

Theory and practice

Sample size calculation in epidemiological studies

V. Kasiulevičius¹, V. Šapoka¹, R. Filipavičiūtė²

¹ Vilnius University

² Institute of Experimental and Clinical Medicine at Vilnius University

Summary

Sample-size determination is often an important step in planning an epidemiological study. There are several approaches to determining sample size. It depends on the type of the study. Descriptive, observational and randomized controlled studies have different formulas to calculate sample size. In this article, we discuss the formulas that can help to estimate sample size in an epidemiological trial. We present a few examples from clinical practice, which may contribute to the understanding of this problem.

Keywords: sample size determination

Determining an appropriate sample size for a clinical trial is an essential step in the statistical design of the project. An adequate sample size helps ensure that the study will yield reliable information, regardless of whether the ultimate data suggest a clinically important difference between the treatments being studied, or the study is intended to measure the accuracy of a diagnostic test or the incidence of a disease. Unfortunately, many studies published in medical literature are conducted with inadequate

sample sizes, making the interpretation of negative results difficult. Conducting a study with an inadequate sample size is not only futile, it is also unethical. Exposing patients to the risks inherent in a research is justifiable only if there is a realistic possibility that the results will benefit those subjects, future subjects, or lead to substantial scientific progress.

How many individuals will I need to study? This question is commonly asked by a clinical investigator and exposes one of many issues that are best to be settled before actually carrying out a study. Consultation with a statistician is worthwhile in addressing many issues of study design, but a statistician is not always readily available.

Sample Size (n) is the number of individuals in a group under study. The larger the sample size, the greater the precision and, thus, power for a given study design to detect an effect of a given size. For statisticians, an $n > 30$ is usually sufficient for the Central Limit Theorem to hold so that normal theory approximations can be used for measures such as the standard error of the mean. However, this sample size ($n = 30$) is unrelated to the clinicians' objective of detecting biologically significant effects, which determines the specific sample size needed for a specific study [1].

Address: V. Kasiulevičius
Santariškių g. 2, Vilnius
Tel. 8 5 2365192
El. paštas Vytautas.Kasiulevicius@santa.lt

Descriptive studies

Descriptive epidemiologic studies examine differences in disease rates among populations in relation to age, gender, race, marital status, occupation and differences in temporal or environmental conditions. In general, these studies can only identify patterns or trends in disease occurrence over time or in different geographical locations, but cannot ascertain the causal agent or degree of exposure. These studies are often very useful for generating hypotheses for further research. Common descriptive designs include case studies, which may be used to describe a disease in an individual patient; case series, which may describe a disease in a group of patients; surveys, which may be used to describe the proportion of a single population that has a condition; and descriptive ecologic studies, which may be used to compare rates of a condition in several populations. Three important uses of descriptive studies include trend analysis, health-care planning, and hypothesis generation. A frequent error in reports of descriptive studies is overstepping the data: studies without a comparison group allow no inferences to be drawn about associations, causal or otherwise. Hypotheses about causation from descriptive studies are often tested in rigorous analytical studies.

Sample size calculation in descriptive study

To calculate the required sample size in a descriptive study, we need to know the *level of precision*, *level of confidence or risk* and *degree of variability* [2, 3].

The *level of precision*, sometimes called *sampling error*, is the range in which the true value of the population is estimated to be. This range is often expressed in percentage points, (e.g., ± 5 percent). The *confidence or risk level* is based on ideas encompassed under the Central Limit Theorem. The key idea encompassed in the Central Limit Theorem is that when a population is repeatedly sampled, the average value of the attribute obtained by those samples is equal to the true population value. Furthermore, the values obtained by these samples are distributed normally about the true value, with some samples having a higher value and some obtaining a lower score than the true population value. In a normal distribution, approximately 95% of the sample values are within two standard deviations of the true population value (e.g., mean).

The third criterion, the *degree of variability* in the attributes being measured, refers to the distribution of attributes in the population. The more heterogeneous a population, the larger the sample size required to obtain a given

level of precision. The more homogeneous a population, the smaller sample size required. Note that a proportion of 50% indicates a greater level of variability than either 20% or 80%. This is because 20% and 80% indicate that a large majority do not or do, respectively, have the attribute of interest. Proportion of .5 indicates the maximum variability in a population, and it is often used in determining a more conservative sample size, that is, the sample size may be larger than if the true variability of the population attribute were used (2–6].

