Review article   # Some Notes on The Sample Size determination

Nasr Ibrahim Rashwan

Faculty of Commerce – Tanta University

nasrrashwan@yahoo.com.

Abstract

Sample size determination is often important step in planning a statistical study. Studies that are too small sizes of samples may lead to inaccurate results but studies that are too large sizes of samples may waste time and resources. Because the appropriate sample size determination depends on the type of data being collected of the study (categorical or continuous) and the purpose of the study (estimate or test). Therefore, in this paper we discuss the important factors that affect on calculating the appropriate sample size. These factors, include: the type of data, level of precision, the confidence level, the degree of variability, effect size, type I error, Type II error, significance level, power of the test and other factors. Also, we discuss and illustrate the different procedures (formulas) for sample size calculation.

Keywords: *Sample size, level of precision, Type I error, significance level, confidence level, Type II error and power.*

Introduction

One of the most frequently asked question in survey statistics is, "what sample size do I need?". The answer is very important because the adequate sample size determination plays an essential role in all a statistical studies and it is the most important design decision that faces the researchers (McClave et al, 2001).

The sample size (n) can be defined as the number of units in a group under study. An adequate sample size allows the researcher to report his results with sufficient degree of confidence and acceptable statistical power and helps ensure that the study will yield reliable information, interpretable results and minimizes research waste, but with inadequate sample size, making the interpretation of negative results difficult and yielding statistically in conclusive results.

In addition to, sample size is important for economic reasons: an undersized study can be waste time and resources without to produce useful information, while an oversized study uses more resources than are necessary. Most of statistical studies are always better when the sample size is adequate relative to the purpose of the study (Russell, 2001).

To avoid undersized and oversized studies, must be calculating of the adequate sample size, but before calculating the sample size of the study, we must answer to some the following questions:

- What is type the collective data of the study?

- What is the purpose of the study?

- How many samples will you calculate?

Because the answering on these questions was related to the important factors that affect on calculating of the sample size. Therefore, in this paper we focus on discussing of these factors. Also we discuss and illustrate the different formulas for sample size calculation and present some examples to show how to calculate the sample size?

- **Factors that affect on the sample size calculation :**

There are several factors must be take in consideration when the sample size determination, knowing these factors and their effect helps the researcher to determine the appropriate the sample size of the study. These factors can be summarized as follows:

(1) The type of collective data of the study.

The statistical methods of sample size determination depend on the type of data being collected. If the data are continuous, our concern will be about studying the means, but if the data are qualitative (categorical data), our concern will be about studying the proportions (Ralph et al, 2002).

(2) The level of precision.

The level of precision refers to the survey error or the sampling error. It is defined to be the difference between the parameter of population and the sample statistic that associated with this parameter. The level of precision is inversely proportional with the sample size. It means, the lower the precision level, the more sample size will be needed to reach the goal. This level is often expressed in percentage (such as $\pm 5\%$ or $\pm 7\%$). Note that, while determining the sample size, the level should be as low as possible without increasing the sample size to more than your resources will allow (Chadha 2006- Israel, 2009).

(3) The confidence level.

The confidence level is necessary when the results will be presented using confidence intervals. This level indicates to the probability value that the sample contains the parameter being estimated. Common confidence levels used in most of studies are 90% or 95%. If 95% confidence level is selected, this means that, if you repeat the experiment 100 times, there are 95 out of 100 times will have the true value of population parameter within the range of precision prespecified. The confidence level is directly proportional with the sample size. That is, the higher the confidence level predetermined, the larger the sample size will be needed to achieve the purpose of the study (Rahul, 2007).

(4) The significant level

The significant level refer to the probability of detecting a statistically significant difference that is the result of chance in other words, this level determines the probability of obtaining an erroneously significant results. Also, this factor refers to the maximum P-value for which a difference is to be considered statistically significant. The significant level is inversely proportional with the sample size, as this level is decreased; the sample size is needed to detect the difference increased. The level of significant is most commonly set a 5% or 1% chance of erroneously reporting a significant effect (Jan et al 2003).

(5) The effect size.

The effect size (minimum expected difference) is the smallest measured difference between comparisons groups that the researcher would like the study to detect. It measures the distance between the null hypothesis and a specified value of the alternative hypothesis. The effect size has a reverse relationship with the sample size. It means, the smaller the minimum expected difference is predetermined; the larger sample size is needed to detect a statistical significance. Note that, estimation of the quantity of effect size is not always straight forward because the setting of this factor is subjective and based on experience with the problem of the study. There are different ways to calculate effect size depending on the evaluation design you use, general size is calculated by taking the difference between the two groups and dividing it by the standard deviation of one of the groups (Adcock 1997, and Eng 2003).

(6) The degree of variability.

The degree of variability in the variables (attributes) being measured refers to the distribution of variables in the population. The more heterogeneous a statistical population (the more the degree of variability), the larger the sample size required to obtain level of precision. The more homogeneous a population, the smaller the sample size. This factor is considered a vital component of the sample size determination and is represented by the expected variance in the primary variables of interest in the study. Cochran(1977) listed several

ways of estimating population variance for sample size determination: 1) Use pilot study results; 2) Use data from previous studies of the similar population; 3) Estimate or guess the structure of the population assisted by some logical; 4) the standard deviation(σ) =(R/4), If our data are normally distributed or (σ) =(R/6), when the data we are dealing with are not normally distributed , where R is the range of the data (Kasiulevicius 2006).

(7) Type I error (α error).

Type I error is the probability that difference revealed by statistical analysis really do not exist or is the probability of rejecting a null hypothesis (H_0) even though it is true, this mean that, α error measures the probability of making this false – positive result $\left(=P\right.$ (reject H_0 \ H_0 True)$\left.\right)$. The α error is inverse relationship with the sample size. That is, the lower the α error is set, the more sample size will be needed to reach the goal of the study. Common type I error that used in determining the adequate sample size in most statistical studies is either 5% or 1% (Kenneth and Daviad; 2005)

(8) Type II error (β error).

