Shanghai Archives of Psychiatry, 2017, Vol. 29, No. 4

•BIOSTATISTICS IN PSYCHIATRY (40)•

Sample Size Calculations for Comparing Groups with Continuous Outcomes

Julia Z. ZHENG [1], Yangyi LI [2], Tuo LIN [3], Angelica ESTRADA [4], Xiang LU [5], Changyong FENG[3]*

Summary: Sample size justification is required for all clinical studies. However, to many biomedical and clinical researchers, power and sample size analysis seems like a magic trick of statisticians. In this note, we discuss power and sample size calculations and show that biomedical and clinical investigators play a significant role in making such analyses possible and meaningful. Thus, power analysis is really an interactive process and scientific researchers and statisticians are equal partners in the research enterprise.

Key words: sample size, continuous outcome, clinical study, power

[*Shanghai Arch Psychiatry.* 2017; **29**(4): 250-256. doi: http://dx.doi.org/10.11919/j.issn.1002-0829.217101]

1. Introduction

Sample size justification is required for all clinical studies. Although commercial and online statistical software have been developed to calculate sample sizes, for many biomedical and clinical researchers, the calculation of sample size seems like a magic trick of the statisticians. When their statisticians ask them for information pertaining to sample size calculations, many do not understand why statisticians ask them for such information.

Sample size, or power analysis, should be done at the design stage of a clinical study. In general, such calculations are based on statistical distributions of test statistics pertaining to study hypotheses. For adaptive designs [1], although sample size may be adjusted according to information accumulated after the study begins, the adjustment plan is pre-specified at the design stage.

Note that for some medical journals, editors often ask authors to calculate power of their completed studies and provide such information in their manuscripts. However, such post-hoc power analysis makes no statistical sense.[2] This is because although outcomes of a real study, along with their associated test statistics, are random quantities in the design stage, they all become non-random once a study is completed and have no probabilistic interpretation. Of course, the information in a completed study can be used for designs of future relevant studies.

As study outcomes are random, what is actually observed after a study is completed may be quite different from what has been proposed in the design. However, this does not mean that the study design is wrong or the study was not executed correctly. For example, suppose X is a standard normal random variable with mean 0 and standard deviation 1. The probability that $X > 1.96$ or $X < -1.96$ is 0.05. Thus, although we usually get a value of X within the range -1.96 to 1.96 when sampling X, there is still a 5% chance that X is outside of this range. Thus, when values of X are observed outside of the range, it does not mean that our assumption about the distribution of X is wrong.

[1] Department of Immunology and Microbiology, McGill University, Montreal, Canada

[2] Department of Mathematics, State University of New York in Stony Brook, Stony Brook, NY, USA

[3] Department of Mathematics, University of California in San Diego, San Diego, CA, USA

[4] Department of Physics, University of California in San Diego, San Diego, CA, USA

[5] Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY, USA

*correspondence: Changyong Feng. Mailing address: Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY 14642, USA. E-Mail: Changyong_Feng@URMC.Rochester.edu

In this manuscript we discuss sample size and power calculations for continuous outcomes. We give the sample size formulas for one group, two independent groups, and two paired groups. We show how preliminary information can be used to power studies. Our paper can demystify sample size justification for biomedical and clinical researchers.

2. Sample size for one group

We first consider sample size calculations for one group. Although relatively simpler, it helps illustrate basic steps for sample size calculations.

Consider a continuous outcome X and assume it has a normal distribution (often called bell-shaped distribution) with mean μ and variance σ^2, denoted by $X \sim N(\mu,\sigma^2)$. It is called the standard normal distribution if $\mu=0$ and $\sigma=1$. For ease of exposition, we assume first that σ is a known constant.

Consider testing the hypothesis,

$$H_0 : \mu=\mu_0 \text{ vs. } H_a : \mu\neq\mu_0, \quad (1)$$

where μ_0 is a known constant, and H_0 and H_1 are called null and alternative hypotheses, respectively. Note that as two-sided alternatives as in (1) are the most popular in clinical research, we only focus on such hypotheses in what follows unless stated otherwise.

