Statistical computing: effect size and sample size determination for two mean difference

Chung-Hong CHAN 陳仲康, Daniel Kwok-Keung NG 吳國強
Department of Paediatrics, Kwong Wah Hospital, Hong Kong

Effect Size

Statistical analysis is not only about p-value. The most important measure is actually the effect size (ES) *vide infra*. Effect size is a measure of the strength of relationship between two variables, e. g. correlation coefficient. ES is better than p-value because it not only conveys the information of "statistically significant", but also the size of the observed effects. An example of ES is Pearson's correlation coefficient (r). The magnitude of Pearson's r implied the strength of the correlation between two variables as well as the direction of the correlation. Pearson's correlation coefficient is an example of standardised ES in which the "r" from different studies can be directly compared.

For the difference between two means, the unstandardised ES such as the raw mean difference (Mean 1-Mean 2) is usually used. This unstandardised ES disallows comparison across studies. Comparison would be possible if this unstandardised ES is standardised as shown below.

Sample Size Determination

A common misconception of sample size determination is that this is a process that generate a sample size that guarantees a "statistically significant" result. However, a large enough sample size only guarantees the type II error rate (false negative).

Therefore, when your analysis gave a non-significant difference and your sample size is large enough, you can safety conclude that your result is not false negative. Large sample size is a two-edge sword. An extremely large sample guarantee a very slim possibility of false negative results, but it will increase the chance of type I error (false positive).

Therefore, one also needs to specify the type 1 error, usually 0.05 or 0.01. Larger than necessary sample size also implies more resources.

Parameters Required for Sample Size Determination

There are three mandatory parameters for sample size determination and they are:

1. Significance level (commonly used value is 0.05)
2. Statistical power (1-Type II error rate, commonly used value is 0.8)
3. Standardised ES

It is required to estimate the standardised ES. It can be found from the previous studies or a pilot study. For example, Marcus et al reported the mean nocturnal diastolic BP index of OSA patients and primary snorers (table). You are planning a similar study and you want to use Marcus et al's results as the blue print of your sample size determination.

Group	Mean ($\bar{x}$)	SD
OSA	-15	8
PS	-23	8

The unstandardised ES is -15 - (-23) = 8 mmHg. We required to transform this unstandardised ES to standardised ES. The standardised ES for mean difference is called Cohen's d. Cohen's d is calculated by:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(SD_1^2 + SD_2^2)/2}}$$

By plugging in the values from Marcus et al's results, the Cohen's d is

$$\begin{aligned} d &= \frac{-15 - -23}{\sqrt{(8^2 + 8^2)/2}} \\ &= 1 \end{aligned}$$

When we have no previous information which allow us to determine the estimated ES, we can use the scale suggested by Cohen (1992). Cohen suggested a "d" of 0.2 for small effect, 0.5 a medium and 0.8 a large ES. For example, if you want to determine the treatment effect of tiramitsu on asthma when compared to frappecino but you find no previous study. You can assume the effect size to be small (d=0.2) based on your clinical experience.

Computer Calculation

Commonly used software for sample size determination is PASS. However, it is an expensive commercial software. An open source software called R (http://www.r-project.org/) is available for sample size determination.

When R is installed in your computer, you need to install a package called pwr (required internet connection) and activate it with the following instructions:

```
>install.packages("pwr",
dependencies=TRUE)
>library(pwr)
```

In R, it is fairly straightforward to perform the sample size determination. In our example, we would like to determine the sample size to detect a ES of Cohen's d being 1 with significance level of 0.05 and power of 0.8. We can use the command pwr.t.test and supply the parameters of d, sig.level and power:

```
>pwr.t.test(d=1,power=0.8,
sig.level=0.05,type="two.sample",
alternative="two.sided")
```

The computer output will read like this:

Two-sample t test power calculation
n = 16.71473
d = 1
sig.level = 0.05
power = 0.8
alternative = two.sided
NOTE: n is number in *each* group

As n = 16.71473, we therefore need at least 17 subjects in each arm.