Type II error occurs when statistical procedures result in a judgment of no significant differences when these differences do indeed exist or is the probability of accepting a null hypothesis even though it is false (=P (accept H_0 \ H_1 true)), where H_1 is alternative hypothesis). This means that, the type II error measures the probability of this false – negative result. The β error has a reverse relationship

6

with the sample size. That is, the smaller the β error is predetermined, the greater sample size is needed. A common β error that used in most statistical studies is 20% (Eng, 2003 – Jones et al 2003).

(9) Statistical power (1 - β)

The power of a statistical test derives from the β error; it is the complement of β (1 - β) and represents the probability of avoiding a false – negative result. This means, the β error is the probability of rejecting a false null hypothesis (=P (reject $H_0 \setminus H_1$ true)). Also the power is defined as its ability to detect a true significant difference between the two different groups, when the differences in fact exist. The power of a test has a direct relationship with the sample size. That is as power is increased, the sample size increases. The power is affected by many factors: the significant level (direct relationship), the effect size (direct relationship) and the sample size (direct relationship). The power also depends on whether one – sided or two – sided test is to be performed and the variability of the data under study. The statistical power is customarily set a number greater than or equal to 80%. A power of 80% means that the researcher is correct 8 items in 10 when accepting his result (Elise and Jonathan 2002, Chadha 2006).

(10) Others factors

There are other factors must be take into consideration when the sample size determination. For example,

- The sampling design

Note that, the approaches used in determining sample size in this paper have assumed that a simple random sample is the sampling design. More complex design, e.g., stratified random samples must take into consideration the variances of subpopulations.

- The sample size formulas provide the number of responses that need to be obtained. Many researches add 10% to the sample size to compensate for persons that researcher is unable to contact. Also the sample size often increased by 30% to compensate for non response (Israel, 2009).

Formulas for sample size determination:

There are several approaches to determine the sample size, these include using a census for small population as the sample, imitating a sample size of similar studies, using published tables and applying formulas to calculate a sample size. The best approach to determine the sample size is the application of one of several formulas. by using approach applying formulas, there are two approaches to sample size calculations : the first is a precision – based sample size calculations and the second is the power – based sample size calculations.

First: Precision – based sample size calculations in the case of continuous data.

1 - Sample size for the estimation of the mean (μ).

Where attention is concentrated on parameter estimation rather than hypothesis testing, the problem is to determine the necessary sample size to estimate a mean by confidence interval guaranteed to

be no longer than a prescribed length. If the population variance (σ^2) is known, the Z-statistic for the standard normal distribution may be used and the d is a half the confidence interval width (sampling error or absolute precision), then, a 100 (1 - α)% confidence interval for μ is :

$$\overline{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \qquad \therefore d = Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Where X is a sample mean (the estimated value of μ), $Z_{\alpha/2}$ is the $\alpha/2$ upper quantite of the stand normal and d is the level(absolute) precision. The sample size needed is (McClave et al 2007)

$$n = \frac{Z_{\alpha/2}^2 \cdot \sigma^2}{d^2} \qquad \dots 1$$

When the σ^2 is unknown, the S^2 (a pilot sample with size n* and variance (S^2) is used instead of σ^2, and the t-distribution should replace the normal distribution, then the sample size needed is :

$$n = \frac{t_{\alpha/2,\ n*-1}^2 \ S^2}{d^2} \qquad \text{if } n^* \text{ is small} \qquad \dots 2$$

but if the n* is large, the sample size is :

$$n = \frac{Z_{\alpha/2}^2 \ S^2}{d^2} \qquad \dots 3$$

Also, we can calculate the sample size based on the coefficient of variation (CV = S / $\overline{X}$, where $\overline{X}$ is a mean of a pilot sample). The sample size needed in this case is :

$$n = \frac{t^2_{\alpha/2 \, , \, n^*-1} \cdot CV^2}{(d/\overline{X})^2} \qquad \dots 4$$

Example : what is the sample size needed to estimate the population mean within d = 2 of the true mean with 95% confidence if a) the variance population = 8. b) There is a pilot sample with size = 5 and values are 5, 2, 3, 7 and 8.

Solution :

a) $\quad n = \dfrac{Z^2_{\alpha/2} \sigma^2}{d^2} = \dfrac{(1.96)^2 \, 8}{(2)^2} = 8$

b) $t_{.025,4} = 2.77$ ، $S^2 = 6.5 \qquad \because n = \dfrac{t^2_{\alpha/2 \, , \, n^*-1} \cdot S^2}{d^2}$

$\therefore n = \dfrac{(2.77)^2 (6.5)}{(2)^2} = 12.5$

Note that, by using iterative approach, we found that the sample size $n = (t^2_{\alpha/2 \, , \, n^*-1} \cdot S^2/d^2)$ is approximate equal to $\quad n = (Z^2_{\alpha/2} S^2/d^2)$

If the population is small (finite), then the sample size can be reduced it is possible to use a finite population correction factor (FPC = (N-n/N-1)), the sample size (n) in the case of infinite populations can be adjusted as follows (Israel, 2009).

$$n(\text{adj}) = \frac{n}{(1 + \frac{n-1}{N})} \quad \text{or} \quad \frac{nN}{n + (N-1)} \qquad \dots 5$$

where N is a population size.