Let X_1,X_2,K,X_n be a random sample from $X \sim N(\mu,\sigma^2)$ and $\overline{X}=n^{-1}\sum_{i=1}^{n}X_i$ be the sample mean. If the null H_0 is true, $\overline{X}$ has a high probability of being close to μ_0. However, because $\overline{X}$ is random, it is still possible for $\overline{X}$ to be far away from μ_0, although such probabilities are small, especially for large n. The type I error α, a quantity introduced to indicate such an error rate, is the probability that measures the likelihood when $\overline{X}$ is too far from μ_0 under H_0. This error rate is typically set at α=0.05 for most studies and at α=0.01 for studies with large sample sizes. Given α, power is the probability that we reject H_0 when H_0 is false.

The decision to reject the null is based on the standardized difference between $\overline{X}$ and μ_0, or the z -score [3]

$$z=\frac{\sqrt{n}(\overline{X}-\mu_0)}{\sigma}. \quad (2)$$

We reject H_0 if $|z|>z_{\alpha/2}$, where $z_{\alpha/2}$ denotes the upper $\alpha/2$ quantile of the standard normal distribution, i.e., $\Phi(z_{\alpha/2})=1-\alpha/2$, with Φ denoting the cumulative distribution function of the standard normal distribution. For example, for α = 0.05, $Z_{\alpha/2}$ = 1.96 . If $H_0 : \mu=\mu_0$ is true, the probability of rejecting H_0, therefore committing a type I error, is readily calculated as

$$P\left(\left|\frac{\sqrt{n}(\overline{X}-\mu_0)}{\sigma}\right|\geq z_{\alpha/2}\,|\,H_0\right)=\alpha. \quad (3)$$

In clinical studies, what we are really interested in is the opposite, i.e., how we can reject the null when the H_0 is false. This is because H_0 usually represents no treatment effect, i.e., a straw man. Statistical power allows one to quantify the chance of rejecting H_0 by specifying the mean μ under the alternative, i.e.,

$$H_0 : \mu=\mu_0 \text{ vs. } H_a : \mu=\mu_1,\ \ \mu_1\neq\mu_0. \quad (4)$$

Without loss of generality, we assume $\mu_1>\mu_0$. Note that unlike the hypothesis stated in (1), we must specify a known value for μ under the alternative H_a if we wish to quantify our ability to reject H_0 when performing power analysis. Such explicit specification is not needed when we only test the null hypothesis after data is observed.

Given type I error α and a specific μ_1 in H_a , we then calculate power, or the probability that (the absolute value of) the standardized difference in (2) exceeds the threshold $z_{\alpha/2}$, i.e.,

$$\text{Power}(n,\alpha,H_0,H_a)=P\left(\left|\frac{\sqrt{n}(\overline{X}-\mu_0)}{\sigma}\right|\geq z_{\alpha/2}\,|\,H_a\right). \quad (5)$$

By comparing the above with (3), we see that the only difference in (5) is the change of condition from H_0 to H_a. The probability is again readily evaluated to yield:

$$\text{Power}(n,\alpha,H_0,H_a)=1-\Phi\left(z_{\alpha/2}-\frac{\sqrt{n}(\mu_1-\mu_0)}{\sigma}\right)+\Phi\left(-z_{\alpha/2}-\frac{\sqrt{n}(\mu_1-\mu_0)}{\sigma}\right). \quad (6)$$

As the above shows, the power, $\text{Power}(n,\alpha,H_0,H_a)$, is a function of sample size n, type I error α and values of μ specified in the null H_0 and alternative H_a.

In most clinical research studies, μ_0 and μ_1 are posited to reflect treatment effects. Thus, once α is selected, power is only a function of sample size n , which increases as n grows and approaches 1 as n grows unbounded. Thus, by increasing sample size, we can have more power to reject the null, or ascertaining treatment effect.

However, as increasing sample size implies higher cost for studies, power is generally set at some reasonable level such as 0.80. Also, although we can detect any small treatment effect, such statistical significance may have little clinical relevance. Thus, it is critical that we specify treatment effects that correspond to clinically meaningful differences.

Sample size justification works the opposite way. Given a type I error α , a pre-specified power $1-\beta$, and H_0 and H_a, we want to find the smallest n such that the test has the given power to reject H_0 under H_a

$$\textbf{Power}(n,\alpha,H_0,H_a)\geq 1-\beta. \quad (7)$$

Although it is generally difficult to find an analytical formula for computing the smallest n satisfying (7), such an n is readily obtained by using statistical packages. Note that power in the literature is typically denoted by $1-\beta$, where β denotes the probability that the null H_0 is rejected when H_0 is false, or type II error rate.

