Review Article

Sample size estimation in epidemiologic studies

Karimollah Hajian-Tilaki (PhD) [*1]

1- Department of Social Medicine and Health, Babol University of Medical Sciences, Babol, Iran.

*** Correspondence:**
Karimollah Hajian- Tilaki, Department of Social Medicine and Health, Babol University of Medical Sciences, Babol, Iran.

E-mail: drhajian@yahoo.com
Tel: 0098 111 2199936
Fax: 0098 111 2199936
Postal Code: 47176-47745

Received: 26 June 2011
Revised: 25 Aug 2011
Accepted: 10 Sep 2011

Abstract

This review basically provided a conceptual framework for sample size calculation in epidemiologic studies with various designs and outcomes. The formula requirement of sample size was drawn based on statistical principles for both descriptive and comparative studies. The required sample size was estimated and presented graphically with different effect sizes and power of statistical test at 95% confidence level. This would help the clinicians to decide and ascertain a suitable sample size in research protocol in order to detect an effect of interest.

Key words: Sample size, Comparative studies, Continuous outcome, Binary outcome, Effect size, Statistical power.

Caspian J Intern Med 2011; 2(4): 289-298

In designing epidemiologic studies, sample size calculation has an important role to detect an effect and to achieve a desired precision in estimates of parameter of interest (1-4). It is a key step that needs to be considered early in a study protocol (2). This particularly helps investigators to devote budget and resources for study. A small sample size will not provide a precise estimate and reliable answers to study hypothesis (5). On the other hand, a large sample size makes difficulty in study management thus, wasting both time and resources (2). Most journals and funding agencies now require a justification for sample size enrolled into a study and investigators must present the principles of sample size calculation to justify these numbers (4).

The clinical researchers frequently ask how many subjects are really needed for a study. Calculations for answering this question are not obviously appealing and sometimes the determination of adequate sample size is mysterious for clinicians. Conceptually, the four major determinants of sample size are: i) the magnitude of effect of interest to be detected in comparative studies or the degree of marginal error of estimate in descriptive design; it is intuitively obvious that if one wishes to detect a small effect size, a higher sample size is needed; ii) the variation (i.e. standard error) of study outcome; with higher variation, a greater sample size is required; iii) confidence level; a higher confidence level in detecting a desired effect, a greater sample size should be included into the study; the confidence level is usually fixed at 95%; iv) study power; given a desired effect size to be detected with a confidence level (e.g. 95%), and how much power is needed (4-5). If one wishes to have more power for statistical test to detect a desired difference, a higher sample size is required. In addition, the width of confidence interval is inversely associated to the number of subjects studied, the more subjects we study, and the more precise we will get about where the true parameter of population lies (4-8). Thus, how many subjects do we need to study in order to get an estimate as close as true value of parameter of interest? In clinical trials, a similar question is raised on how many subjects are needed to be treated in order to get a clinical useful treatment effect (9-12). Despite the conceptual framework of these four essential elements of sample size calculation, the formula for sample size calculation may vary in different study designs with different outcomes and hypotheses.

Thus, the researchers and clinicians may have trouble for sample size calculation in their studies. This review was provided conceptually for clinicians who have background of 2 credits of biostatistics course in their academic curriculum. In this article, the critical elements of sample size were discussed conceptually and the formula for sample size calculation was drawn and classified with respect to the study design and outcome of interest. An illustration was provided of how and why these equations are derived. We also calculated sample sizes for various effect sizes for some conditions and then, the required sample sizes were tabulated and presented graphically.

Sample size in estimating the mean of continuous outcome- descriptive studies

Suppose an investigator may wish to estimate the mean of continuous outcome in a population in which the marginal error in estimates (i.e. the difference between true parameter and its estimate) does not exceed from pre-determine value of d. This marginal error is sometimes called as precision of estimate. For example, if we denote μ as parameter of mean of study population and $\bar{X}$ denotes its estimate, then with maximum marginal error of d in estimate, we have $|\bar{X}-\mu| \leq d$. Based on confidence interval for mean (13-15), we have $|\bar{X} - \mu| = \frac{Z_{\alpha/2}\,\sigma}{\sqrt{n}} \leq d$. Finding the necessary sample size requires solving this equation with respect to n. The result is