The previous method in (1) and (2) can be applied to estimate a mean difference based on paired samples (μ_d). The sample size required is:

$$n = \frac{Z^2_{\alpha/2} \cdot \sigma^2_d}{d^2} \quad \text{or} \quad n = \frac{t^2_{\alpha/2, \, n^*-1} S^2_d}{d^2} \qquad \dots 6$$

Where σ^2_d is the population variance of the differences and S^2_d is the sample variance of the paired samples

2) **Sample size estimation for the comparison of two means:**

Assume that $n_1 = n_2 = n$, independent samples, normal populations and a common population variance $\sigma^2 = \sigma^2_1 = \sigma^2_2$, a 100 (1-α)% confidence interval for $\mu_1 - \mu_2$ is :

$$(\overline{X}_1 - \overline{X}_2) \pm Z_{\alpha/2} \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}} \quad \therefore d = Z_{\alpha/2} \sigma \sqrt{\frac{2}{n}}$$

Then, the sample size needed is $\quad n = \dfrac{2Z^2_{\alpha/2} \sigma^2}{d^2} \qquad \dots 7$

if $\sigma^2_1 \neq \sigma^2_2$, Then $\qquad n = \dfrac{Z^2_{\alpha/2} [\sigma^2_1 + \sigma^2_2]}{d^2} \qquad \dots 8$

Example : Suppose you want to be 95% sure that the true difference in mean is within ± 5 of your estimated difference in means. How many observations will you need to include this study if $\sigma_1 = \sigma_2 = \sigma = 20$?

Solution : $n = \dfrac{2Z_{\alpha/2}^2 \, \sigma^2}{d^2} = \dfrac{2(1\cdot96)^2 (20)^2}{(5)^2} = 123.$

If $\sigma_1^2 = \sigma_2^2$ and unknown, then we estimated these variances, by using a two pilot samples with sizes n_1^* and n_2^* (each sample size is large) and with variances S_1^2 and S_2^2.

$$n = \frac{2Z_{\alpha/2}^2 \, S_P^2}{d^2} \qquad \qquad \dots 9$$

where $S_P^2 = \dfrac{(n_1^* - 1)S_1^2 + (n_2^* - 1)S_2^2}{n_1^* + n_2^* - 2}$

but, if each sample size is small, then the sample size required is :

$$n = \frac{2t_{\alpha/2, \, n_1^* + n_2^* - 2}^2 \cdot S_p^2}{d^2} \qquad \qquad \dots 10$$

if $\sigma_1^2 \neq \sigma_2^2$ and unknown, the sample size needed is give by

$$n = \frac{Z_{\alpha/2}^2 \, (S_1^2 + S_2^2)}{d^2} \qquad \begin{array}{l} \text{If each sample of} \\ \text{two pilot samples is large} \quad \dots 11 \end{array}$$

or

$$n = \frac{t_{\alpha/2, \, df}^2 \, (S_1^2 + S_2^2)}{d^2} \qquad \begin{array}{l} \text{if each sample of two pilot} \\ \text{samples is small} \qquad \dots 12 \end{array}$$

Where df $= \dfrac{(S_1^2 / n_1^* + S_2^2 / n_2^*)^2}{\dfrac{(S_1^2 / n_1^*)^2}{n_1^* - 1} + \dfrac{(S_2^2 / n_2^*)^2}{n_2^* - 1}}$

Second :   Precision – based sample size calculation in the case of Categorical data.

(1) **Sample size for the estimation of the population proportion (P) :** when the population size is infinite or unknown, population is normal, the population proportion (P) is known from the previous similar studies, Then a 100 (1 - α)% confidence interval for P is given by :

$$\hat{P} \pm Z_{\alpha/2} \sqrt{\frac{P(1-P)}{n}} \quad \therefore d = Z_{\alpha/2} \sqrt{\frac{P(1-P)}{n}}$$

Where $\hat{P}$ is the estimated value of P (sample proportion)

Then, the sample size required is (kasiulevicius et al 2006, Geoffery, 2009) :

$$n = \frac{Z_{\alpha/2}^2 \, P(1-P)}{d^2} \qquad \dots 13$$

Where d is the level precision

Note that the determination of the needed size depends on P. in most case, the proportion P is unknown, we need a pilot study to give as estimate of P. The sample size in this case is

$$n = \frac{Z_{\alpha/2}^2 \, \hat{P}(1-\hat{P})}{d^2} \qquad \dots 14$$

Where $\hat{P}$ is a pilot sample proportion

If the actual variability value of an attribute of interest in the population is unknown and no pilot study performed beforetime, use

P= 0.5, since it will give us the most conservative estimate. The sample size required is :

$$n = \frac{0.25 \ Z_{\alpha/2}^2}{d^2} \qquad \dots 15$$

Example : A study is run to estimate the proportion of women who smoke in a certain city. The study involved a random of 100 women, and reported that they were smokes. Suppose that we wanted the estimate to be within 3% of the true proportion with a 95% confidence level. Determine the sample size required?

Solution : $P = \dfrac{28}{100} = 0.28$ from pilot study

$$\therefore n = \frac{Z_{\alpha/2}^2 \ P(1 - P)}{d^2} = \frac{(1.96)^2 \ 0.28(0.72)}{(.03)^2} = 860$$

If no pilot study, use $P = 0.5$ $\therefore n = \dfrac{(1.96)^2 (0.25)}{(.03)^2} = 1067$

If the population is finite, then the sample size required is the same formula in (5).

