MEDICAL STATISTICS

Practical Issues in Calculating the Sample Size for Prevalence Studies

L. Naing[1,2*], T. Winn[2], B.N. Rusli[1,2]

[1]Department of Community Dentistry, School of Dental Sciences, [2]Department of Community Medicine, School of Medical Sciences, Universiti Sains Malaysia, Health Campus, 16150 Kubang Kerian, Kelantan, Malaysia
*Corresponding author: naing@kck.usm.my

ABSTRACT
The sample size calculation for a prevalence only needs a simple formula. However, there are a number of practical issues in selecting values for the parameters required in the formula. Several practical issues are addressed and appropriate recommendations are given. The paper also suggests the application of a software calculator that checks the normal approximation assumption and incorporates finite population correction in the sample size calculation.

Keywords: sample size calculator, prevalence study

INTRODUCTION

"*How big a sample do I require*?" is one of the most frequently asked questions by investigators. Sample size calculation for a study estimating a population prevalence has been shown in many books (Daniel, 1999, Lwanga and Lemeshow, 1991). The aim of the calculation is to determine an adequate sample size to estimate the population prevalence with a good precision. It can be calculated using a simple formula as the calculation needs only a few simple steps. However, the decision to select the appropriate values of parameters required in the formula is not simple in some situations. In this paper, we highlight the problems commonly encountered, and give recommendations to handle these problems.

HOW TO CALCULATE THE SAMPLE SIZE

The following simple formula (Daniel, 1999) can be used:

$$n = \frac{Z^2 P(1-P)}{d^2}$$

where n = sample size,
Z = Z statistic for a level of confidence,
P = expected prevalence or proportion
(in proportion of one; if 20%, P = 0.2), and
d = precision
(in proportion of one; if 5%, d = 0.05).

Z statistic (Z): For the level of confidence of 95%, which is conventional, Z value is 1.96. In these studies, investigators present their results with 95% confidence intervals (CI). Investigators who want to be more confident (say 99%) about their estimates, the value of Z is set at 2.58.

Expected proportion (P): This is the proportion (prevalence) that investigators are going to estimate by the study. Sometimes, investigators feel a bit puzzled and a common response is that '*We don't know this P. That is why we are going to conduct this study*'. We need to understand that the scale of P is from zero to one, and the sample size varies depending on the value of P (Figure 1). Therefore, we have to get an estimate of prevalence (P) in order to calculate the sample size. In many cases, we can get this estimate from previous studies. In this paper, P is in proportion of one, not using a percentage in all formulae. For example, if prevalence is 20%, then P is equal to 0.2.

Precision (d): It is very important for investigators to understand this value well. From the formula, it can be conceived that the sample size varies inversely with the square of the precision (d^2).

At the end of a study, we need to present the prevalence with its 95% confidence interval. For instance, the prevalence in a sample is 40% and 95% CI is 30% to 50%. It means that the study has estimated the population prevalence as between 30% and 50%. Please notice that the precision (d) for this estimate is 10% (i.e. 40% ±10% = 30%~50%). It shows that the width of CI is two times of the precision (width of CI = 2d).

If the width of the CI is wide like in this example (30% to 50%, the width of the interval is 20%), it may be considered as a poor estimate. Most investigators want a narrower CI. To obtain a narrower CI, we need to design a study with a smaller d (good precision or smaller error of

estimate). For instance, if investigators want the width of CI as 10% (0.1), *d* should be set at 0.05. Again, *d* in the formula should be a proportion of one rather than percentage.

PRACTICAL ISSUES IN DETERMINING SAMPLE SIZE PARAMETERS

Determining Precision (*d*)

What is the appropriate precision for prevalence studies? Most of the books or guides show the steps to calculate the sample size but there is no definite recommendation for appropriate *d*. Investigators generally ends up with the ball-park figures of the study sizes usually based on their limitations such as financial resources, time or availability of subjects. However, we should calculate the sample size with a reasonable or acceptable precision and then allowing for other limitations.

In our experience, it is appropriate to have a precision of 5% if the prevalence of the disease is going to be between 10% and 90%. This precision will give the width of 95% CI as 10% (e.g. 30% to 40%, or 60% to 70%). However, when the prevalence is going to be below 10% or more than 90%, the precision of 5% seems to be inappropriate. For example, if the prevalence is 1% (in a rare disease) the precision of 5% is obviously crude and it may cause problems. The obvious problem is that 95% CIs of the estimated prevalence will end up with irrelevant negative lower-bound values or larger than 1 upper bound values as seen in the Table 1.

Therefore, we recommend *d* as a half of *P* if *P* is below 0.1 (10%) and if *P* is above 0.9 (90%), *d* can be {0.5(1-*P*)}. For example, if *P* is 0.04, investigators may use *d*=0.02, and if *P* is 0.98, we recommend *d*=0.01. Figure 1 is plotted with this recommendation. Investigators may also select a smaller precision than what we suggest if they wish.