Although $\mu_1-\mu_0$ measures treatment difference between the means of X under H_0, this difference depends on the scale of X and may change when different scales are used. For example, if X represents distance, $\mu_1-\mu_0$ will have different values if different scales are used such as mile and kilometer. Thus, effect size is used to remove such dependency:

$$\text{es} = \frac{|\mu_1-\mu_0|}{\sigma}. \quad (8)$$

The above is often referred to as Cohen's d and is widely used in clinical research. In the example of distance, effect size is the same regardless of whether mile or kilometer is used.

Note that for simplicity, we have assumed that σ^2 is known. In practice, σ^2 is also unknown and is estimated by the sample variance, $s^2 = n^{-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$. In this case, the above arguments still apply, but the cumulative normal distribution Φ will be replaced by the cumulative t distribution to account for sampling variability when estimating σ^2 by s^2.

3. Sample Size for Two Independent Groups

Now consider two independent samples and let X_{ij} ($i=0,1; j=1,\ldots,n_i$) denote the random outcomes from the two samples. We assume that both group outcomes follow normal distributions, $X_{ij}\sim N(\mu_i,\sigma_i^2)$, with unknown means μ_i and known variances σ_i^2 ($i=0,1$).

Considering testing the hypothesis,

$$H_0 : \mu_1-\mu_0=d_0 \text{ vs. } H_a : \mu_1-\mu_0\neq d_0 \quad (10)$$

Let $\overline{X}_{i\cdot}=n_i^{-1}\sum_{j=1}^{n_i}X_j$ denote the sample mean of the ith group ($i=0,1$). As in the one-group case, the difference between the two sample means $\overline{X}_{1\cdot}-\overline{X}_{0\cdot}$ should be close to d_0 if H_0 is true. Again, because $\overline{X}_{1\cdot}$ and $\overline{X}_{0\cdot}$ are random, it is still possible for $\overline{X}_{1\cdot}-\overline{X}_{0\cdot}$ to be very different from d_0, although such probabilities are small, especially for large n. The level of such type I error rate a is also set equal to 0.01 or 0.05 depending on sample size as discussed earlier.

Although most clinical trials allocate equal number of subjects into groups, some studies may assign more patients to a group.[4] We assume that the number of subjects in group 0 and group 1 are n_0 and n_1, respectively. If $H_0 : \mu_1-\mu_0=d_0$ is true, the probability of rejecting H_0, therefore committing type I errors, is readily calculated as:

$$P\left(\left|\frac{\overline{X}_{1\cdot}-\overline{X}_{0\cdot}-d_0}{\sqrt{\frac{\sigma_0^2}{n_0}+\frac{\sigma_0^2}{n_1}}}\right|\geq z_{\alpha/2}\,|\,H_0\right)=\alpha, \quad (11)$$

where α is the type I error level set *a priori* and $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution.

For power analysis, we again need to specify $\mu_1-\mu_0$ under H_a to quantify the ability to reject the null when performing power analysis, i.e.,

$$H_0 : \mu_1-\mu_0=d_0 \quad \text{vs.} \quad H_a : \mu_1-\mu_0=d_1,\ d_1\neq d_0, \quad (12)$$

Without loss of generality, we assume $d_1>d_0$. Given a significance level α, H_0 and H_a, we then calculate power, or the probability that (the absolutely value of) the standardized difference in (11) exceeds the threshold $z_{\alpha/2}$, i.e.,

$$\text{power}(n,\alpha,H_0,H_1)=P\left(\left|\frac{\overline{X}_{1\cdot}-\overline{X}_{0\cdot}-d_0}{\sqrt{\frac{\sigma_0^2}{n_0}+\frac{\sigma_1^2}{n_1}}}\right|\geq z_{\alpha/2}\,|\,H_a\right).$$

As in the one-group case, we use effect size as a measure of treatment effect when calculating power. In this case, Cohen's d is given by:

$$\text{es} = \frac{|d_1-d_0|}{\sqrt{\frac{1}{2}\left(\sigma_0^2+\sigma_1^2\right)}}.$$

In many studies, group variances are assumed the same, in which case the effect size reduces to

$$\text{es} = \frac{|d_1-d_0|}{\sigma}.$$

Given a type I error α, a power $1-\beta$, and H_0 and H_a, we can also find the smallest n such that the test has the given power to reject the null H_0 under H_a, i.e.,

$$\text{Power}(n,\alpha,H_0,H_a)\geq 1-\beta.$$

Again, statistical packages are readily applied to find such an n.

