$$n_1 : n_2 = \sqrt{\frac{q_1}{p_1C_1}} : \sqrt{\frac{q_2}{p_2C_2}} = \sqrt{\frac{q_1C_2}{p_1}} : \sqrt{\frac{q_2C_1}{p_2}} \quad (10)$$

Citation: Luo J, Wang Y and Meza J. Optimal Sample Size Allocation under Financial Constraint. Austin Biom and Biostat. 2015;2(3): 1024.

C_1 and C_2 are extra terms compared to the corresponding result in [5].

Example: Now consider the design for an experiment of binary data with p_1=*01*, p_2=*0.05*, C_1=*40*, C_2=*10* and C=*21750*. To minimize the variance of $\hat{p}$ we choose n_1=*399* and n_1=*579* according to our formula (7). This sample size allocation will give us *84%* power at significant level of *0.05* and precision $Var\left(\hat{P}\right)=0.000308$. If we choose n_1=n_2=*435* then we can only get *80%* power at level *0.05* and precision $Var\left(\hat{P}\right)=0.000316$.

Maximal power with financial constraints

For two independent samples of continuous data with hypotheses

$$H_0: \mu_1=\mu_2 \text{ vs } H_1: \mu_1\neq\mu_2 \tag{11}$$

the critical point for the power can be written as

$$Z_\beta=\frac{|\mu|-Z_\alpha\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}=\frac{|\mu|}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}-Z_\alpha \tag{12}$$

which is maximized when

$$\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}$$

is minimized. Therefore the test for hypotheses (11) reaches maximal power under financial constraint (4) when the sample sizes are allocated according to (5) (Guo et al. [3]). Namely the solution (5) simultaneously minimizes the precision and maximizes the power.

The critical point for the power of hypotheses:

$$H_0: p_1=p_2 \text{ vs } H_1: p_1\neq p_2 \tag{13}$$

for dichotomous data is given by Fleiss ([7])

$$Z_\beta=\frac{|p|-z_\alpha\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\overline{pq}}}{\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}} \tag{14}$$

with

$$\overline{p}=\frac{n_1p_1+n_2p_2}{n_1+n_2},\ \overline{q}=1-\overline{p},\ p=p_1-p_2$$

and Z_α and Z_β are the cut off points for type I and II errors, respectively, in normal distribution. The optimal solution of (14) under the constraint (4) has no closed form and we can only use iterative algorithm to get it. The details can be found in [6]. But as sample size n_1+n_2→∞ the solution of (14) with the constraint (4) is the same as (7).

Let us consider the survival analysis with two independent samples. First assume we are going to follow the subjects until the events. Then there is no censoring. For testing the hypotheses

$$H_0: \lambda_1=\lambda_2 \text{ vs } H_1: \lambda_1\neq\lambda_2 \tag{15}$$

Pasternack and Gilbert have given the following formula ([8])

$$Z_\beta=\frac{|\lambda_1-\lambda_2|-z_\alpha\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\overline{\lambda}^2}}{\sqrt{\frac{\lambda_1^2}{n_1}+\frac{\lambda_2^2}{n_2}}} \tag{16}$$

where

$$\overline{\lambda}=\frac{n_1\lambda_1+n_2\lambda_2}{n_1+n_2} \tag{17}$$

Again the optimal solution for (16) under constraint (4) has no closed form and asymptotic solution as n_1+n_2→∞ is given by

$$n_1=\frac{C\lambda_1}{\sqrt{C_1}\left(\lambda_1\sqrt{C_1}+\lambda_2\sqrt{C_2}\right)},\ n_2=\frac{C\lambda_2}{\sqrt{C_2}\left(\lambda_1\sqrt{C_1}+\lambda_2\sqrt{C_2}\right)} \tag{18}$$

Now we assume that there is censoring in the data since most of time we cannot follow all subjects until the events. Under some regular conditions we have

$$Z_\beta=\frac{|\lambda_1-\lambda_2|-z_\alpha\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\varphi\left(\overline{\lambda}\right)}}{\sqrt{\frac{\varphi(\lambda_1)}{n_1}+\frac{\varphi(\lambda_2)}{n_2}}} \tag{19}$$

where λ is given by (17) and

$$\varphi(\lambda)=\frac{\lambda^3T}{\lambda T-1+e^{-\lambda T}}$$

(see [9] for details). Obviously (16) is a special case of (19) with $\phi(\lambda)=\lambda^2$. We need to use the iterative Lagrange multiplier method to get the optimal solution of (19) under (4). The asymptotic solution is given by

$$n_1=\frac{C\sqrt{\varphi(\lambda_1)}}{\sqrt{C_1}\left(\sqrt{\varphi(\lambda_1)C_1}+\sqrt{\varphi(\lambda_2)C_2}\right)},\ n_2=\frac{C\sqrt{\varphi(\lambda_2)}}{\sqrt{C_2}\left(\sqrt{\varphi(\lambda_1)C_1}+\sqrt{\varphi(\lambda_2)C_2}\right)} \tag{20}$$

