**Current Topic**

# IS 30 THE MAGIC NUMBER? ISSUES IN SAMPLE SIZE ESTIMATION

**Sitanshu Sekhar Kar[1], Archana Ramalingam[2]**

[1]Assistant Professor; [2]Post- graduate, Department of Preventive and Social Medicine, JIPMER, Puducherry
**Correspondence** Dr Sitanshu Sekhar Kar, Email: drsitanshukar@gmail.com

## ABSTRACT

Research has become mandatory for career advancement of medical graduates. Researchers are often confounded by issues related to calculation of the required sample size. Various factors like level of significance, power of the study, effect size, precision and variability affect sample size. Also design issues like sampling technique and loss to follow up need to be considered before calculating sample size. Once these are understood, the researcher can estimate the required sample size using softwares like Open Epi. Correct estimation of sample size is important for the internal validity of the study and also prevents unnecessary wastage of resources.

**Keywords**: sample size, estimation, epidemiological studies, Open Epi

## INTRODUCTION

MCI has recently amended, both the 'PG medical education regulations' and the 'minimum qualifications for teachers in medical institutions'.[1,2] These have made active participation in research mandatory, be it for getting a PG degree or for promotions in medical institutions. The statement 'publish or perish' sums up the situation well.

The path towards quality research is not one without hurdles. Many researchers face difficulty in the various steps of conducting a study, starting with framing the research question/hypothesis up to the analysis of data and interpretation of results. In this section we would like to focus on sample size estimation, one of the steps which invariably confound most researchers.

Even though statistical textbooks give formulae for sample size estimation, the wide range of formulae that can be used for specific situations and study designs makes it difficult for most investigators to decide which method to use.[3]

Many a time, questions like "Is there a magic number?", "Is it not okay if I include only 30 subjects in each group?", "How to know whether the number studied is adequate or not?" perplex many researchers.

## NEED FOR SAMPLE SIZE ESTIMATION

*Why at all are we so concerned about sample size?*

Research is always resource intensive. Hence, it is not always possible to study the entire population. So, we conduct the study on a sample and then generalize the results to the study population. In order to do so, our sample should be 'representative' or in other words not different from population.

If our sample size is too small then we may fail to detect what we intended to. On the other hand if we study a large sample then we would unnecessarily invest more resources in the form of manpower, materials, money and minutes (time). Also, we would be subjecting more number of people than required to the adverse effects of drugs.

There are a few principles which govern the estimation of sample size. These are: Level of significance, Power of the study, Effect size, Variability and Precision.

*Level of significance*:

Level of significance (alpha level) is the probability of rejecting the null hypothesis when it is actually true. This type of error in hypothesis testing is called Type I error. Simply put, it is the probability of saying that the drug/intervention has an effect when it actually does not, or getting a positive result in a diagnostic test when in reality the disease is absent i.e. false positive result. We usually set the level of significance at 5% by convention, though 1% and 0.1% are also used by researchers.

When we decrease the level of significance from 5% to 1% we are reducing the chances of committing type I error and to do so we require a larger sample size. So, smaller the level of significance, larger the required sample size (provided other factors are kept constant).

*Power of the study:*

The power of a statistical test is the probability that the test will correctly reject a false null hypothesis. In lay man's terms, it is the probability of detecting the true effect of treatment after administration of a drug/intervention. So, if we choose the power to be 80% then the study will be able to detect a true effect of the drug 80% of times i.e. false negative results will occur only 20% of times. Failure to reject the null hypothesis when it is false is called type II error. Higher the power of the study, greater will be the required sample size.

*Effect Size:*

Effect size provides the magnitude of association between a predictor and an outcome variable. In simple terms, it gives the magnitude of treatment effect of a drug (for example: reduction of mean BP by 2%). Usually the effect size can be found from review of literature, through a pilot study or by asking an expert in the field. To correctly identify small treatment effects, we need a larger sample size.

*Variability:*

Variability indicates the spread of a continuous variable. Usually the variability is measured using standards deviation (SD) or standard error of mean (SEM), the latter being a better measure of variability than the former. When the variability is high, the required sample size is more.