(2) Sample size estimation for the difference between two population proportions $(P_1 - P_2)$:

Assume that $n_1 = n_2 = n$, Normal populations, independent samples and P_1 and P_2 are known from the previous similar studies, Then the sample size required is (Chadha, 2006)

$$n = \frac{Z^2_{\alpha/2}\,[P_1(1-P_1) + P_2(1-P_2)]}{d^2} \qquad \dots 16$$

If no pilot study, use $P_1 = P_2 = 0.5$, Then the sample size is:

$$n = \frac{Z^2_{\alpha/2}\,[0.5(0.5) + 0.5(0.5)]}{d^2} = \frac{0.5\,Z^2_{\alpha/2}}{d^2} \qquad \dots 17$$

Example : The desired margin of error = 5% with 95% confidence. If the $P_2 = 0.3$ and $P_2 = 0.25$ from a pilot study, then

$$n = \frac{(1.96)^2\,[0.3(0.7) + 0.25(0.75)]}{(.05)^2} = 611$$

If not pilot study, then $\quad n = \dfrac{0.5(1.96)^2}{(.05)^2} = 768$

Third : power-based sample size calculations in the case of quantitative (continuous) data :

We have seen above that precision – based sample size calculations relate to estimation. On the other hand, power – based sample size

calculations relate to hypothesis testing. The most frequently used formulas can be summarized as follows:

1 -   Sample size determination for testing a population mean (μ).

Let μ is the true population mean, which can be estimated sample mean $\overline{X}$. if the test is two − sided, then the hypotheses of interest are given by :

$H_0 : \mu = \mu_0$ VS $H_1 : \mu \neq \mu_0$ or $H_0 : \Delta = 0$ VS $H_1 : \Delta \neq 0$

Where $\Delta = \mu - \mu_0$ is the difference between the true mean and the reference value (the expected mean difference or effect size). For a given level of significance (α) and the population variance σ^2, the null hypothesis H_0 is rejected if $Z = \left| \dfrac{\overline{X} - \mu_0}{\sigma / \sqrt{n}} \right| > Z_{\alpha/2}$ under the alternative hypothesis H_1, the power (1 - β) is equal to $\phi\left(\dfrac{\sqrt{n}|\Delta|}{\sigma} - Z_{\alpha/2} \right)$

where ϕ (Z) represent the area under the curve to the left of Z on a standard normal curve (a cumulative function).

The sample size needed for achieving the power can be obtained by solving the following equation (Chow et al, 2003).

$$Z_\beta = \frac{\sqrt{n}|\Delta|}{\sigma} - Z_{\alpha/2}$$

$\therefore$ The sample size $n = \dfrac{(Z_{\alpha/2} + Z_\beta)^2 \sigma^2}{\Delta^2}$ … 18

let $\alpha = 5\%$ and $\beta = 20\%$, you find $Z_{\alpha/2} = 1.96$, $Z_\beta = 0.84$

and $\left(Z_{\alpha/2} + Z_{\beta}\right)^2 = 8$ $\therefore n = \dfrac{8\sigma^2}{\Delta^2}$

If the test is one – sided, the sample size needed by using the same previous method is given by :

$$n = \dfrac{\left(Z_{\alpha} + Z_{\beta}\right)^2 \sigma^2}{\Delta^2} \qquad \dots 19$$

Note that, if σ^2 unknown, we can used S^2 (a pilot sample variance) instead of σ^2. The sample size needed is (David and Anthony, 2004) :

$$n = \dfrac{\left(Z_{\alpha/2} + Z_{\beta}\right)^2 S^2}{\Delta^2} \text{ (two – sided) or } \dfrac{\left(Z_{\alpha} + Z_{\beta}\right)^2 S^2}{\Delta^2} \text{(one sided)} \dots 20$$

This case is correct if a pilot sample size is large but if the pilot sample size (say n*) is small, the sample size needed is :

$$n = \dfrac{\left(t^2_{\alpha/2,\,n*-1} + t_{\beta,n*-1}\right)^2 S^2}{\Delta^2} \text{ (two-sided) or}$$

$$\dfrac{\left(t_{\alpha,n*-1} + t_{\beta,n*-1}\right)^2 S^2}{\Delta^2} \text{ (one – sided)} \qquad \dots 21$$

Example : Suppose from the similar study that it is estimated the standard deviation is equal 10. The sample size needed to achieve on 90% power at the 5% significance level for detecting the expected difference of 5 is given by :

- If the test is two – sided

$$\therefore n = \frac{(Z_{0.025} + Z_{0.10})^2 (10)^2}{(5)^2} = \frac{(1.96 + 1.28)^2 \, 100}{25} = 42$$

- If the test is one – sided

$$\therefore n = \frac{(Z_{.05} + Z_{.10})^2 (10)^2}{(5)^2} = \frac{(1.65 + 1.28)^2 \, 100}{25} = 34$$

(2) Sample size determination for testing the difference between two population means. In this case, the parameter for interest is given by $\mu_1 - \mu_2$, which can be estimated by $\overline{X_1} - \overline{X_2}$. Assume that $n_1 = n_2 = n$, independent samples, normal populations and a common population variance $\sigma^2 = \sigma_1^2 = \sigma_2^2$. The hypotheses of interest are given by:

$$H_0 : \Delta = 0 \quad \text{VS} \quad H_1 : \Delta \neq 0$$

Where $\Delta = \mu_1 - \mu_2$ (minimum expected difference). For a given α, H_0 is rejected if

$$Z = \left| \frac{\overline{X}_1 - \overline{X}_2}{\sigma \sqrt{\dfrac{1}{n} + \dfrac{1}{n}}} \right| \geq Z_{\alpha/2}$$

Under H_1, the power $= \varphi \left(\dfrac{|\Delta|}{\sigma \sqrt{\dfrac{2}{\sigma}}} - Z_{\alpha/2} \right)$. The sample size needed to achieve the desired power can be obtained by solving the following equation (Janet 2002, Chow et al 2003, Charan and Biswas 2013) :

$$Z_\beta = \frac{|\Delta|}{\sigma\sqrt{\dfrac{2}{n}}} - Z_{\alpha/2}$$

Then, the sample size (in each group) $n = \dfrac{2(Z_{\alpha/2} + Z_\beta)^2 \sigma^2}{\Delta^2}$ ... 22

or n (in each group, one-sided test) $= \dfrac{2(Z_\alpha + Z_\beta)^2 \sigma^2}{\Delta^2}$ ... 23