However, if there is a resource limitation, investigators may use a larger *d*. In case of a preliminary study, investigators may use a larger *d* (e.g. >10%). However, justification for the selection of *d* should be stated clearly (e.g. limitation of resources) in their research proposal so that reviewers will be well informed. In addition, the larger *d* should meet the assumption of normal approximation that we will discuss later.

Estimating *P*

Speculating *P* may be intriguing in practice. The investigator may get several *P*s from the literature. Preferably, *P* from the studies with similar study design and study population from the most recent studies would be most preferable.

If we have a range of *P*, for instance, 20% to 30%, we should use 30% as it will give a larger sample size (Figure 1). If the range is 60% to 80%, we should use 60% as it will give a larger sample size. If the range is 40% to 60%, 50% will give a larger sample size. Macfarlane (1997) also suggested that if there was doubt about the value of *P*, it is best to err towards 50% as it would lead to a larger sample size.

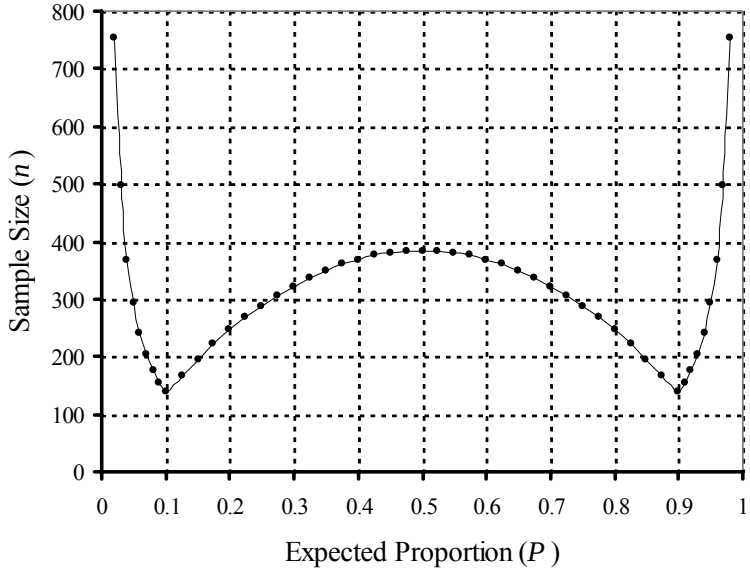

Figure 1 Relationship between Sample Size and Expected Proportion (Prevalence) (Plotted using Microsoft Excel)

Table 1 95% CI of rare diseases (P =< 0.05) and common diseases (P => 0.95) with a precision (d) set at 0.05

P	*n*	95% CI	
		Lower	Upper
0.01	16	-0.04	0.06
0.02	31	-0.03	0.07
0.03	45	-0.02	0.08
0.04	60	-0.01	0.09
0.05	73	0	0.10
0.95	73	0.90	1.00
0.96	60	0.91	1.01
0.97	45	0.92	1.02
0.98	31	0.93	1.03
0.99	16	0.94	1.04

Setting *P*=0.5 does not necessarily provide the biggest sample size

Some books or guides suggest that if it is impossible to come up with a good estimate for P, one may set P equal to 0.5 to yield the maximum sample size (Daniel, 1999, Lwanga and Lemeshow, 1991). In our opinion, this suggestion should be taken with caution. If P is between 10% and 90%, it is a good guide to take P as 0.5 (if it is impossible to make a better estimate) as it will give the biggest sample size. However, if P is quite small (<10%) or very large (>90%), we may need a larger sample size than those calculated using P=0.5 (Figure 1). Our arguments are as follows. Firstly, for instance, calculation was done using P=0.5 and d=0.05 as an investigator could not estimate P, so that the sample size was 385. However, if the real P is unfortunately 1%, we may get, on average, 3 or 4 cases (diseases) from 385 subjects or you may not get any disease case at all. Secondly, with this small number of cases (diseases), the assumption of normal approximation that is used in this sample size calculation is not met. Similarly, if P is too large (e.g. 99%), with the sample size of 385, you may get only a few non-cases (non-diseases) or perhaps none, and again, the normal approximation assumption that is discussed later, may not be met.

In practice, investigators should be a little cautious before applying this 'P=0.5' suggestion. It is not very difficult for an investigator to estimate whether P is below 10%, between 10% and 90% or above 90% with his or her experience. Otherwise, a very crude pilot study (e.g. with a sample size of 20~30) can also easily determine this P. If within 10%-90%, it is safe to apply the 'P=0.5' suggestion.