*Precision:*

Precision is a measure of how close our sample estimate is to the true value of a population parameter. It is of two types: absolute or relative. Let us take the prevalence of hypertension in the population as 20%. An absolute precision of 5% means that the prevalence of hypertension in our sample population will be between 15% and 20%. If we take relative precision of 5%, then the prevalence in the sample population will be between 19% and 21 %.( 5% of 20 is 1; hence the prevalence in the sample will be between 19 and 21). By convention the relative precision is taken between 5% and 20%. The closer we need our sample estimate to be to the population mean, the greater should be the size of the sample we use.

Calculations on how, an increase or decrease in each of these principles affect the sample size is beyond the scope of this review. To read further, please refer to: Statistics for the behavioral sciences 8th edition. [4]

## HOW DO WE ESTIMATE THE REQUIRED SAMPLE SIZE?

Do we always require an expert for calculation of sample size? The answer is no, however it is always better to cross check the result from an expert.

The sample size can be estimated from:
1. Statistical packages
2. Formulae and tables from standard books and
3. Nomograms (not used these days).

The formulae for calculation of sample size for common study designs are given in the table 1[5].

The most easy and preferred way of calculating sample size is by using an appropriate statistical package. The popular ones are "OPEN EPI", "Stat Cal" and "STATA".

Open Epi is a free, web-based, open source, operating system-independent series of programs for use in epidemiology, biostatistics, public health, and medicine, providing a number of epidemiologic and statistical tools for summary data. The Open Epi developers have had extensive experience in the development and testing of Epi Info, a program developed by the Centers for Disease Control and Prevention (CDC) and widely used around the world for data entry and analysis. It is freely downloadable from the web address: http://www.**openepi**.com/[6]

The information required for sample size calculation using Open Epi for various study designs is given in Table 2.

**Table 1: Formulae for sample size calculation**

| Type of study | Formula for minimum sample size | Required information |
|---|---|---|
| **Descriptive study: Mean** | $N=[\{Z_{1-\alpha/2}\}^2 s^2]/d^2$ | $Z_{1-\alpha/2}$: Value of normal deviate at considered level of confidence<br>d : Expected absolute allowable error in the mean<br>s: Expected standard deviation of the variable in the group |
| **Descriptive study: Proportion** | $N=Z^2{}_{1-\alpha/2}p(1-p)/d^2$ | $Z_{1-\alpha/2}$: Value of normal deviate at considered level of confidence<br>p : Expected prevalence of the event in the study group<br>d : Expected absolute allowable error in the mean |
| **RCT: Equivalence of two means** | $N= \frac{(Z_{1-\alpha}+ Z_{1-\beta})^2[ v_1+v_2]}{[d-(m_1-m_2)]^2}$ | $Z_{1-\alpha}$: Value of normal deviate at considered level of confidence (one sided)<br>$Z_{1-\beta}$: Value of normal deviate at considered power of study<br>$m_1$: Anticipated mean of the variable in the standard treatment group<br>$m_2$: Anticipated mean of the variable in the new treatment group<br>$v_1$: Anticipated variance of the variable in the standard treatment group<br>$v_2$: Anticipated variance of the variable in new treatment group |
| **Cohort Study** | $N= \{ Z_{1-\alpha/2}\sqrt{[2p'(1-p')]}+ Z_{1-\beta}\sqrt{[p_1(1-p_1)+p_2 (1-p_2)]^2} /(p_1-p_2)^2$ | $Z_{1-\alpha/2}$: Value of normal deviate at considered level of confidence (two sided<br>$Z_{1-\beta}$: Value of normal deviate at considered power of study<br>$p_1$ : Anticipated probability of disease/event in the people exposed to factor of interest<br>$p_2$ : Anticipated probability of disease/event in the people not exposed to factor of interest<br>Anticipated relative risk: RR: $p_1/ p_2$<br>$p'$ : $(p_1- p_2)/2$ |
| **Case control** | $N= \{ Z_{1-\alpha/2}\sqrt{[2 p_2 (1- p_2)]}+ Z_{1-\beta}\sqrt{[p_1(1-p_1)+p_2 (1-p_2)]}\}^2 /(p_1-p_2)^2$ | $p_1$: Anticipated probability of exposure for cases<br>$p_2$: Anticipated probability of exposure for controls<br>Anticipated odds ratio: OR= $[p_1/(1-p_1)]/[ p_2/(1- p_2)]$<br>$Z_{1-\alpha/2}$: Value of normal deviate at considered level of confidence (two sided<br>$Z_{1-\beta}$: Value of normal deviate at considered power of study |