Example : assume that $\sigma_1 = \sigma_2 = \sigma = 8$, $\alpha = 5\%$, $\beta = 10\%$ and the minimum difference we wish to detect between the two means $(\Delta) = 2$. we have

$$n \ (two - sided) = \frac{2(Z_{0.025} + Z_{0.10})^2 (8)^2}{(2)^2} = \frac{2[1.96 + 1.29]^2 (8)^2}{(2)^2} = 336$$

and $$n \ (one - sided) = \frac{2(Z_{0.05} + Z_{0.10})^2 (8)^2}{(2)^2} = \frac{2[1.65 + 1.29]^2 (8)^2}{(2)^2} = 274$$

This mean that, a total number of 672 subjects are needed for achieving a 90% power for detection D = 2 at the 5% level of significance if two – sided test. But a total number of 548 subjects when one – sided test.

In most cases, the two samples are not equal. Assume that there is relationship between n, and n_2, say $n_1 = Kn_2$. The sample size is given by (two – sided) :

$$n_1 = K n_2 \quad \text{and} \quad n_2 = \frac{(Z_{\alpha/2} + Z_\beta)^2 \sigma^2 [1 + (1/k)]}{\Delta^2} \qquad \dots 24$$

or in one sided – test

$$n_1 = K n_2 \quad \text{and} \quad n_2 = \frac{(Z_\alpha + Z_\beta)^2 \sigma^2 [1 + (1/k)]}{\Delta^2} \qquad \dots 25$$

If $\sigma_1^2 = \sigma_2^2$ and unknown, then by using a pilot samples to estimate σ_1^2 and σ_2^2. When a pilot samples with sizes n_1^* and n_2^* and each is large, sample size. Needed is given by :

$$n \text{ (in each group)} = \frac{2(Z_{\alpha/2} + Z_\beta)^2 S_P^2}{\Delta^2} \text{(two-sided test)} \qquad \dots 26$$

and

$$n \text{ (in each group)} = \frac{2(Z_\alpha + Z_\beta)^2 S_P^2}{\Delta^2} \text{(one-sided test)} \qquad \dots 27$$

where

$$S_P^2 = \frac{(n_1^* - 1)S_1^2 + (n_2^* - 1)S_2^2}{n_1^* + n_2^* - 2}$$

but if each sample size is small, then the sample size required is (Wei et al 2010l; Chow et al 2003) :

$$n \text{ (in each group)} = \frac{2S_P^2[t_{\alpha/2, n_1^* + n_2^* - 2} + t_{\beta, n_1^* + n_2^* - 2}]^2}{\Delta^2} \quad \begin{matrix} \text{Two-sided} \\ \text{test} \end{matrix} \dots 28$$

and

$$n = \frac{2S_P^2[t_{\alpha, n_1^* + n_2^* - 2} + t_{\beta, n_1^* + n_2^* - 2}]^2}{\Delta^2} \quad \text{(one – sided test)} \quad \dots 29$$

when $\sigma_1^2 \neq \sigma_2^2$ and unknown, and using a pilot study with each sample size is large, the sample size is :

$$n = \frac{(Z_{\alpha/2} + Z_\beta)^2 [S_1^2 + S_2^2]}{\Delta^2} \quad (\text{Two} - \text{sided test}) \qquad \ldots 30$$

and

$$n = \frac{(Z_\alpha + Z_\beta)^2 [S_1^2 + S_2^2]}{\Delta^2} \quad (\text{one} - \text{sided test}) \qquad \ldots 31$$

but each a pilot sample is small, the sample size required is (David and Anthony, 2004) :

$$n = \frac{(S_1^2 + S_2^2)[t_{\alpha/2, df} + t_{\beta, df}]^2}{\Delta^2} \quad (\text{Two} - \text{sided test}) \qquad \ldots 32$$

and

$$n = \frac{(S_1^2 + S_2^2)[t_{\alpha, df} + t_{\beta, df}]^2}{\Delta^2} \quad (\text{one sided test}) \qquad \ldots \quad 33$$

Fourth : power-based sample size calculations in he case of categorical data.

(1) Sample size determination for testing a population proportion.

Let P is the true population proportion which can be estimated by $\hat{P}$ .for a given significance level α and the population proportion P from a previous study is known. The hypotheses of interest are given by :

$$H_0 : \Delta = 0 \qquad VS \qquad H_1 : \Delta \neq 0 \quad (Two - sided\ test)$$

Where $\Delta = P - P_0$ is the parameter of interest (difference between the true proportion and reference value). The H_0 is rejected if

$$Z = \left| \frac{\sqrt{n}\Delta}{\sqrt{P(1-P)}} \right| > Z_{\alpha/2} \text{ . under the } H_1, \text{ the power } = \varphi \left(\frac{\sqrt{n}|\Delta|}{\sqrt{P(1-P)}} - Z_{\alpha/2} \right)$$

The sample size needed can be obtained by solving the following equation (Chow et al, 2003) :

$$Z_\beta = \frac{\sqrt{n}|\Delta|}{\sqrt{P(1-P)}} - Z_{\alpha/2} \quad \text{which leads to}$$

$$n = \frac{(Z_{\alpha/2} + Z_\beta)^2 P(1-P)}{\Delta^2} \qquad (two - sided\ test) \qquad \dots 34$$

$$\text{and} \quad n = \frac{(Z_\alpha + Z_\beta)^2 P(1-P)}{\Delta^2} \qquad (one - sided\ test) \qquad \dots 35$$

22

(2) Sample size determination for testing the difference between two population proportions.