Assumption of Normal Approximation

The above sample size calculation formula is based on the assumption of normal approximation. It says that nP and $n(1-P)$ must be greater than 5 (Daniel, 1999). In other words, both cases and non-cases in the selected sample must be greater than 5. Small sample sizes might not fulfill this assumption, and we should check this assumption after calculating the sample size. The recommendation that we have made to apply the precision (d) of half of P and $0.5(1-P)$ will also ensure to meet this assumption (Table 2).

For those who wish to know more about the normal approximation assumption, the following is a worked example:

Suppose we wish to estimate a proportion of population (P) who are regular cigarette smokers in a village. We would like our sample estimate (p) to have a high probability of falling between $P-d$ and $P+d$. (Please notice that P is a population proportion and p is a sample estimate). Before calculating the sample size, we must take what seems to be an ironical step. We must speculate what this proportion (P) is. The basic reason for this step is that sample size depends on the standard error (SE) of the distribution of prevalence of smokers (p). As a matter of fact, the sample size calculation formula shown earlier has been derived from the following SE equations.

Table 2 Checking assumption, *nP* and *n*(1-*P*) for calculated sample sizes

P	*d*	*n*	*np*	*n*(1-*P*)
0.01	0.005	1521	15.2	1506.1
0.02	0.010	753	15.0	737.9
0.03	0.015	497	14.9	481.9
0.04	0.020	369	14.8	354.0
0.05	0.025	292	14.6	277.4
0.06	0.030	241	14.4	226.3
0.07	0.035	204	14.3	189.9
0.08	0.040	177	14.1	162.6
0.09	0.045	155	14.0	141.4
0.10	0.050	138	13.8	124.5
0.20	0.050	246	49.2	196.7
0.30	0.050	323	96.8	225.9
0.40	0.050	369	147.5	221.3
0.50	0.050	384	192.1	192.1
0.60	0.050	369	221.3	147.5
0.70	0.050	323	225.9	96.8
0.80	0.050	246	196.7	49.2
0.90	0.050	138	124.5	13.8
0.91	0.045	155	141.4	14.0
0.92	0.040	177	162.6	14.1
0.93	0.035	204	189.9	14.3
0.94	0.030	241	226.3	14.4
0.95	0.025	292	277.4	14.6
0.96	0.020	369	354.0	14.8
0.97	0.015	497	482.0	14.9
0.98	0.010	753	737.9	15.1
0.99	0.005	1521	1506.1	15.2

$d = Z \text{ x } \text{SE}_{(p)}$ and

$\text{SE}_{(p)} = \sqrt{\frac{P(1-P)}{n}}$, and therefore

$d = Z\sqrt{\frac{P(1-P)}{n}}$.

If we judge the proportion of smokers in the village (*P*) to be 0.3, *d* as 0.05, and *Z* as 1.96 then,

$$0.05 = 1.96\sqrt{\frac{0.3 \text{ x } 0.7}{n}}$$

The final step in our derivation expresses the prior opinion that we would like 1.96 standard errors of our estimate equal 0.05. If $1.96*\text{SE}_{(p)}$ is equal to 0.05, then our sample estimate has a 95% chance of being within five percentage points of the true (*P*). This assumes that *n* is large enough to make distribution of all possible *p* approximately normal. By reading off the row in Table 2 at *P*=0.3 (30%) and *d*=0.05, the sample size *n* is 323. It means that if we take a random sample of 323 villagers and measure smoking prevalence using a survey questionnaire, we would get a smoking prevalence (*p*) as low as 25% or as high as 35% (30%±5%). The idea behind this example is that if we repeat this survey 100 times, hundred different values of smoking prevalence (*p*) will be obtained. Out of these 100 values, 95 of them would take the values between 25% and 35%. All these values (*p*) are approximately normally distributed (Figure 2). This example also illustrates why do we need to have the values of the parameters (*Z*, *P* and *d*) before we start calculating the sample size.

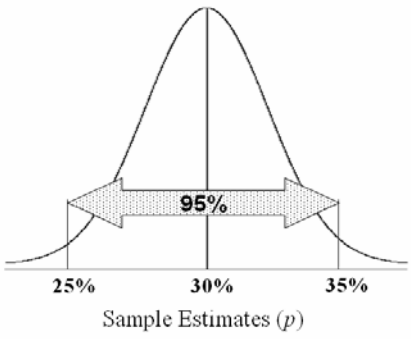

Figure 2 Normally distributed sample estimates (*p*)

Finite Population Correction

The above sample size formula is valid if the calculated sample size is smaller than or equal to 5% of the population size ($n/N \leq 0.05$) (Daniel, 1999). If this proportion is larger than 5% ($n/N > 0.05$), we need to use the formula with finite population correction (Daniel, 1999) as follows.

$$n' = \frac{NZ^2P(1-P)}{d^2(N-1) + Z^2P(1-P)}$$

where

n' = sample size with finite population correction,
N = Population size,
Z = Z statistic for a level of confidence,
P = Expected proportion (in proportion of one), and
d = Precision (in proportion of one).