Following are a few examples for calculating sample size using OPEN EPI[7]

*Example 1: Case control study*

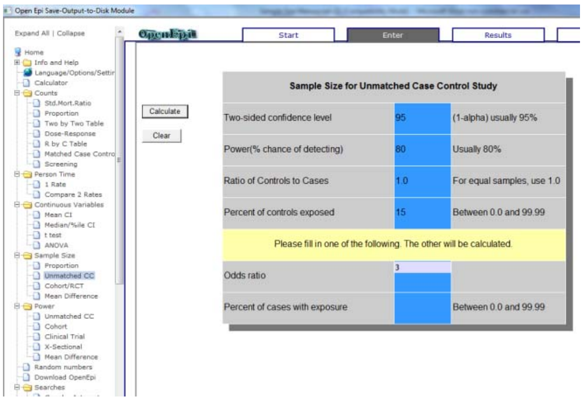
Calculation of the sample size for studying the association of obesity with breast cancer using a hospital based case control design. The list of information required to calculate the sample size is given in table 2.

After literature search let us say that we found the proportion of controls with obesity to be 15% and the odds ratio to be 3. Let us take the level of significance to be 5% and power to be 80%. After inputting these data into Open Epi we can calculate the required sample size (Figures 1 and 2). The sample size for this example comes to 170 (85 in each group) using Open Epi software

**Table: 2. Information required for calculating sample size for various study designs**

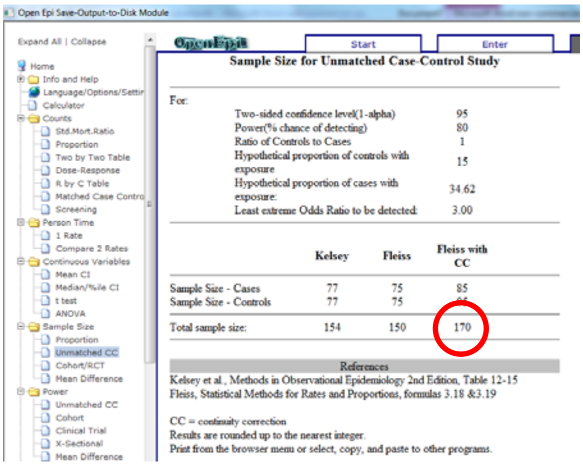
| Descriptive (Prevalence) | Anticipated frequency | Confidence limits | Precision |
|---|---|---|---|
| Case control | Percentage of controls exposed | Percentages of cases exposed/OR | Ratio of cases and controls |
| RCT (Proportion) | Percent of outcome in control group | Percent of outcome in intervention group | Ratio of subjects in control & intervention |
| RCT (Continuous ) | Mean & SD of Control Group | Mean & SD of Intervention Group | Ratio of subjects in control & intervention |

**Figure 1: Sample size calculation for Case control studies using Open Epi : "Enter Data Page"**



| Sample Size for Unmatched Case Control Study | | |
|---|---|---|
| Two-sided confidence level | 95 | (1-alpha) usually 95% |
| Power(% chance of detecting) | 80 | Usually 80% |
| Ratio of Controls to Cases | 1.0 | For equal samples, use 1.0 |
| Percent of controls exposed | 15 | Between 0.0 and 99.99 |
| Please fill in one of the following. The other will be calculated. | | |
| Odds ratio | 3 | |
| Percent of cases with exposure | | Between 0.0 and 99.99 |

**Figure 2: Sample size calculation for Case control studies using Open Epi: Results Page**



Sample Size for Unmatched Case-Control Study

For:

| | |
|---|---|
| Two-sided confidence level(1-alpha) | 95 |
| Power(% chance of detecting) | 80 |
| Ratio of Controls to Cases | 1 |
| Hypothetical proportion of controls with exposure | 15 |
| Hypothetical proportion of cases with exposure: | 34.62 |
| Least extreme Odds Ratio to be detected: | 3.00 |

| | Kelsey | Fleiss | Fleiss with CC |
|---|---|---|---|
| Sample Size - Cases | 77 | 75 | 85 |
| Sample Size - Controls | 77 | 75 | 85 |
| Total sample size: | 154 | 150 | 170 |