In this case, the parameter of interest is given by $P_1 - P_2$ which can be estimated by $\hat{P}_1 - \hat{P}_2$ .To calculate the sample size needed to the difference between two proportions (two binary variables). We assume that the samples are independent and because the sample size in the two groups are equal, the average proportion is $P = (P_1 + P_2)/2$. Under these assumptions, the binary variable has a binomial distribution, the samples size for each the two groups is given by (Geoffrey, 2009)

$$n = \frac{2[Z_{\alpha/2} + Z_\beta]^2 \, \overline{P}(1 - \overline{P})}{\Delta^2} \quad \text{(Two-sided test)} \qquad \dots 36$$

or

$$n = \frac{2[Z_\alpha + Z_\beta]^2 \, \overline{P}(1 - \overline{P})}{\Delta^2} \quad \text{(one-sided test)} \qquad \dots 37$$

This simple formula uses the same variance under both H_0 and H_1. if the variances differ, there is a more accurate formula is given by :

$$n = \frac{\left[Z_\alpha \sqrt{Z\overline{P}(1 - \overline{P})} + Z_\beta \sqrt{P_1(1 - P_1) + P_2(1 - P_2)} \right]^2}{\Delta^2} \quad \text{(two-sided test)} \quad \dots 38$$

or

$$n = \frac{[Z_\alpha \sqrt{Z\overline{P}(1 - \overline{P})} + Z_\beta \sqrt{P_1(1 - P_1) + P_2(1 - P_2)}]^2}{\Delta^2} \quad \text{(one-sided test)} \quad \dots 39$$

Note that, these formulas does not include a continuity correction (bring a normal distribution in closer agreement with binomial distribution) applying the correction, the sample size needed is given by (Fleiss et al, 1980) :

$$n = \frac{n}{4}\left[1 + \sqrt{1 + \frac{4}{n|\Delta|}}\right]^2 \qquad \dots 40$$

all the previous formulas for testing $P_1 - P_2$, we can be approximated as the following formula (Elise and Jonathan, 2002) :

$$n = \frac{(Z_{\alpha/2} + Z_\beta)^2 [P_1(1 - P_1) + P_2(1 - P_2)]}{\Delta^2} \quad (two - sided\ test) \qquad \dots 41$$
or

$$n = \frac{(Z_\alpha + Z_\beta)^2 [P_1(1 - P_1) + P_2(1 - P_2)]}{\Delta^2} \quad (one - sided\ test) \qquad \dots 42$$

if the two samples are net equal, let $n_1 = Kn_2$, then the sample size is given by :

$$n_1 = kn_2 \quad , \quad n_2 = \frac{(Z_{\alpha/2} + Z_\beta)^2 [P_1(1 - P_1) + P_2(1 - P_2)]}{\Delta^2} \qquad \dots 43$$

or

$$n_1 = kn_2 \quad , \quad n_2 = \frac{(Z_\alpha + Z_\beta)^2 [P_1(1 - P_1) + P_2(1 - P_2)]}{\Delta^2} \qquad \dots 44$$

Example : Suppose, we wish to design a study to compare two treatments. A two – sided test is planned at a 5% significance level. Based on a pilot study, $P_1 = 0.20$ and $P_2 = 0.10$. What is the sample size would be needed per group to detect the difference between P_1 and P_2 with a 90% power ?

Solution

- by using $n = \dfrac{(Z_{\alpha/2} + Z_\beta)^2 \overline{P}(1-\overline{P})}{\Delta^2} = \dfrac{2[1.96 + 1.28]^2 (.15)(.85)}{(0.20 - 0.10)^2} = 267$

where $\overline{P} = \dfrac{0.20 + 0.10}{2} = 0.15$

- by using $n = \dfrac{[Z_{\alpha/2}\sqrt{2\overline{P}(1-\overline{P})} + Z_\beta\sqrt{P_1(1-P_1) + P_2(1-P_2)}]^2}{\Delta^2}$

$= \dfrac{[1.96\sqrt{2(.15)(.85)} + 1.28\sqrt{.20(.80) + .10(.90)}]^2}{(0.10)^2} = 265$

- by using $n = \dfrac{(Z_{\alpha/2} + Z_\beta)^2 (P_1(1-P_1) + P_2(1-P_2))}{\Delta^2}$

$= \dfrac{(1.96 + 1.28)^2 [0.20(.80) + .10(.90)]}{(.10)^2} = 262$

Fifth : Sample size determination for more complex study design

- For example, comparison of k groups (one – way ANOVA). Consider first the comparison of k treatments with n subjects studied in each. Let $\mu_1, \mu_2, \ldots, \mu_k$ denote the k underlying means, σ^2 denote the assumed common variance, and assume that the subjects in each group are independently and normally distributed. The following hypotheses are usually considered :

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k \quad VS \quad H_1 : \mu_i \neq \mu_j \quad 1 \leq i < j \leq k$$

For a given α, the H_0 is rejected if $F = \dfrac{SSB/K-1}{SSE/K(n-1)} > F_{\alpha, k-1, k(n-1)}$

Where SSB is the sum squares between treatments, SSE is the sum of squares error (within treatments) and $F_{\alpha, k-1, k(n-1)}$ denote the α upper quantile of the F distribution with (k-1) and k (n-1) degrees of freedom.