$n= \frac{Z^2_{\frac{\alpha}{2}}\sigma^2}{d^2}$ where for α=0.05 (i.e. 95% confidence level) and $Z_{\frac{\alpha}{2}}$=1.96 for two side tests; σ^2 represents the variance of continuous outcome (σ is the standard deviation (SD). Notice that the large value of σ yields n to be large as does a small value of d. The use of this formula is required that an estimate of SD of study outcome is provided. One possibility is to carry out a pilot study and use the resulting sample SD. Another possibility is to use published data or even simply to make guess about the value of SD which might result and to be used in calculating n. For a population distribution that is not too skewed (13), dividing the normal range by 4 gives a rough idea of the SD ($\sigma = \frac{Range}{4}$). As it is intuitively appealing, the sample size has inversely related with square of marginal errors (i.e. d^2). If the marginal error becomes half, the sample size increases 4 times. It has also a positive association with variation of study outcomes (i.e. σ^2) and the square of Z score for confidence level as well. For example, if a researcher knows the normal range of cholesterol in population as 180-220 mg/dl and he wants to estimate the mean of cholesterol in population while the marginal error in estimate does not exceed than 1 mg/dl, then the required sample size is as follows:

$$\sigma= \frac{40}{4}=10. \ n=\frac{1.96^2}{1^2} \times 10^2=385$$

One could consider $\delta= \frac{d}{\delta}$ as an effect size; then

$$n= \frac{Z^2_{\frac{\alpha}{2}}}{\delta^2} =\frac{1.96^2}{\delta^2}$$

Sample size in estimating the proportion of binary outcome- descriptive study

A similar discussion is relevant for sample size calculation in estimating the prevalence (or proportion) of a binary outcome in population. Again, assuming a study objective was focused on estimating the prevalence (or proportion) of an outcome while the marginal error of estimate does not exceed from a pre -determine value of d with 95% confidence level. Let P denote the proportion of interest in population and $\hat{P}$ denotes its estimates. The investigator constrains the marginal error as $|\hat{P} - P| \leq d$. Then based on (1-α) % confidence interval for parameter P (13-15), the result is

$$Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \leq d$$

By solving this equation with respect to n, the required sample size is as follows:

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

Where $\hat{P}(1-\hat{P})$ is a variance component of binary outcome and it is substituted as σ^2 (13). Unfortunately, the use of this formula requires a pre ascertained value of $\hat{P}$. One possibility is to use a similar published data or to carry out a pilot study. Also, a pre determined value of maximum marginal error (d) in estimate should be allocated by investigator from clinical and statistical judgment. In fact, the degree of this error depends on magnitude of proportion (or prevalence), for example by statistical judgment, d should not exceed from a quarter of P ($d \leq \frac{1}{4}P$).

For example, a researcher wants to estimate the required sample size for prevalence of hyperlipidemia in population in which the marginal error of its estimate does not exceed than 2% with 95% confidence level. If the previous published data showed that this rate was 20%. Then, the sample size is calculated as

$$n=\frac{1.96^2 \times 0.20 \times 0.8}{0.02^2}=1537$$

For unknown P, a conservative solution is replacing P=0.5, then the maximum sample size is estimated with a given marginal error and confidence level since P (1-P) never exceeded than 0.25. This conservative estimate is only recommended for a common outcome. Thus

$$n=\frac{Z^2_{\frac{\alpha}{2}} \; 0.5 \times 0.5}{d^2}=\frac{Z^2_{\frac{\alpha}{2}}}{4d^2} \approx \frac{1}{d^2}$$

Sample size for comparative studies: testing population mean with fixed value

In comparative study, a researcher wants to detect a specific effect that is measured by difference in mean of a population and a fixed hypothesized value (μ_1-μ_0) with (1-α) % (e.g. 95%) confidence level and (1-β) % power of statistical test. In testing the null hypothesis H_0: $\mu=\mu_0$ versus an alternative hypothesis H_1: $\mu \neq \mu_0$ (e.g. $\mu=\mu_1$), the two type errors may occur in making decision. The type 1 error occurs when Ho is rejected while H_0 is really true. We denote this type of error as α (e.g. α=0.05) and 1-α (e.g. 95%) is called confidence level. The other error is called type 2 error that occurs when we accept H_0 while H_1 is really true and it is denoted by β and 1-β is called statistical power; it means the probability of rejecting H_0 when H_0 is not true (In fact, H_1 is true, i.e. $\mu=\mu_1$). A widely used convention for acceptable levels of power is 80% (i.e. β=0.20). Conceptually, this means assuming the null hypothesis is true, the researcher has 80% chance of finding statistical significant differences. On the other hand, indicating that the researcher has only 20% chance of failing to find significant differences, if in fact they exist. If we have low power (i.e. high β error), then an effect that is ascertained by H_1 (i.e. μ_1-μ_0) can not be detected. Thus, for the detection of a specific effect by statistical test, in addition to type 1 error, type 2 errors (its compliment denotes as power) also threats study findings. These two types of errors simultaneously influence the required sample size.