References

Kelsey et al., Methods in Observational Epidemiology 2nd Edition, Table 12-15
Fleiss, Statistical Methods for Rates and Proportions, formulas 3.18 &3.19

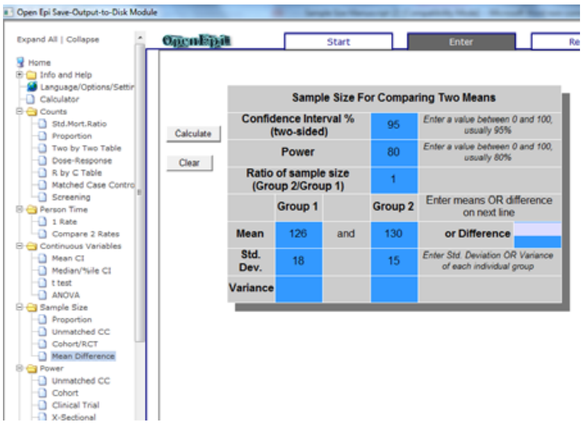
CC = continuity correction
Results are rounded up to the nearest integer.
Print from the browser menu or select, copy, and paste to other programs.

* Circled in red is the required sample size

*Example 2: Randomised control trial*

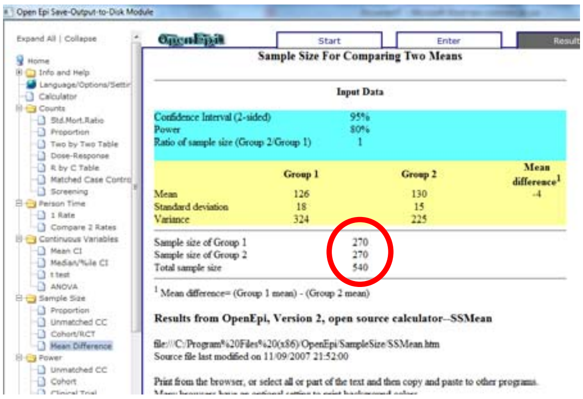
Calculation of sample size for studying the efficacy of Drug 'P' in lowering BP levels using a randomized placebo control trial. Since drug 'P' is a new drug, no data about the required information (as mentioned in table 1) is available. So after doing a pilot study, let us say that we found the mean BP after giving drug 'P' to be 126 mmHg ± 18mm Hg and after giving placebo to be 130mmHg± 15 mmHg. Using this data and keeping the level of significance at 5% and power of the study as 80% the sample size is calculated to be 540 (270 in each group) by using Open Epi software. (Figures 3 and 4)

**Figure 3: Sample size calculation for RCT using Open Epi: "Enter Data Page"**



| Sample Size For Comparing Two Means | | | | |
|---|---|---|---|---|
| Confidence Interval % (two-sided) | 95 | | | *Enter a value between 0 and 100, usually 95%* |
| Power | 80 | | | *Enter a value between 0 and 100, usually 80%* |
| Ratio of sample size (Group 2/Group 1) | 1 | | | |
| | Group 1 | | Group 2 | Enter means OR difference on next line |
| Mean | 126 | and | 130 | or Difference |
| Std. Dev. | 18 | | 15 | *Enter Std. Deviation OR Variance of each individual group* |
| Variance | | | | |

**Figure 4: Sample size calculation for RCT using Open Epi: "Results page"**



Sample Size For Comparing Two Means

Input Data

| | | | |
|---|---|---|---|
| Confidence Interval (2-sided) | 95% | | |
| Power | 80% | | |
| Ratio of sample size (Group 2/Group 1) | 1 | | |

| | Group 1 | Group 2 | Mean difference[1] |
|---|---|---|---|
| Mean | 126 | 130 | -4 |
| Standard deviation | 18 | 15 | |
| Variance | 324 | 225 | |

| | |
|---|---|
| Sample size of Group 1 | 270 |
| Sample size of Group 2 | 270 |
| Total sample size | 540 |

[1] Mean difference= (Group 1 mean) - (Group 2 mean)

**Results from OpenEpi, Version 2, open source calculator--SSMean**

file:///C:/Program%20Files%20(x86)/OpenEpi/SampleSize/SSMean.htm
Source file last modified on 11/09/2007 21:52:00

Print from the browser, or select all or part of the text and then copy and paste to other programs.