There are several approaches to determining the sample size. These include imitating a sample size of similar studies, using published tables, and applying formulas to calculate a sample size. Effortless approach is to use the same sample size as those of studies similar to the one you plan. Without reviewing the procedures employed in these studies you may run the risk of repeating errors that were made in determining the sample size for another study. Although tables can provide a useful guide for determining the sample size, you may need to calculate the necessary sample size for a different combination of levels of precision, confidence, and variability. The best approach to determining sample size is the application of one of several formulas. For populations that are large, Cochran developed the formula to yield a representative sample for proportions [2]:

$$n = Z^2 \frac{p(1-p)}{e^2},$$

which is valid where n is the sample size, Z^2 is the abscissa of the normal curve that cuts off an area at the tails ($1 -$ equals the desired confidence level, e.g., 95%), e is the desired level of precision, p is the estimated proportion of an attribute that is present in the population. The value for Z is found in statistical tables which contain the area under the normal curve; e is level of precision.

Finite Population Correction factor

When population sizes are less than 10 times the estimated sample size, it is possible to use a finite population correction factor (*fpc*) [6]. The finite population correction factor measures how much extra precision we achieve when the sample size becomes close to the population size. The formula for *fpc* is

$$fpc = \sqrt{\frac{N-n}{N-1}},$$

where N is the size of the population and n is the size of the sample. If fpc is close to 1, then there is almost no effect. When fpc is much smaller than 1, then sampling a large fraction of the population is indeed having an effect on precision.

When the sample size is 50, it does not matter much whether the population is 10 thousand or 10 million. When the sample size is four thousand, then we have about 23% more precision with a population of ten thousand than we would for a population of ten million.

It is possible to calculate sample size directly, without the fpc calculation:

$$n = \frac{n}{1 + \frac{n-1}{N}}$$

Researchers must be cautious when using the fpc . Frequently we want to generalize results to a larger population. We may have restricted the population for convenience, but we are interested in more than just a convenient population. This extrapolation will add to the uncertainty of our estimates, so the last thing we would want to do is to use the fpc to make your confidence intervals narrower. The finite population correction factor really applies only to “warehouse” type studies, where we are trying to characterize all the data in a single physical or conceptual location. Warehouse studies are quite common in accounting, but they are unusual in medical research. In a descriptive study we also need to know the likely response rate. For example, if our calculations indicate that we need a minimum sample size of 384, but we only expect a 80% response rate, then we will need a minimum sample size of 480 to allow for a possible non-response [4].

There are two methods to determine sample size for variables that are polytomous or continuous. One method is to combine responses into two categories and then use a sample size based on proportion. The second method is to use the formula for the sample size for the mean. Yamane (1967) provides a simplified formula to calculate sample sizes for proportions [2]:

$$n = \frac{N}{1 + Ne^2}$$

where N is the size of the population and n is the size of the sample, e is the level of precision.

The formula of the sample size for the mean is similar

to that of the proportion, except for the measure of variability. The formula for the mean is shown in the equation.

$$n_0 = \frac{Z^2 \sigma^2}{e^2}$$

Where n_0 is the sample size, z is the abscissa of the normal curve that cuts off an area at the tails, e is the desired level of precision (in the same unit of measure as the variance), and σ^2 is the variance of an attribute in the population.

Case-control studies

In a case-control study, patients who have developed a disease are identified and their past exposure to suspected etiological factors is compared with that of controls or referents who do not have the disease. This permits estimation of odds ratios (but not of attributable risks). Allowance is made for potential confounding factors by measuring them and making appropriate adjustments in the analysis. This statistical adjustment may be rendered more efficiently by matching cases and controls for exposure to confounders, either on an individual basis (for example, by pairing each case with a control of the same age and sex) or in groups (for example, choosing a control group with an overall age and sex distribution similar to that of the cases). Unlike in a cohort study, however, matching, on its own, does not eliminate confounding. Statistical adjustment is still required.