Under the H_0, Z-score is defined as $Z_{\frac{\alpha}{2}}=\frac{\bar{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$ and thus, $Z^2_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}=\bar{X}-\mu_0$

While under H_1, Z-score is $Z_{1-\beta}=\frac{\bar{X}-\mu_1}{\frac{\sigma}{\sqrt{n}}}$ and since $Z_{1-\beta}=-Z_\beta$; thus, $-Z_\beta \frac{\sigma}{\sqrt{n}}=\bar{X}-\mu_1$

By subtracting the two above equations, one can easily derive it as

$$\frac{\sigma}{\sqrt{n}}(Z_{\frac{\alpha}{2}}+Z_\beta)=\mu_1-\mu_0$$

By solving the above equation with respect to n, thus

$$n=\frac{(Z_{\frac{\alpha}{2}}+Z_\beta)^2 \times \sigma^2}{(\mu_1-\mu_0)^2}$$

We consider $\delta=\frac{\mu_1-\mu_0}{\sigma}$ as an effect size. Thus

$$n=\frac{(Z_{\frac{\alpha}{2}}+Z_\beta)^2}{\delta^2}$$

Where $\alpha=0.05, Z_{\frac{\alpha}{2}}=1.96$ and $\beta=0.20, Z_\beta$=0.84.

Sample size for comparative study of two population means- continuous outcome

In hypothesis, testing of two populations means H_0: $\mu_1=\mu_2$ versus H_1: $\mu_1 \neq \mu_2$ (i.e. $\mu_2-\mu_1 \neq 0$), as we explained previously, the Z-score under two alternative hypotheses is as follows:

Under H0: $Z_{\frac{\alpha}{2}}=\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{\sigma^2_1}{n_1}+\frac{\sigma^2_2}{n_2}}}$

Under H1: $-Z_\beta=\frac{\bar{X}_1-\bar{X}_2-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma^2_1}{n_1}+\frac{\sigma^2_2}{n_2}}}$

Let us assume $\sigma_1=\sigma_2$ and $n_1=n_2$ =n then by subtracting the two above equations and solving with respect to n, the required sample size for each group would be drawn as follows:

$$n=\frac{2(Z_{\frac{\alpha}{2}}+Z_\beta)^2 \times \sigma^2}{(\mu_1-\mu_2)^2}$$

where $Z_{\frac{\alpha}{2}}$=1.96 (α=0.05), Z_β=0.84 (β=0.20)

One would consider $\delta=\frac{\mu_1-\mu_2}{\sigma}$ as an effect size of interest. Thus, for detecting the difference of two population means with effect size of δ and 95% confidence level and

80% power, the required sample size for each group are as follows:

$$n=\frac{2(1.96+0.84)^2}{\delta^2}$$

Sample size for testing proportion with fixed value-binary outcome

Let us suppose an investigator wishes to test a proportion of binary outcome in a single population with a fixed value: Testing hypothesis H_0: $P=P_0$ versus H_1: $P\neq P_0$ (i.e. $P=P_1$). This investigator needs to know how many sample sizes should be included in the study. For this question, he should answer first to the four determinants of sample size: 1) type 1 error under H_0, 2) study power i.e. the probability of rejecting H_o when a real difference exists (i.e. H_1 is true); 3) the magnitude of difference (P_1-P_0) should be apparent as real difference (or significant); 4) an estimate of SD of binary outcome that is determined under H_0: P=P0 and it is formulated by P_0 (1-P_0). With similar principles as discussed for continuous outcome, the Z-score under the H_0 and H_1 can easily be written and the required sample size is drawn as follows:

$$n=\frac{[Z_{\frac{\alpha}{2}}\sqrt{P_0(1-P_0)}+Z_\beta\sqrt{P_1(1-P_1)}]^2}{[P_1-P_0]^2}$$