Note that for simplicity, we have again assumed that the group variances σ_i^2 are known. In practice, σ_i^2 are generally unknown and are estimated by the sample variance, $s_i^2=n_i^{-1}\sum_{j=1}^{n_i}(X_j-\overline{X}_i)^2$. In this case, the above arguments still apply, but the cumulative normal distribution Φ will be replaced by the cumulative t distribution to account for sampling variability when estimating σ_i^2 by s_i^2.

4. Sample Size for Paired Groups

In the last section, data from the two groups are assumed independent. When groups are formed by different subjects, they are generally independent. In practice, we may be interested in changes before and after an intervention. For example, suppose we are interested in the effect of a newly developed drug

on high blood pressure. We measure blood pressure of each subject before and after administering the drug and compare mean blood pressure between the two assessments. Since subjects with their blood pressure above the mean before the intervention are likely to stay above the mean blood pressure after the intervention, the two measures of blood pressure are not independent. As a result, the two independent group *t*-test does not apply to this paired group, or pre-post study, setting.

Let (X_{0j}, X_{1j}) denote the two paired outcomes from the j th pair. For each pair, treatment difference is $D_j = X_{1j} - X_{0j}$. If the difference D_j has a mean $d = 0$, then there is no treatment effect. In general, we are interested in testing the hypothesis

$$H_0 : d = 0 \text{ vs. } H_a : d \neq 0. \quad (13)$$

In the two independent group case, X_{0j} and X_{1k} are assumed to have their own means and the hypothesis (12) involves both group means. In the current paired-group case, it is not necessary to identify the means of X_{0j} and X_{1j}, since only the mean of difference D_j is of interest in the hypothesis (12). By comparing (4) and (13), it is readily seen that the sample size and power calculation is simply a special case of the one-group case with $H_0 : \mu = 0$.

5. Illustrations

In this section, we illustrate power and sample size calculations for the one group, two independent and two paired groups discussed using G*Power, a free program for power analysis, and R, a free package for statistical analysis, which also includes functions for power and sample size calculations for our current as well as more complex study settings.

Example 1. The mean weight of men aged 20-29 in a study population of interest in 1970 was $\mu_0 = 170$ lbs with a standard deviation σ=40 lbs. A researcher proposes that the mean weight of this subpopulation has increased to $\mu_1 = 190$ lbs in 2010 with the same standard deviation. The researcher wants to determine the sample size n so that there is 0.8 power to detect this difference.

The statistical hypotheses is

$$H_0 : \mu_0 = 170 \text{ vs. } H_a : \mu_1 = 190.$$

We set α=0.05. Although the alternative shows an increased weight, we compute power under a two-sided test. To compute power, we first convert the parameters into effect size:

When using the G*Power package, choose the following options (see Figure 1):

$$\text{es} = \frac{|\mu_1 - \mu_0|}{\sigma} = 0.5.$$

Test family > t tests

Statistical test > Means: Difference from constant (one sample case)

Type of power analysis > A priori: Compute required sample size

Tails > Two

Effect size d > 0.5

α err prob > 0.05

Power (1 - β err prob) 0.80

We obtain n=34 under Total sample size in the G*Power screen.

In R, we may use the pwr package to compute power. For *t*-tests, use the function:

pwr.t.test(n = , d = , sig.level = , power = , type = c("two.sample", "one.sample", "paired"))

where n is the sample size, d is the effect size, and type indicates a two-sample t-test, one-sample t-test or paired t-test. For each function, entering any three of the four quantities (effect size, sample size, significance level, power) and the fourth is calculated.

Using the function pwr.t.test (d = 0.5 , sig.level = 0.05 , power = 0.8 , type = "one.sample"), we obtain $n = 33$ after rounding to the nearest integer.

