Education in Medicine Journal ISSN 2180-1932

EDUCATIONAL RESOURCE
Volume 6 Issue 2 2014
DOI: 10.5959/eimj.v6i2.252
www.eduimed.com

Calculating standard deviation of difference for determination of sample size for planned paired t-test analysis.

Wan Nor Arifin

Medical Lecturer, Unit of Biostatistics and Research Methodology, School of Medical Sciences, Universiti Sains Malaysia

ARTICLE INFO

Received : 28/01/2014
Accepted : 10/03/2014
Published : 01/06/2014

KEYWORD

Sample size
Paired-t test
Standard deviation of difference

ABSTRACT

Introduction: For pre-post and cross-over design analysis of numerical data, paired t-test is the simplest analysis to perform. In planning such studies, it is imperative to calculate appropriate sample size required for the test to detect hypothesized difference. However, the sample size formula requires determination of standard deviation of difference, which is not commonly reported. In this article, the author guides the reader to calculation of standard deviation of difference from standard deviation of each separate occasion.

CORRESPONDING AUTHOR: Dr. Wan Nor Arifin, Unit of Biostatistics and Research Methodology, School of Medical Sciences, Universiti Sains Malaysia, 16150 Kubang Kerian, Kelantan, Malaysia.
E-mail: wnarifin@usm.my

Introduction

In pre-post and cross-over design studies, essentially the aim is to determine whether there is any significant difference or change in values for a particular numerical variable (e.g. systolic blood pressure, weight) between two occasions for same subjects. For analysis of paired numerical data in the studies, the simplest analysis would be paired t-test. When such studies are planned, it is imperative to calculate required sample size for the use of paired t-test to detect hypothesized or clinically meaningful mean difference between the two repeated observations. Sample size formula for paired t-test [1,2] is given as

$$n = \left[\frac{\left(z_{1-\alpha} + z_{1-\beta}\right)\sigma_d}{\Delta} \right]^2$$

where

n sample size/number of pair
$z_{1-\alpha}$ corresponding z-value at chosen significance level, α
$z_{1-\beta}$ corresponding z-value at chosen power, 1–β
σ_d standard deviation of difference
Δ hypothesized mean difference/change

The values for all components of the formula are easily determined, with exception of value for standard deviation (SD) of difference, σ_d which is not commonly reported in journal articles. Often in journal articles, for analysis of paired-t, SDs for each occasion are given instead. As such, apart from conducting a pilot study just to determine the value, researchers are left in the dark looking for the elusive SD of difference to calculate the sample size for their proposals. The SD of difference is almost never known in

advance [2]. The researchers may use the largest of the SDs from separate occasions, which can be considered as an approximate value to SD of difference. However, it may lead to underestimation of sample size whenever the repeated observations are poorly correlated to each other. Thus in this short article, I would show that it is possible to calculate SD of difference using *variance sum law*.

Standard Deviation of Difference

Variance sum law for uncorrelated observations

Variance sum law [3] states that the variance of sum (σ^2_{X+Y}) or difference (σ^2_{X-Y}) of two uncorrelated variables equals the sum of the variance of the variables (σ^2_X, σ^2_Y),

$$\sigma^2_{X\pm Y} = \sigma^2_X + \sigma^2_Y$$

It does not matter whether we sum up X and Y or we subtract Y from X, the variance of X+Y and X–Y for a number of uncorrelated observations is similar. From this law, we are concerned with formula for variance of difference between X and Y,

$$\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y$$

To put the formula in our context, consider X as our post intervention values and Y as our pre-intervention values, thus variance of difference,

$$\sigma^2_d = \sigma^2_{post-pre} = \sigma^2_{post} + \sigma^2_{pre}$$

as we know, SD is the square root of variance, thus SD of difference,

$$\sigma_d = \sqrt{\sigma^2_{post-pre}} = \sqrt{\sigma^2_{post} + \sigma^2_{pre}}$$

As an example, in a pre-post intervention to study the effect of a drug on systolic blood pressure (SBP) reduction, it was reported that the SD of pre- and post-intervention SBP was 14.70mmHg and 11.66mmHg respectively. SD of difference was not reported. The calculated SD of SBP difference is

$$= \sqrt{\sigma^2_{post-pre}} = \sqrt{\sigma^2_{post} + \sigma^2_{pre}}$$
$$= \sqrt{11.66^2 + 14.70^2} = \sqrt{216.09 + 135.96}$$
$$= \sqrt{352.05} = 18.76$$

which can be used for sample size calculation using the formula mentioned previously. However, it would give you relatively large sample size because of the large SD of difference calculated. This is due to our use of *variance sum law*, which is meant for uncorrelated observations. In our context, the observations are actually correlated to some extend because of repeated observations on same subjects. Thus, if we know the degree of correlation between the observations, which is the Pearson's correlation coefficient *r*, we could use an extension to the law as discussed in following section.

Variance sum law for correlated observations

When the variables X and Y are correlated, according to *variance sum law* [3], the variance of difference is given as

$$\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y - 2\rho\sigma_X\sigma_Y$$

where ρ is our Pearson's correlation coefficient, *r*. To put it in our context, variance of difference

$$\sigma^2_d = \sigma^2_{post-pre} = \sigma^2_{post} + \sigma^2_{pre} - 2\rho\sigma_{post}\sigma_{pre}$$

and SD of difference

$$\sigma_d = \sqrt{\sigma^2_{post-pre}} = \sqrt{\sigma^2_{post} + \sigma^2_{pre} - 2\rho\sigma_{post}\sigma_{pre}}$$

By using our previous example, it is known that σ_{post} = 11.66, σ_{pre} = 14.70. Suppose that it was also reported for the study that the correlation between pre- and post-intervention SBP was *r* = 0.159. Thus, our calculated SD of SBP difference is

$$= \sqrt{\sigma^2_{post-pre}} = \sqrt{\sigma^2_{post} + \sigma^2_{pre} - 2\rho\sigma_{post}\sigma_{pre}}$$
$$= \sqrt{11.66^2 + 14.70^2 - 2\times 0.159\times 11.66\times 14.70}$$

$= \sqrt{352.05 - 54.51} = \sqrt{297.54}$

$= 17.25$

When correlation coefficient *r* between the observations is known, or if *r* can be estimated reliably based on *r* for other variables in same study, or also from expert opinion, we can easily obtain accurate SD of difference. In turn, it would reduce the calculated sample size for the planned study as compared to using the formula for uncorrelated observations.

Conclusion

In this article, I have shown the related formulas to determine the SD of difference. When *r* is known, it is recommended to use the formula for SD of difference for correlated observations. It would result in more accurate SD of difference and smaller sample size. In case when *r* is not known, but it can be estimated reliably, then formula for correlated observations can be used with caution. When r is not known and it is not possible to estimate it reliably, the safest way to determine the SD of difference is by using the formula for uncorrelated observations. Although it would give larger SD value and consequently larger sample size, we can avoid the possibility of underestimating the required sample size.

Reference

1. Naing N. A practical guide on the determination of sample size in health sciences research. Kota Bharu, Malaysia: Pustaka Aman Press; 2010.
2. Norman GR, Streiner DL. Biostatistics The Bare Essentials. 3rd ed. Ontario: BC Decker Inc; 2008.
3. Lane DM. Online Statistics Education: A Multimedia Course of Study. Rice University; 2013 [cited 2013, 15 December]; Available from: http://onlinestatbook.com/.

Further Reading

1. Machin D, Campbell MJ, Beng TS, Tan SH. Sample size tables for clinical studies. Singapore: Wiley-Blackwell; 2009.