Sample size for comparative study of two population proportions- binary outcome

Suppose a researcher has a plan to compare two proportions using two independent samples from two populations. For testing hypothesis Ho: P_1=P_2 versus H1: $P_1\neq P_2$, for detecting a specific effect P_1-P_2 with 80% power and 95% confidence level, he wants to know how many samples should be recruited in the study. Let n be the sample size for each group and P_1 denotes the proportion in study group and P_2 for control, a similar analogy and principle as discussed above can be applied for comparison of two proportions. Z-score under H0 and H1 can be written easily and then the formula for sample size drawn is as follows:

$$n=\frac{[Z_{\frac{\alpha}{2}}\sqrt{2\bar{P}(1-\bar{P})}+Z_\beta\sqrt{P_1(1-P_1)+P_2(1-P_2)}]^2}{[P_1-P_2]^2}$$ where

$$\bar{P}=\frac{P_1+P_2}{2}$$

A formula that is simpler than above, and for practical purposes an approximately equivalent sample size of each group is given by

$$n=\frac{2\bar{P}(1-\bar{P})[Z_{\frac{\alpha}{2}}+Z_\beta]^2}{[P_1-P_2]^2}$$

In analogy to comparison of two means, the effect size (or standard difference) for comparison of two proportions is

$$\delta=\frac{P_1-P_2}{\sqrt{\bar{P}(1-\bar{P})}}$$

Thus $n=\frac{2(Z_{\frac{\alpha}{2}}+Z_\beta)^2}{\delta^2}$

Sample size and study design

In cross sectional studies, P_1 and P_2 represent the prevalence of outcome of interest in two populations. For example, P_1 and P_2 are the prevalence of hypertension in obese and non-obese populations, respectively. In prospective study (either cohort or clinical trials), P_1 and P_2 are the risk of developing an outcome (or incidence rate) in study group (exposed) and control group (non-exposed), respectively. While in case-control design, P_1 and P_2 are the proportion of exposure among cases and control, respectively. An investigator may have no idea about the proportion of exposure among cases in retrospective studies or the risk of developing an outcome among exposed in prospective studies while he knows about these risks in control group. It is possible to estimate P_2 using P_1 (the proportion of exposure in control group or the risk of outcome in non-exposed group) and odds ratio (OR) or risk ratio (RR). These two latter indexes are the measures of association between exposure and outcome in case control and prospective studies, respectively and the researchers may know P_1 and the estimates of these effect measures from literature. Then P_2 can be estimated using the following formula (16):

$$P_2=\frac{P_1\times OR}{1+P_1(OR-1)}$$

Or

$$P_2=\frac{P_1\times RR}{1+P_1(RR-1)}$$

OR is the odds ratio determined in the case-control study and RR is the risk ratio in cohort study or clinical trial. Thus, by computing P_2 using the above formula, one can apply the formula for comparison of two proportions to calculate the sample size for binary outcome either in cross- sectional study or case control / prospective study. Sample size calculation has been extended in regression and correlation analysis (6). This has been paid a little attention for practical purpose in clinical researches. A similar strategy can also be applied to derive sample size formula for diagnostic studies in the analysis of receiver operating characteristic (ROC) curve in estimating and testing of diagnostic accuracy for single modality and comparative study of two modalities (17-18). This is behind the scope of this article and for more detailed information; the interested readers are referred to some published articles (19-23).

Results of sample size calculation

We calculated the sample size for different combinations of effect size and power using excel software. The calculated sample sizes were presented in tables 1, 2, 3 and figures 1, 2, 3 as well.