Example 2. A researcher, who wants to study the possible difference in hemoglobin between smokers μ_1 and non-smokers μ_0, would be interested to find any mean differences $d_1 = |\mu_1 - \mu_0| \geq 2$ mmol/L between the two study populations, with 80% power. The standard deviation in each group is assumed to be $\sigma = 5$ mmol/L for both groups (from other published studies).

The statistical hypothesis is

$$H_0 : \mu_1 - \mu_0 = 0 \text{ vs. } H_a : \mu_1 - \mu_0 = 2.$$

Again, we set α=0.05 and compute power for a two-sided test. Under the assumptions, the effect size is

$$\text{es} = \frac{|d_1 - d_0|}{\sigma} = \frac{|2 - 0|}{5} = 0.4.$$

We also assume a common group size so that $n_0 = n_1$. In G*Power package, choose the following options (see Figure 2):

Test family > t tests

Statistical test > Means: Difference between two independent means (two groups)

Type of power analysis > A priori: Compute required sample size

Tails > Two

Effect size d > 0.4

α err prob > 0.05

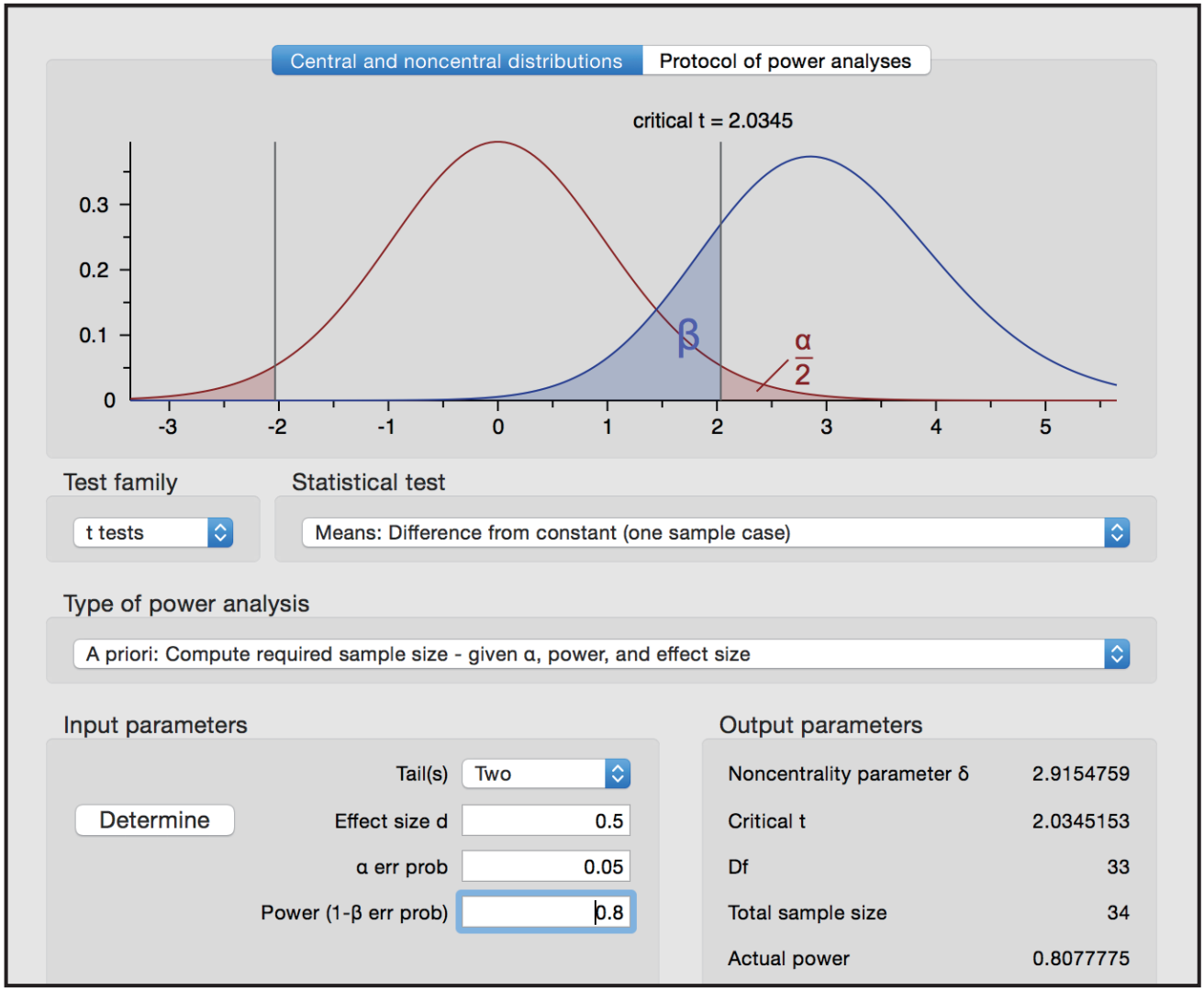
Power (1 - β err prob) 0.80

Allocation ratio N2/N1 > 1

From G*Power, we obtain $n_0 = n_1 = 100$ for each group or the total sample size $n_0 + n_1 = 200$.