Cluster or Multistage Sampling

The above sample size formulae are valid only if we apply the simple random or systematic random sampling methods. Cluster or multistage sampling methods require a larger sample size to achieve the same precision. Therefore, the calculated sample size using the above formulae need to be multiplied by the design effect (*deff*) (Cochran, 1977). For example, in immunization coverage cluster surveys, the design effect has been found to be approximately two (Macfarlane, 1997). This means that such cluster sampling requires double the sample size of above calculation.

However, in practice, investigators rarely report their design effects in the literature. What one may do is to contact the authors who have published these articles and request for their design effect. We strongly recommend reporting the design effect if investigators apply cluster or multistage sampling method in their studies. If the design effect is not available at the end, a pilot study can be done to estimate the design effect. Normally, the cluster or multistage sampling is applied in large-scale surveys, and it is worth to conduct a pilot study for several reasons at the first stage. Investigators should consult a statistician before conducting such a study.

Table 3 Gain in precision (error reduction) by increasing the sample size while Z (1.96) and P (0.5) remain constant

n	*d*	% Gain in the precision
97	0.100	-
194	0.070	30.0
291	0.057	43.0
388	0.050	50.0

"*The larger the sample size the better the study*" is not always true

One of the aims of applying appropriate sample size calculation formula is not to obtain the biggest sample size ever. The aim is to get an optimum or adequate sample size. Unnecessarily large sample is not cost-effective. In some circumstances it is unethical. In drug trials, for instance, a very large sample would lead to a conclusion that the new drug is significantly better that an old drug in the statistical sense although the difference may be clinically insignificant. If one examines Table 3 carefully, a two-fold increase in sample size improves the precision by 30% (i.e. reducing the error of estimate by 30%). After the sample size is quadrupled, the precision becomes halved (i.e. reducing the error of estimate by 50%).

Other Objectives of the Study

This article describes and discusses on the sample size calculation for estimating a prevalence as one of the study objectives. Most studies, however, have more than one objectives. It is therefore recommended to also calculate the sample size required for other study objectives appropriately, and the biggest sample size obtained out of all calculations should be taken as the sample size that would accommodate all study objectives.

Anticipating Non-Response or Missing Data

The calculated sample size is for the desired precision or CI width assuming that there is no problem with non-response or missing values. If this is the case, the investigators will not achieve the desired precision. Therefore, it is wise to oversample by 10% to 20% of the computed number required depending on how much the investigators would anticipate these discrepancies.

Role of Statisticians and Reviewers

In this calculation, statisticians may assist investigators in order to ensure correct application of formulae, considering for the assumption and finite population correction. The appropriate precision must be well understood and decided by the investigators with the up-to-date knowledge of the research area that they are going to study. Similarly, specialists (in the study area) in research committees and reviewers of the proposals must have a good understanding of precision and critically look into it to ensure that it is worth to conduct the study with the precision proposed by the investigators.

A Helpful Calculator

In many situations, investigators need to calculate repeatedly, ensure the assumption, and check the necessity of applying the finite population correction. The authors have developed a calculator (Naing *et al*. 2006) using Microsoft Excel and it can be downloaded freely from http://www.kck.usm.my/ppsg/stats_resources.htm. The calculator is designed to give the sample size for various precisions (error of estimate) with or without finite population correction, and also will suggest the need to apply the finite population correction. It will also determine whether the normal approximation assumption is met or not.

CONCLUSION

The paper highlights a number of practical issues in making decision for the parameters applied in the sample size calculation formulae. Investigators may use precision 0.05 if P is between 0.1 and 0.9. However, a smaller d should be applied if $P<0.1$ or $P>0.9$. A range of P should be obtained and apply the P which gives the biggest sample size. Investigators should be cautioned that setting $P=0.5$ doesn't necessarily provide the biggest sample size. In addition, the calculation should consider the assumption of normal approximation, finite population correction and sampling method. Using a calculator considering these issues will be helpful in this sample size calculation.

REFERENCES

Cochran WG (1977). Sampling Techniques, 3rd edition. New York: John Wiley & Sons.

Daniel WW (1999). Biostatistics: A Foundation for Analysis in the Health Sciences. 7th edition. New York: John Wiley & Sons.

Lwanga SK and Lemeshow S (1991). Sample Size Determination in Health Studies: A Practical Manual. Geneva: World Health Organization.

Macfarlane SB (1997). Conducting a Descriptive Survey: 2. Choosing a Sampling Strategy. *Trop Doct*, **27**(1): 14-21.

Naing L, Winn T and Rusli BN (2006). Sample Size Calculator for Prevalence Studies. Available at: http://www.kck.usm.my/ppsg/stats_resources.htm