Table 1. The calculated sample sizes for different combination of prevalence and the maximum marginal errors in estimating prevalence rate (or proportion)

Maximum marginal errors

Prevalence	d=1/4 P	d=1/5 P	d=1/6 P	d=1/7 P	d=1/8 P	d=1/9 P	d=1/10 P
0.001	61404	95944	138159	188050	245617	310858	383776
0.005	12232	19112	27521	37459	48927	61923	76448
0.01	6085	9508	13691	18636	24340	30806	38032
0.02	3012	4706	6777	9224	12047	15247	18824
0.03	1987	3105	4472	6086	7950	10061	12421
0.04	1475	2305	3319	4518	5901	7468	9220
0.05	1168	1825	2628	3577	4671	5912	7299
0.06	963	1505	2167	2949	3852	4875	6019
0.07	817	1276	1837	2501	3266	4134	5104
0.08	707	1104	1590	2165	2827	3578	4418
0.09	621	971	1398	1903	2486	3146	3884
0.10	553	864	1245	1694	2213	2801	3457
0.12	451	704	1014	1380	1803	2282	2817
0.14	378	590	850	1156	1510	1911	2360
0.16	323	504	726	988	1291	1634	2017
0.18	280	438	630	858	1120	1418	1750
0.20	246	384	553	753	983	1245	1537
0.22	218	341	490	667	872	1103	1362
0.24	195	304	438	596	779	985	1217
0.26	175	273	394	536	700	886	1093
0.28	158	247	356	484	632	800	988
0.30	143	224	323	439	574	726	896
0.32	131	204	294	400	522	661	816
0.34	119	186	268	365	477	604	746
0.36	109	171	246	335	437	553	683
0.38	100	157	226	307	401	508	627
0.40	92	144	207	282	369	467	576

Table 2. The required sample size for estimation and comparative study of mean with respect to effect size and power

Effect Size	n for estimation of mean	n for each group in comparative study	
		80% power	90% power
0.01	38416	156800	209952
0.02	9604	39200	52488
0.03	4268	17422	23328
0.04	2401	9800	13122
0.05	1537	6272	8398
0.06	1067	4356	5832
0.07	784	3200	4285
0.08	600	2450	3281
0.09	474	1936	2592
0.1	384	1568	2100
0.2	96	392	525
0.3	43	174	233
0.4	24	98	131
0.5	15	63	84
0.6	11	44	58
0.7	8	32	43
0.8	6	25	33
0.9	5	19	26
1	4	16	21

Table 3. The calculated sample size for each group with respect to risk ratio - RR (or OR) and P_1 ├ with 95% confidence level and 80% power

RR	n	n	n	n	n	n	n
or OR	P_1=0.05	P_1=0.10	P_1=0.25	P_1=0.5	P_1=0.75	P_1=0.90	P_1=0.95
1.5	1687	910	465	387	570	1258	2429
2	515	282	152	136	214	491	960
2.5	271	151	84	80	133	313	618
3	176	99	58	58	99	239	474
3.5	129	73	44	46	81	199	396
4	100	58	36	38	70	174	348
4.5	82	48	30	34	63	157	315
5	69	41	26	30	58	145	292
5.5	60	35	23	28	54	136	274
6	53	31	21	26	50	129	261
6.5	47	28	19	24	48	123	250
7	43	26	18	23	46	119	241
7.5	39	24	17	22	44	115	233
8	36	22	16	21	43	112	227
8.5	33	21	15	20	42	109	221
9	31	19	14	19	41	107	217
9.5	29	18	14	19	40	104	213
10	27	17	13	18	39	103	209