Sample size for independent case-control studies

The estimated sample size n for independent case-control study is calculated as

$$n = \frac{[Z_{\alpha} \sqrt{(1+m)\bar{p}'(1-\bar{p}')} + Z_{\beta} \sqrt{p_1(1-p_1) + m p_0(1-p_0)}]^2}{(p_1 - p_0)^2}$$

$$\bar{p}' = \frac{p_1 + p_0 / m}{1 + 1/m}$$

$$p_1 = \frac{p_0 \psi^m}{1 + p_0(\psi^m - 1)}$$

$$n_c = \frac{n}{4} \left(1 + \sqrt{1 + \frac{2(m+1)}{nm|p_0 - p_1|}} \right)^2$$

where $\alpha =$ alpha, $\beta = 1 -$ power, $\psi =$ odds ratio (odds ratio of exposures between cases and controls), $m =$ number of control subjects per case subject, $p_1 =$ probability of expo-

sure in controls. p_0 can be estimated as the population prevalence of exposure, n_c is the continuity corrected sample size and Z_p is the standard normal deviate for probability p . If possible, choose a range of odds ratios that you want to detect the statistical power.

The formulas give the minimum number of case subjects required to detect a real odds ratio or case exposure rate with power and two-sided type I error probability α . This sample size is also given as a continuity-corrected value intended for the use with corrected chi-square and Fisher's exact tests [7, 8, 10].

Sample size for matched case-control studies

The estimated sample size n for matched case-control study is calculated as

$$n = \frac{[(1/\sigma_w)Z_{\alpha/2} + Z_\beta]^2}{d^2}$$

$$\sigma_w = \sqrt{\sum_{k=1}^m \frac{k t_k \psi^k (m - k + 1)}{(k \psi^k + m - k + 1)^2}}$$

$$t_k = p_1(k-1)p_{0+}^{k-1}(1-p_{0+})^{m-k+1} + (1-p_1)k p_{0-}^k - (1-p_{0-})^{m-k}$$

$$p_{0+} = \frac{p_1 p_0 + r \sqrt{p_1(1-p_1)p_0(1-p_0)}}{p_1}$$

$$p_{0-} = \frac{p_0(1-p_1) - r \sqrt{p_1(1-p_1)p_0(1-p_0)}}{1-p_1}$$

$$d = \frac{\left[\sum_{k=1}^m \frac{k t_k \psi^k}{k \psi^k + m - k + 1} \right] - 1}{\sigma_w}$$

where α = alpha, β = 1 – power, ψ = odds ratio, r – correlation coefficient for exposure between matched cases and controls, p_0 – probability of exposure in the control group, m – number of control subjects matched to each case subject. When r is not known from previous studies, some authors state that it is better to use a small arbitrary value for r , e. g., 0.2, than it is to assume independence (a value of 0) [9]. p_0 can be estimated as the population prevalence of exposure. Note, however, that due to matching, the control sample is not a random sample from the population; therefore, population prevalence of exposure can be a poor estimate of p_0 (especially if confounders are strongly associated with exposure (9)). If possible, choose a range of odds ratios that you want to detect the

statistical power. If you are using more than one control per case, then this function also provides the education in sample size relative to a paired study that you can obtain using your number of controls per case [9].

Cohort studies

The starting point of a cohort study is the recording of healthy subjects with and without exposure to the putative agent or the characteristic being studied. Individuals exposed to the agent under study (index subjects) are followed over time and their health status is observed and recorded during the course of the study. In order to compare the occurrence of disease in exposed subjects with its occurrence in non-exposed subjects, the health status of a group of individuals not exposed to the agent under study (control subjects) is followed in the same way as that of the group of index subjects.