Using the function pwr.t.test (d = 0.4 , sig.level = 0.05 , power = 0.8 , type = "two.sample") in R and rounding to the nearest integer, we obtain $n_0 = n_1 = 99$.

Figure 1. Screen shot from G*Power for Example 1



Example 3. A weight loss study using food diary wants to find a difference between pre- and post-intervention mean weight loss of $d = 2$ kg. The standard deviation of the difference d is assumed $\sigma_d = 5$ kg.

The statistical hypotheses is

$$H_0 : d = 0 \text{ vs. } H_a : d = 2.$$

We set α=0.05 and compute power for a two-sided test. The effect size is

$$\text{es} = \frac{|d_1 - d_0|}{\sigma} = \frac{|2-0|}{5} = 0.4.$$

By viewing the paired-group setting as a special case of the one-group setting, we readily obtain sample size using the following options in G*Power (see Figure 3):

Test family > t tests

Statistical test > Means: Difference from constant (one sample case)

Type of power analysis > A priori: Compute required sample size

Tails > Two

Effect size d > 0.4

α err prob > 0.05

Power (1 - β err prob) 0.80

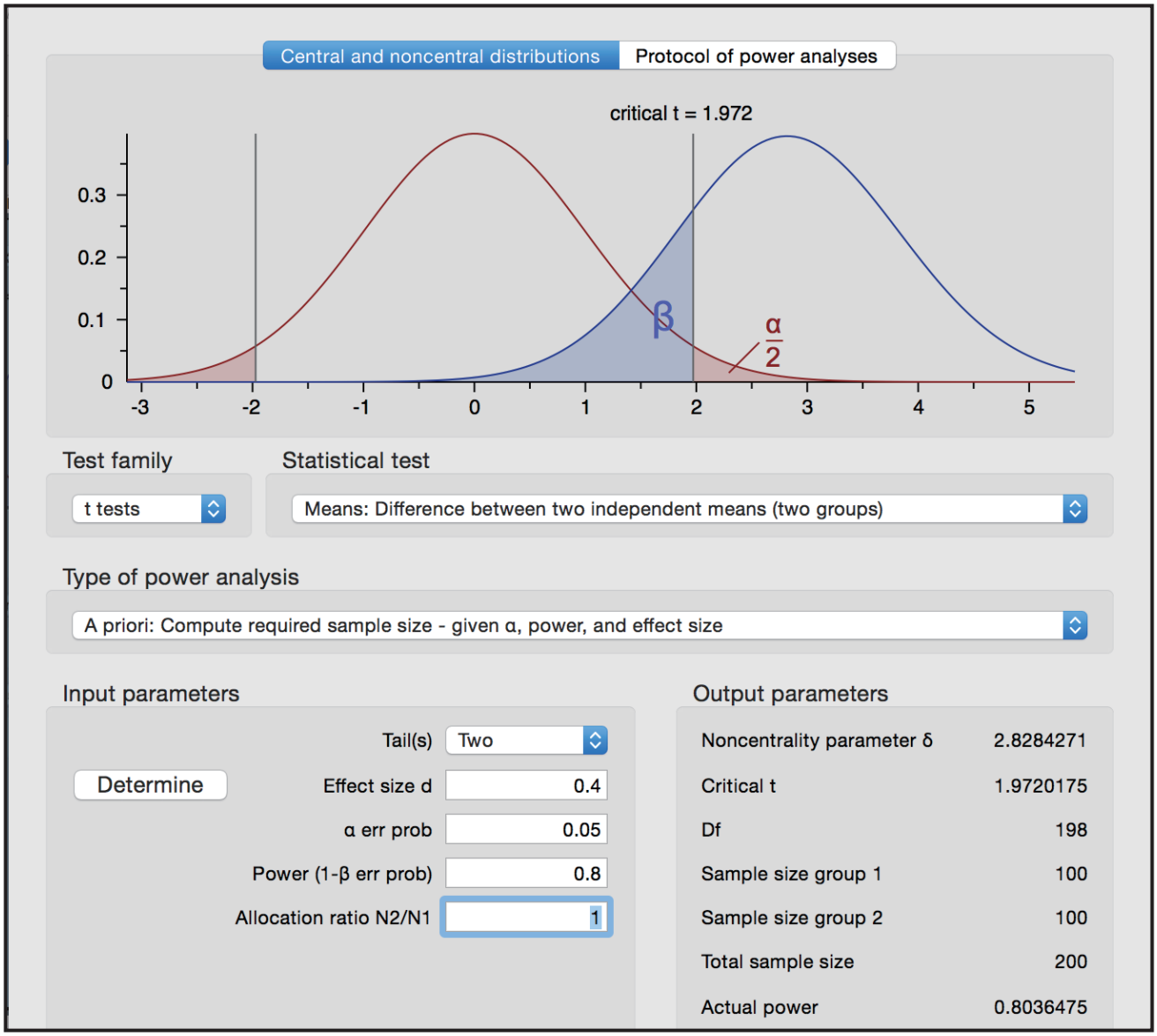
From G*Power, we obtain n=52.