├ P_1 represents the proportion for reference group

Table 1 shows that the sample size substantially varies with respect to marginal errors and prevalence rate in estimating purpose. For a rare outcome that its rate is about 1 per 1000, a very high sample size is required and it strongly depends on the marginal error and varies from 61404 to 383776 for moderate ($d\leq\frac{1}{4}P$) and very low marginal error ($d\leq\frac{1}{10}P$) respectively. As prevalence rate increases to 10%, the requirement sample size substantially decreases to 553 for $d\leq\frac{1}{4}$ P and 3457 for d $\leq\frac{1}{10}$P. For a more common outcome (P=30%), the sample size was calculated as 143 and 896 for moderate and very low marginal errors, respectively. figure 1 also shows the variation of sample size with respect to maximum marginal error of estimate. The second column of table 2 represents the required sample size for purpose in estimating the mean of continuous outcome in a single population with respect to effect size. The more reasonable effect size based on clinical and statistical judgment is about 0.1 that yielded sample size of 384 while the least effect size (0.01) corresponded to very low marginal errors, increased sample size substantially is not of practical convenience. In contrast, high effect size produced very small sample size and the results imprecise estimate. The columns 3 and 4 represents sample size in comparative study of two modalities for comparison of two means with different statistical powers. As expected, in a comparative study, the testing hypotheses H_0: $\mu_1=\mu_2$ versus H_1: $\mu_1\neq\mu_2$, given a similar effect size, a greater sample size of each group is needed compared with estimating purpose. For moderate effect size 0.2-0.4, the required sample size varies from 96 to 382 for 80% power and from 130 to 525 for 90% power. Obviously, for testing hypotheses of single modality with fixed value: H_0: $\mu=\mu_0$ versus H_1: $\mu\neq\mu_0$, the calculated sample size becomes half of those shown in column 3 and 4 (it was not shown in table 2). Table 3 represents the calculated sample size with respect to effect measure of RR and OR is used in cohort studies/ clinical trials and case-control studies respectively. As a general rule, for a given power (80%) and confidence level (95%), in detecting a lower effect measure (RR<2), a greater sample size is required. In addition, the sample size is influenced by the rate of outcome (prospective study) or rate of exposure (case-control study) in reference group (P_1). For a given effect measure, a higher sample size is required for P_1=0.95. With respect to various P_1 examined, a smaller sample size was calculated for P_1=0.5 when other factors were fixed.

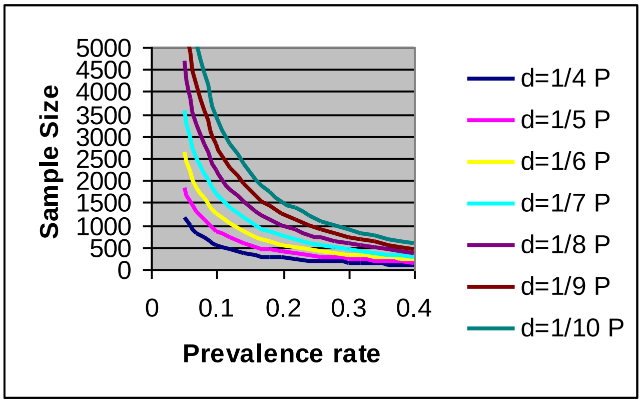

Figure 1. Sample size with respect to prevalence rate and maximum marginal error for estimation of prevalence

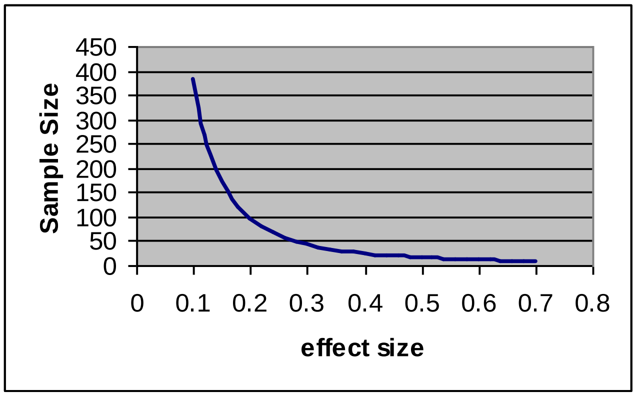

Figure 2. Sample size with respect to effect size (d/σ) for estimation of mean

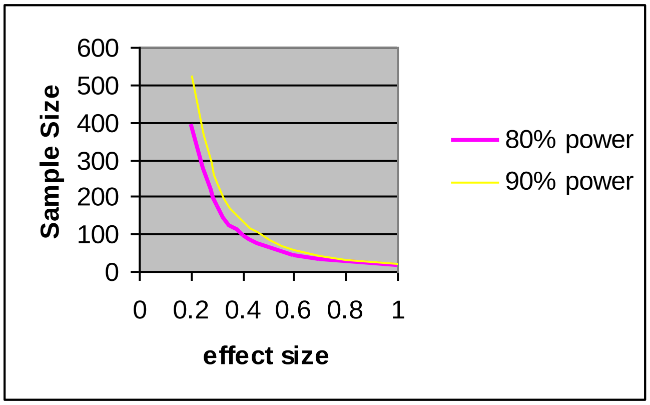

Figure 3. Sample size with respect to effect size (µ1- µ2/σ) and power in comparative study of two population means