In many applications it is important to detect possible difference in correlations. Suppose we have two independent samples with correlations r_1 and r_2, respectively. Our hypotheses are

$$H_0: r_1=r_2 \text{ vs } H_1: r_1\neq r_2 \tag{21}$$

Then

$$Z_\beta=\frac{\left|z_{(r_1)}-z_{(r_2)}\right|}{\sqrt{\frac{1}{n_1-3}+\frac{1}{n_2-3}}}-Z_\alpha \tag{22}$$

where

$$Z_{(r)}=\frac{1}{2}\ln\frac{1+r}{1-r} \tag{23}$$

according to Fisher's arctanh transformation [10]. The hypotheses reach maximal power under constraint (4) when

$$n_1=3+\frac{C-3C_1-3C_2}{\sqrt{C_1}\left(\sqrt{C_1}+\sqrt{C_2}\right)},\ n_2=3+\frac{C-3C_1-3C_2}{\sqrt{C_2}\left(\sqrt{C_1}+\sqrt{C_2}\right)} \tag{24}$$

As an example we are going to prove (24) and show how to use Lagrange multiplier theory to prove similar results. In fact, to maximize the power we must maximize Z_β in (22). Equivalently, we only need to minimize

$$\frac{1}{n_1-3}+\frac{1}{n_2-3} \tag{25}$$

under constraint of (4). So the corresponding Lagrange multiplier function is

$$Q(n_1,n_2,\lambda)=\frac{1}{n_1-3}+\frac{1}{n_2-3}+\lambda\left(n_1C_1+n_2C_2-C\right) \tag{26}$$

and

$$\frac{\partial Q}{n_1}=-\frac{1}{(n_1-3)^2}+\lambda C_1=0$$

$$\frac{\partial Q}{n_2}=-\frac{1}{(n_2-3)^2}+\lambda C_2=0$$

imply

$$n_1=\sqrt{\frac{1}{\lambda C_1}}+3,\ n_2=\sqrt{\frac{1}{\lambda C_2}}+3 \tag{27}$$

Plugging (27) in (4) and solving forλ, we get

$$\lambda=\frac{\left(\sqrt{C_1}+\sqrt{C_2}\right)^2}{\left(C-3C_1-3C_2\right)^2} \tag{28}$$

Now we plug (28) in (27) and obtain (24). Since the Hessian matrix of Q with respect to n_1 and n_2 is

$$\begin{bmatrix}\frac{\partial^2Q}{\partial n_1^2} & \frac{\partial^2Q}{\partial n_1\partial n_2}\\ \frac{\partial^2Q}{\partial n_2\partial n_1} & \frac{\partial^2Q}{n_2^2}\end{bmatrix}=\begin{bmatrix}\frac{2}{(n_1-3)^3} & 0\\ 0 & \frac{2}{(n_2-3)^3}\end{bmatrix}$$

which is positive definite, (24) is the minimum solution for (26). Therefore it maximizes the power.

Other results can be proved similarly.

Example: Suppose $p_1=0.6$, $q_1=0.4$, $p_2=0.2$, $q_2=0.8$, $C_1=\$400$, $C_2=\$100$, and $C=\$1000$. If we want to test the difference at significant level 0.05 and 80% power with equal sample size, then $n_1=n_2=23$. This sample size allocation is going to give us the total cost of $1150, which is over the budget. If we use $n_1=n_2=20$ our power will be 75%, which is usually not acceptable. Now plug all the parameters in our formula (7), then we get $n_1=17.75$ and $n_2=28.99$. Choosing $n_1=18$ and $n_2=28$ we will get power of 80% and the actual type I error is 0.043, which is what we want.

Conclusion and Discussion

Cost constraints have important impact in the design of experimental studies. We have studied optimal sample allocation to achieve maximal precision and power under total financial constraints for comparison of two samples. The results are easy to program and therefore have broad applications. But we must point out that there are limits, say, the problems have been simplified and we do not consider recruitment and related costs. For rare disease, minimal sample size for fixed power and false positive rate is more important than fixed cost due to difficulty in recruiting. The applicability of the results is quite obvious.

Acknowledgement

Thanks are given to Dr. James Anderson for reading the manuscripts and providing useful suggestions. We appreciate the suggestive comments of an anonymous referee.

References

1. Cochran W. Sampling Techniques, 3rd edn. Wiley, New York. 1977; 448.
2. Allison DB, Allison RL, Faith MS, Paultre F, Pi-Sunyer FX. Power and Money: Designing Statistically Powerful Studies While Minimizing Financial Costs, Psychological Methods. 1997; 2: 20-33.
3. Guo JH, Chen HJ, Luh WM. Sample size planning with the cost constraint for testing superiority and equivalence of two independent groups. Br J Math Stat Psychol. 2011; 64: 439-461.
4. Guo JH, Luh WM. Optimum sample size allocation to minimize cost or maximize power for the two-sample trimmed mean test. Br J Math Stat Psychol. 2009; 62: 283-298.
5. Brittain E, Schlesselman JJ. Optimal allocation for the comparison of proportions. Biometrics. 1982; 38: 1003-1009.
6. Bertsekas DP. Nonlinear Programming, Athena Scientific, Belmont, MA. 1999.
7. Fleiss J. Statistical Methods for Rates and Proportions, Wiley, New York. 1973.
8. Pasternack BS, Gilbert HS. Planning the duration of long-term survival time studies designed for accrual by cohorts. J Chronic Dis. 1971; 24: 681-700.
9. Gross A, Clark V. Survival Distributions: Reliability Applications in Biomedical Science, Wiley, New York. 1975.
10. Freiman JA, Chalmers TC, Smith H Jr, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 "negative" trials. N Engl J Med. 1978; 299: 690-694.

Citation: Luo J, Wang Y and Meza J. Optimal Sample Size Allocation under Financial Constraint. Austin Biom and Biostat. 2015;2(3): 1024.