Using the function pwr.t.test (d = 0.4 , sig.level = 0.05 , power = 0.8 , type = "paired") in R, we obtain n=51 after rounding to the nearest integer.

6. Conclusion

Sample size justification is an important consideration and a necessary component for clinical research studies. It provides critical information for assessing feasibility and clinical implications of such studies. Although power and sample size analysis relies on solid statistical theory and requires advanced computing methods, scientific investigators also play a critical role in this endeavor by providing relevant data. Without reliable input parameters, not only may power and sample

Figure 2. Screen shot from G*Power for Example 2



size analysis be less informative, but more important potentially yield misleading information for study planning and execution.

Funding statement
This study received no external funding.

Conflicts of interest statement
The authors have no conflict of interest to declare.

Authors' contributions
Julia Zhang, Yingyi Li, Tuo Li and Changyong Feng: Theoretical derivation and manuscript drafting.
Angelica Estrada, Xiang Lu and Changyong Feng: Computations of power and manuscript editing.

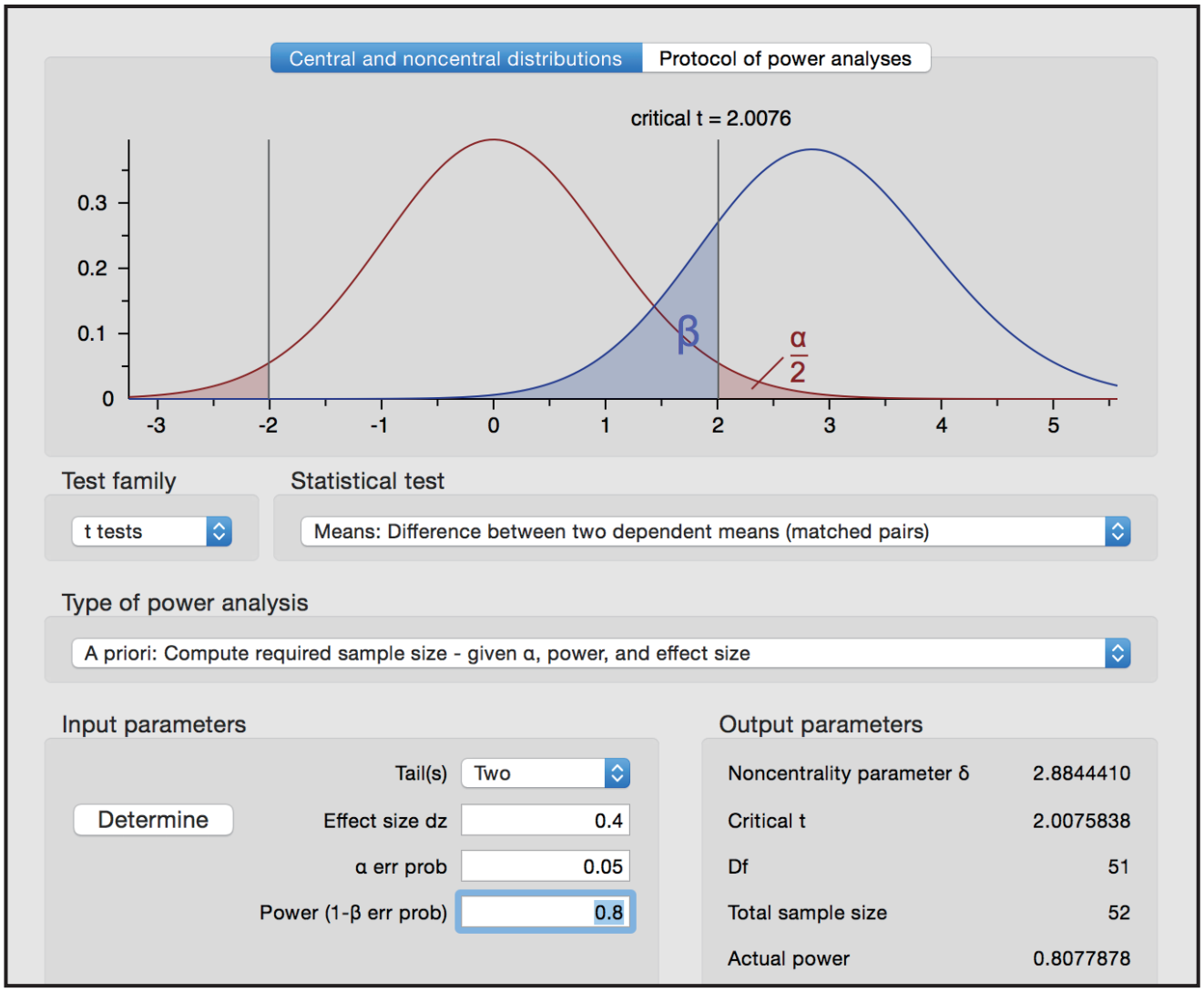
比较两组连续性结果的样本计算

Zheng JZ, Li Y, Lin T, Estrada A, Lu X, Feng C