Discussion

We calculated the required sample size with different magnitudes of effect size and the power of statistical test. The results show that detecting a small effect size, a higher sample size is needed for a given statistical power. In particular, for an effect size of 0.01, the sample size increases dramatically that is inconvenient for practical purpose of sampling and data management. It seems that moderate effect size of 0.1- 0.4 is more reasonable from practical conveniences for clinical trials that yields an intermediate level of sample size. The effect size of near to 1 is very high and close to 0.01 is very low for the purpose of sample size calculation. Thus, it is not recommended to detect very low effect or to allocate very low marginal error in estimates for sample size calculation because of practical inconvenience, although, such calculated sample sizes result to a more precise estimates and high statistical power. Obviously, for a given effect size, a higher statistical power yields greater sample size. On the other hand, with a small sample size, even a real effect exists in comparative study; the statistical test has low power to detect it (24-25). Thus, conceptually, one should ascertain enough sample size in order to detect a given effect with a given power and confidence level. Alternatively, one may calculate the power of statistical test for a given effect size and allocated sample size as well in order to be confident that the sample size is good enough for the investigation of hypotheses under study.

As we have shown, for detecting a small difference of effect (e.g. the difference in means of two populations or two proportions), the sample size increases dramatically for a given power and confidence level. Alternatively, for detecting a large effect, we do not need a large sample size but if such expected effect was not revealed in study findings, the estimated sample size has low power to detect the difference that was apparent in study. In particular, type 2 error (β) occurs when study suffers from lack of enough sample size (24). This error is the probability of accepting H0 when alternative hypothesis (H_1) is in fact true and the compliment of β (i.e. 1-β) is the statistical power i.e. the probability of rejecting H_0 when H_0 is not true (i.e. H_1 is true). We used two side tests for sample size calculation instead of one side test since two side tests are most often employed in medical and behavioral research. The use of one side test remains controversial for some statisticians arguing their use should never be applied when it is in doubt. Nevertheless, in doubtless instances, the unidirectional hypothesis may be appropriate (1). Furthermore, for a given sample size and effect size, a greater statistical power can be achieved in detecting effect size by employing one side test.

A comparison of sample size for three different scenarios: estimation, testing the mean of single population with fixed value and a comparative study of two modalities, as expected, the latter condition results to a larger sample size for a given effect size and confidence level since the sample size formula includes the variation of two populations. Generally, the required sample size is larger for testing purposes versus estimating, since both type 1 and type 2 errors simultaneously are incorporated in sample size calculation for testing.

In some senses, it seems that the sample size may depend on a study design for binary outcome. While the proportions P_1 and P_2 measure the rate of study outcome for prospective study, they determine the rate of exposure in case-control study. But the sample sizes are more influenced by study hypothesis, type of outcome and the effect measure to be detected. In particular, when a study involving with binary outcome (or exposure), the sample is affected by RR (or OR) as the effect measure used in prospective and retrospective studies. As expected, for detecting a higher RR (e.g. RR=10), we have shown that the required sample is very small except for situation $P_1 \geq 0.90$. Such an effect should not be considered in sample size calculation since it may not be revealed in study findings. The author recommends the size of effect of RR (or OR) as 1.5-2.5 in sample size calculation unless a strong evidence exists for a higher effect. Furthermore, for a given effect size (RR) and power, a larger sample size is required for $P_1=0.95$ since there is no more room for P_2, (assuming $P_2>P_1$) and the maximum effect to be detected is about 0.05.

Lack of sufficient sample size in epidemiologic studies specifically in clinical trials does not yield a valid conclusion (9-11). Some studies in clinical trials failed to show a significant effect (3). One possible explanation is the low power of statistical test because of small sample size used. In addition, the estimation of sample size early in research protocol helps the investigator to provide enough resources such as budgets, human resources, and time. On the other hand, a very large sample size is allocated arbitrary into the study, it might waste the resources and it is time consuming and does not provide additional information. As addressed, the sample size must be calculated based on statistical principles with reasonable effect size not based on previously published studies.

One might argue the sample size was ascertained based on the number of subjects used in previously published studies. If these studies did not reach a significant valid result, how can one be sure the allocated sample size is enough unless one uses the statistical principles for sample size calculation? In addition, the difference of variation of study outcome due to socioeconomic, cultural status, genetical and biological variations add the difficulty of such sample size in detection of effect. Even the effect size might be different in various conditions. Thus, for each study protocol, the sample size should be calculated independently. This review will help the clinicians to decide and calculate a suitable sample size in their research protocol in order to detect an effect of interest with respect to study design and outcome of interest.