*Circled in red is the required sample size

## DO SAMPLE SIZE CALCULATIONS DIFFER BASED ON SAMPLING TECHNIQUES AND ISSUES LIKE LOSS TO FOLLOW UP?

*a) Design Effect:*

All along we have discussed about the sample size required if simple random sampling is followed. However if we use other sampling techniques like cluster sampling or multistage sampling then the required sample size will change as we have to take into account the fact that each member of the sampling frame may not have an equal chance of getting selected. So we multiply the calculated sample size by design effect. Formulae for calculation of design effect can be found in statistics textbooks and beyond the scope of this article, but by convention we take design effect to be between 1.5 and 3.

*b) Adjustments for loss to follow up or non-response:*

Sometimes it may so happen that some of the recruited participants may not continue in the study and are termed as 'loss to follow up'. In other cases some participants may not respond to our questionnaire and they will come under the non-response group. These issues also must be considered while calculating sample size. Let us say that we expect x% of the participants to fall under non-response or loss to follow up category then the required sample size will be:

Adjusted sample size = Unadjusted sample size * (100/ [100-x]) [8]

x= expected percent of loss to follow up /non-response

## CONCLUSION

This paper gives insight into basic principles of sample size estimation. Sample size calculations can be done with the help of statistical soft wares, once the principles behind these are clearly understood. Various factors like level of significance, power, effect size, variability and precision play an important role in determining sample size of a particular study. We must remember that information on these should be gathered by the researcher through literature search, pilot study and consulting experts in the field. Hence, there is no such thing as a magic number when it comes to sample size calculations and arbitrary numbers such as 30 must not be considered as adequate. Also calculation of sample size using Open Epi software is discussed. Several situations like calculation of sample size in matched case control study, diagnostic tests and designs dealing with clustered data are not dealt in this manuscript. It is better to take the help of an expert when one is in doubt or while dealing with complex study designs.

## REFERENCES:

1. Medical council of India Postgraduate medical education regulations, 2000 (amended up to December, 2010) [Internet]. [Cited on 16/2/2012]. Available from: http://www.mciindia.org/rules-and-regulation/Postgraduate-Medical-Education-Regulations-2000.pdf
2. Minimum Qualifications for Teachers in Medical Institutions Regulations, 1998 (amended up to November, 2010). MCI. [Internet]. [Cited on 16/2/2012]. Available from: http://www.mciindia.org/rules-and-regulation/Teachers-Eligibility-Qualifications-Rgulations-1998.pdf
3. Marlies Noordzij, Giovanni Tripepi, Friedo W. Dekker, Carmine Zoccali, Michael W. Tanck,Kitty J. Jager. Sample size calculations: basic principles and common pitfalls. Nephrol Dial Transplant (2010) 25: 1388–1393
4. Frederick J Gravetter, Larry B Wallnau. Statistics for the behavioral sciences 8th edition. Wadsworth, Cengage learning. 2009.
5. Betty R Kirkwood, Jonathan A C Sterne 2nd edition. Massachusetts. Blackwell Science Ltd. 2003. P 420
6. OpenEpi [Internet]. Wikipedia, the free encyclopedia. 2012 [cited 2012 Oct 7]. Available from: http://en.wikipedia.org/w/index.php?title=OpenEpi&oldid=486949640
7. Dean AG, Sullivan KM, Soe MM. OpenEpi: Open Source Epidemiologic Statistics for Public Health, Version 2.3. [Home page from Intenet] [updated 2011/23/06 ; cited on 2012/02/16]. Available from: http://www.openepi.com/OE2.3/Menu/OpenEpiMenu.htm
8. Betty R Kirkwood, Jonathan A C Sterne 2nd edition. Massachusetts. Blackwell Science Ltd. 2003. p423.