Under the H_1 the power is equal (approximated) (Chow et al, 2003).

$$P\left[F > F_{\alpha, k-1, k(n-1)}\right] \approx P\left[nSSB > \sigma^2 \chi^2_{\alpha, k-1}\right]$$

Where σ^2 can be estimated by $\hat{\sigma}^2 = SSE / k(n-1)$ and $\chi^2_{\alpha, k-1}$ represents the α upper quantile for a χ^2 distribution with k-1 degree of freedom. Under the H_1, $nSSB / \sigma^2$ is distributed as a noncentral χ^2 with (k-1) degree of freedom and the noncentrality parameter $\lambda = n \Delta$, then the sample size needed (in each group) is given by

$$n = \frac{\lambda}{\Delta} \qquad \ldots 45$$

where

$$\Delta = \frac{1}{\sigma^2} \sum_{i=1}^{k} (\mu_i - \bar{\mu})^2 \quad , \quad \bar{\mu} = \frac{1}{k} \sum_{i=1}^{k} \mu_i$$

and the value of λ can be obtained by solving the following equation

$$\chi^2_{k-1}(\chi^2_{\alpha,k-1} / \lambda) = \beta$$

where $\chi^2_{k-1}(\cdot \setminus \lambda)$ is the cumulative distribution function of the noncentral χ^2 distribution with (k-1) degree of freedom and noncentrality parameter λ.

- Sample size determination for estimating correlation coefficient (ρ)

  let is the correlation coefficient between two variables, say X and Y. For testing the hypotheses $H_0 : \rho = 0$ VS $H_1 : \rho \neq 0$ the test statistic is of the fisher's Z transformation can be written as :

$$t = \frac{1}{2} \ln \left[\frac{1+r}{1-r} \right] \sqrt{n-3}$$

where r is the estimate of the correlation coefficient between X and y. it can be show that t is normally distributed the sample size needed is given by (Hsieh et al 1998)

$$n = \frac{(Z_{\alpha/2} + Z_\beta)^2}{\frac{1}{4} \left[\ln(\frac{1+r}{1-r}) \right]^2} + 3 \qquad \dots 46$$

Note that, the sample size in equation 46 is the same the sample size needed for a simple linear regression model.

Example : the sample size required to detect a correlation coefficient of 0.7 at the significance level of 5% and power 80% is

$$n = \frac{(1.96 + 0.84)^2}{\frac{1}{4}\left[\ln(\frac{1+0.7}{1-0.7})\right]^2} + 3 = 14$$

- Sample size determination for multiple regression model.

  For multiple regression, the formula used to determine the planned sample size depend on confidence intervals around population regression coefficient β_j (we use standardized regression coefficient), which can be written as follows:

$$\hat{\beta}_j \pm t_{\frac{\alpha}{2}, n-p-1} \sqrt{\frac{1 - R^2}{(1 - R^2_{xx_j})(n - p - 1)}} \qquad J = 1, 2, \dots, p$$

Where P is the number of predictors variables. R^2 : the coefficient of determination and $R^2_{xx_j}$ is the coefficient of determination in the regression of predictor variable X_j on the remaining predictors variables (p-1) of the model (Green 1991, Scott 2000, Ken and Scott, 2002) by using $Z_{\frac{\alpha}{2}}$ instead of $t_{\frac{\alpha}{2}, n-p-1}$ (this approximation lead to increase in precise of estimates), then the previous equation become as follows :

$$\hat{\beta}_j \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{1 - R^2}{(1 - R^2_{xxj})(n - p - 1)}}$$

$$\text{Let, } d = Z_{\alpha/2} \sqrt{\frac{1-R^2}{(1 - R^2_{xx_j})(n - p - 1)}}$$ where d is the desired half-width of

confidence interval. By solving this equation for n, then the sample

size needed is given by :

$$n = (\frac{Z_{\alpha/2}}{d})^2 (\frac{1-R^2}{1 - R^2_{xx_j}}) + P + 1 \qquad \dots 47$$

Also, Tabachnick and Fidell (2001) used the following rules to

determine the sample size :

- A rule of Thumb for testing β's is given by : $n \geq 104 + P$ … 48
- A rule of Thumb for (forecasting) testing R^2 is : $n \geq 50 + 8P$ … 49
- If you are using stepwise regression   : $n \geq 40 P$ … 50
- When the dependent variable is skewed, the sample size needed is
  given by :

$$n \geq \left(\frac{8}{F^2}\right) + (P-1) \qquad \dots 51$$

where f^2 = .01,0.15 and 0.35 for small, Medium and large effect size.

 Example: If P =2, then the sample size for (structural analysis)
testing β's is n ≥ 106 observations, the sample size for (forecasting)
testing R^2 is n ≥ 66 observations, the sample size for using stepwise
regression is n ≥ 80 observations and when the dependent variable is

skewed, the sample size needed if $f^2 = 0.02$, 0.15 and 0.35 is 401 , 55 and 24 observations respectively

- Sample size for logistic regression model.

Let the simple logistic regression model is log $(P/1-P) = \beta_0 + \beta_1 X$, where Y is the binary response variable (Y = 0 or 1), X is the predictor variable, β_1 is the regression coefficient (is the change in log odds for an increase of one unit of X) and P = prob [Y = 1]. To test $H_0: \beta_1 = 0$ VS $H_1 : \beta_1 = \beta^*$ where $\beta^* \neq 0$. when X is the continuous variable, the sample size required is given by (more details in Hsieh et al 1998).

$$n = \frac{(Z_{\alpha/2} + Z_\beta)^2}{P_1(1-P_1)\beta^*} \qquad \dots 52$$

where β^* is the effect size and P_1 is the event rate at the mean of X.

if X is the binary variable (X = 0 or 1), the sample size is given by :

$$n = \frac{[Z_{\alpha/2}\sqrt{P(1-P)B} + Z_\beta\sqrt{P_1(1-P_1) + P_2(1-P_2)/B}}{(P_1-P_2)^2(1-B)} \qquad \dots 53$$

where P = $(1 - B) P_1 + B P_2$ is the overall event rate, B is the proportion with X = 1, P_1 and P_2 are the events rates at X = 0 and X=1 respectively.