References

1. Houle TT, Penzien DB, Houle CK. Statistical power and sample size estimation for headache research: An overview and power calculation tools. Headache 2005; 45: 414-8.
2. Fitzner K, Heckinger E. Sample size calculation and power analysis: a quick review. Diabetes Educ 2010; 36: 701-7.
3. Detskey AS, Sackett DL. When was a negative clinical trial big enough? How many patients you needed depends on what you found. Arch Intern Med 1985; 145: 709-12.
4. Livingston EH, Cassidy L. Statistical power and estimation of the number of required subjects for a study based on the t-test. J Surg Res 2005; 126: 149-59.
5. Jones SR, Carley S, Harrison M. An introduction to power and sample size estimation. Emerg Med J 2003; 20: 453-8.
6. Devore J, Peck R. Statistics: The exploration and analysis of data. 2nd ed. California: Wadsworth Publishing Company 1993; pp: 398-406.
7. Corty EW, Corty RW. Setting sample size to ensure narrow confidence intervals for precise estimation of population values. Nurs Res 2011; 60: 148-53.
8. Springate SD. The effect of sample size and bias on the reliability of estimates of error: a comparative study of Dahlberg’s formula. Eur J Orthod 2011; Mar 29 [ahead of print].
9. Cook RJ, Sackett DL. The number needed to treat a clinically useful measure of treatment effect. BMJ 1995; 310: 452-4.
10. Abdul Latif L, Daut Amadera JE, Pimentel D, Fregni F. Sample size calculation in physical medicine and rehabilitation: a systematic review of reporting, characteristics and results in randomized control trials. Arch Phys Med Rehabil 2011; 92: 306-15.
11. de Craen AJ, Vickers AJ, Tijssen JG, Kleijnen J. Number-needed –to-treat and placebo-controled trials. Lancet 1998; 351: 310.
12. Tylor S, Muncer S. Readdressing the power and effect of significance. A new approach to an old problem: teaching statistics to nursing students. Nurs Edue Today 2000; 20: 358-64.
13. Munaro BH, William F. Statistical methods for health care research. Fifth ed. Philadelphia: Lippincott Wiliams & Wilkins 1997; pp: 261-2.

14. Armitage P. Statistical methods in medical research. Fourth ed. Oxford: Blackwell scientific publications 1977; pp: 184-189.
15. Kramer MS. Clinical epidemiology and biostatistics. First ed. Berlin: Springer-Verlag 1988; pp: 157.
16. Schlesselman JJ. Case-control studies: design, conduct, analysis. New York: Oxford University press 1982; pp: 144-69.
17. Obuchowski NA. Sample size tables for receiver operating characteristic studies. AJR Am Roentgenol 2000; 175: 603-8.
18. Liu A, Schisterman EF, Mazumadar M, Hu J. Power and sample size calculation of comparative diagnostic studies with multiple correlated test results. Boim J 2005; 47: 140-50.
19. Hajian-Tilaki KO, Hanley JA. Comparison of three methods for estimation the standard error of the area under the curve in ROC analysis of quantitative data. Acad Radiol 2002; 9: 1278-85.
20. Hajian-Tilaki KO, Hanley JA, Joseph L, Collet JP. A comparison of parametric and nonparametric approaches to ROC analysis of quantitative diagnostic tests. Med Decis Making 1997; 17: 94-102.
21. Hanley JA. Receiver operating characteristic (ROC) methodology: the state of the art. Crit Rev Diagn Imaging 1989; 29: 307-35.
22. Kumar R, Indrayan A. Receiver operating characteristic (ROC) curve for medical researchers. Indian Pediatr 2011; 48: 277-87.
23. Hanley JA, Hajian-Tilaki KO. Sampling variability of nonparametric estimates of the area under receiver operating characteristic curves: An updated. Acad. Radiol 1988; 4: 49-58.
24. Freiman JA, Chalmers TC, Smith H jr, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 “negative” trails. N Engl J Med 1978; 299: 690-4.
25. Nikerson RS. Null hypothesis significant testing: A review of an old and continuing controversy. Psychol Methods 2000; 5: 241-301.