概述： 所有的临床研究都需要对样本量进行辨证。然而，对于众多生物医学和临床研究人员来说，把握度和样本量看起来就像一个统计学家的魔术。在本文中，我们讨论了把握度和样本量的计算，并说明生物医学和临床研究人员在该分析的可行性和意义中具有重要作用。因此，把握度分析的确是一个互动的过程，并且科学研究人员和统计人员在研究团队中是平等合作的伙伴。

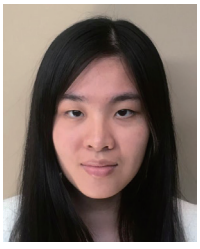
关键词： 样本量、连续性结果、临床研究、把握度

Figure 3. Screen shot from G*Power for Example 3



References

1. Chow SC, Chang M. *Adaptive design methods in clinical trials*. New York: Chapman & Hall / CRC; 2007
2. Heonig JM, Heisey DM. The abuse of power: the pervasive fallacy of power calculations for data analysis. *Am Stat*. 2001; **55**(1): 19-24. doi: http://dx.doi.org/10.1198/000313001300339897
3. Kreyszig E. *Advanced Engineering Mathematics (Fourth ed.)*. New York: Wiley; 1979
4. Moss AJ, Zareba W, Hall WJ, Klein H, Wilber DJ, Cannom DS, et al. Prophylactic implantation of a defibrillator in patients with myocardial infarction and reduced ejection fraction. *N Engl J Med. 2002;* **346**: 877--883. doi: http://dx.doi.org/10.1056/NEJMoa013474



Julia Zheng is currently completing her BS in Immunology and Microbiology at McGill University, Montreal, Canada. She is preparing to expand her interest in maths and computer science by pursuing a Bachelor's in Computer Science at University of Windsor, Windsor, Canada. In the future, Julia hopes to engage in a Master or PhD in Biostatistics to pursue her research interests in the fields of life sciences, computing biology, and biostatistics.