In a multiple logistic regression model, the sample size needed is (approximate)

$$n_p = n / 1 - R^2_{1.23\dots p} \qquad \dots 54$$

30

Where p is the number of predictor variables, $R^2_{1.23...p}$ a multiple determination coefficient relating X_1 with X_2, X_3, …, X_P and n_p and n are the sample sizes required for a logistic regression model with p and 1 variables.

Conclusion:

The calculation of the adequate sample size plays on important role during the design stage of all statistical studies and allows to ensure the validity results of studies. in this study, we discuss that the accuracy of sample size calculations depend on a set of factors including, type of data, effect size, significant level, level of precision, type I error, level of variability, type II error, confidence level, power of test and other considerations. Knowing these factors and their effect help us to determine the sample size needed. Also we discuss and reviews many of procedures that used in calculating the sample size for continuous and categorical data, and procedures for determine the sample size for complex situations (for example one – way ANOVA, Linear regression (simple and multiple), and logistic regression (simple and multiple)). Unnecessary large sample size is a huge waste of the resources; therefore, the objective of the sample size calculations is to select a minimum sample size for achieving a level of precision or desired power of test. In this study, we didn't take the cost factor into account when the sample size is determined. Thus, we

recommend this factor to be taken seriously because of its importance. Finally, whatever the used formula in determining the sample size, the researcher must to be convinced by the data collected for study to obtain satisfactory results.

Reference

1 - Adcock, C.J., (1997), "Sample size determination : A Review", Statistician, Vol. (46), pp. 261 – 283.

2 - Chadha, V.K., (2006), "Sample size determination In Health studies", National Tuberculosis Institute Bulletin, Vol. (42), pp. 55 – 62.

3 - Charan,J. and Biswas,T. (2013) "How to calculate sample size for different study designs", Indian journal of psychological medicine, Vol. (35), pp. 121-126.

4 - Chow, S.C., Shao, J., and Wang, H., (2003), "Sample size calculation in clinical research", Marcel Dekker, New York.

5 - Cochran, W.G., (1977), "Sampling Techniques", 3rd ed. John Wiley and Sons, New York.

6 - David, A.H., and Anthony, R.B., (2004), "Sample size and power calculations using the noncentral t-distribution", The state Journal, Vol. (4), PP. 142 – 153.

7 – Elise W., and Jonathan, B., (2002), "Statistics Review 4 : Sample size calculations", Available on line http://ccforum.com/content/6/4/335.

8 – Eng, J.M., (2003), "Sample size estimation : How Many Individuals should be studied", Radiology, Vol. (277), pp. 309 – 313.

9 – Fleiss, J., Tytun, A., and Ury, H., (1980), "A Simple approximation for calculation sample size for comparing independent proportions", Biometrics, Vol. (36), pp. 343 – 346.

10 - Geoffrey, T.F., (2009), "Practical sample calculations for surveillance and diagnostic investigations", J.Vet.Diagn.Invest., Vol. (21), PP. 3 – 14.

11 - Green, S.B., (1991), "How many subjects does it take to do a regression analysis", Multivariate Behavioral Research, Vol. (26), pp. 499 – 510.

12- Hsieh, F.Y., Daniel, A.B., and Michael, D. L., (1998), "A Simple Method of sample size calculation for linear and logestic regression", Statis. Med., Vol. (17), PP. 1623 – 1634.

13– Israel, G.D., (1992), "Determining sample size", Program Evaluation and organizational development, IFAS, PEOD-6, university of Florida.

14- Janet, W., (2002), "Sample size calculations for Randomized controlled Trials", Epidemiologic Reviews, Vol.(24), PP. 39–53.

15- Jon, K.M., Lars, E.M., and Katie, D.M., (2003), "Considerations on sample size and power calculations in Randomized clinical trails", The Journal of Arthroscopic and Related surgery, Vol. (19), pp. 997-999.

16- Jones, S.R., Carley, S., and Harrison, M., (2003), "An Introduction to power and sample size Estimation", Emerg. Med. J., Vol. (20), pp. 453 – 458.

17- Kasiulevicius, V., Sapoka, V., and Filipaviciute, R., (2006), "Sample size calculation in Epidemiological studies", Gerontologija, Vol. (7), pp.225-231.

18- Ken, K., and Scott, E.M., (2003), "Sample Size for Multiple Regression obtaining Regression Coefficients that are accurate, Not simply significant psychological methods, Vol. (8), pp.305-321.

19- Kenneth, F.S., and David, A.G., (2005), "Sample size calculations in Randomized Trials : Mandatory and Mystical", Lancet, Vol. (365), PP.1348-1353.

20- Ln,J.(2016)," Considerations when calculating the sample size for an inequality test, Anesthesiology, vol 69, pp.327–331.

21- McClave, J.T., Benson, P.G., and Sincich, T., (2007), "Statistics for Business and Economics", 10th ed., upper saddle River (NJ) : Prentice Hall International.

22 - Rahul, M.D., (2007), "Sample size Calculations", Available at www.RavenAnalytics.com/articles.php.

23 - Ralph, B.D., Steve, H., and Rajasekhar, R., (2002), "Sample size Determination", ILAR Journal, Vol. (43), PP. 207 – 213.

24- Russel, V.L., (2001), "Some Practical Guidelines for effective sample size determination", The American Statistician", Vol. (55), PP. 187 – 193.

25- Scott, E.M., (2000), "Sample size and Multiple Regression Analysis", Psychological Method, Vol. (5), pp. 201 – 226.

26- Tabachnick, G.B., and Fidell, S.L., (2001), "Using Multivariate Statistics", 4th Ed. Needham Heights, MA : Allyn and Bacon.

27- Wei, J.L. Huey, M.H., and James, J.C., (2010), "Power and sample size estimation in Microarry studies", Available at http://www.biomedcentral.com/1471/2015/11/48.