Sample size for independent cohort studies

The estimated sample size n for independent cohort studies is calculated as

$$n = \frac{[Z_\alpha \sqrt{(1+1/m)\bar{p}(1-\bar{p})} + Z_\beta \sqrt{p_0(1-p_0)/m + p_1(1-p_1)}]^2}{(p_0 - p_1)^2}$$

This formula gives the minimum number of case subjects required to detect a true relative risk or experimental event rate with power and two-sided type I error probability α (alpha). 5% is the usual choice for α . Usual values for power (probability of detecting a real effect) are 80%, 85% and 90%. β = 1 – power, n_c is the continuity corrected sample size, m is the number of control subjects per experimental subject, p_0 is the probability of event in controls, p_1 is the probability of event in experimental subjects, and Z_p is the standard normal deviate for the probability p .

$$\bar{p} = \frac{p_1 + m p_0}{m + 1}$$

$$n_c = \frac{n}{4} \left(1 + \sqrt{1 + \frac{2(m+1)}{n m |p_0 - p_1|}} \right)^2$$

p_0 (probability of event in controls) can be estimated as the population prevalence of the event under investigation. If possible, choose a range of relative risks that you want to detect the statistical power. This sample size is also given

as a continuity-corrected value intended for use with corrected chi-square and Fisher’s exact tests [7–10].

Sample size for paired cohort studies

The estimated sample size n for paired cohort studies is calculated as

$$n = \frac{\left[\frac{Z_{\alpha/2}}{2} + Z_{\beta} \sqrt{p_a(1 - p_a)} \right]^2}{(p_a - 0.5)^2(p_x p_y)}$$

$$p_a = \frac{p_y}{p_x + p_y}$$

$$p_y = p_1(1 - p_0) - r * \sqrt{p_1(1 - p_1)p_0(1 - p_0)}$$

$$p_x = p_0(1 - p_1) - r * \sqrt{p_1(1 - p_1)p_0(1 - p_0)}$$

where α = alpha, β = 1 – power, r – correlation coefficient for failure between paired subjects, p_0 – event rate in the control group, p_1 – event rate in experimental group. p_0 can be estimated as the population event rate. Note, however, that due to matching, the control sample is not a random sample from the population; therefore, population event rate can be a poor estimate of p_0 (especially if confounders are strongly associated with the event). r can be estimated from previous studies; note that r is the phi (correlation) coefficient given for a two by two table by the chi-square function. When r is not known from previous studies, some authors state that it is better to use a small arbitrary value for r , e. g., 0.2, than to assume independence (a value of 0). This formula gives you the minimum number of subject pairs required to detect a true relative risk [7, 8, 11].

Randomized controlled study

Randomized Controlled Clinical Trial (RCT): a prospective, analytical, experimental study using primary data generated in the clinical environment. Individuals similar at the beginning are randomly allocated to two or more treatment groups and the outcomes of the groups, are compared after a sufficient follow-up time. Traditionally, the *control* in randomized controlled trials refers to studying a group of treated patients not in isolation but in comparison to other groups of patients, the *control groups*, who by

not receiving the treatment under study give investigators important clues to the effectiveness of the treatment, its side effects, and the parameters that modify these effects. In the hierarchy of research designs, the results of randomized, controlled trials are considered to be evidence of the highest grade, whereas observational studies are viewed as having less validity because they reportedly overestimate treatment effects [16, 17].

Sample size calculations in randomized trials

For scientific and ethical reasons, the sample size for a trial needs to be planned carefully, with a balance between clinical and statistical considerations. Ideally, a study should be large enough to have a high probability (power) of detecting, as statistically significant, a clinically important difference of a given size if such a difference exists. The size of an effect considered to be important is inversely related to the sample size necessary to detect it; that is, large samples are necessary to detect small differences. Elements of the sample size calculation are [13]:

1. the estimated outcomes (P_1 and P_2) in each group (which implies the clinically important target difference between the intervention groups),
2. the *alpha* (Type I) error level,
3. the statistical power (or the *beta* (Type II) error level), and
4. for continuous outcomes, the standard deviation of the measurements [14, 15].

The errors are of two types: type-1 error (alpha) and type-2 (beta) error. Type-1 error is the false positive rate or probability of declaring a treatment difference where none exists, also known as the significance level or α -level, usually, fixed at $\alpha = 5\%$ (two-sided) by regulatory agencies. Type-2 error (β) is a false negative rate or probability of failing to detect a treatment difference that actually exists. It is also called the β -error, and $1 - \beta$ is known as the power of the study [14–18].

Unlike the statistical power and the level of significance, which are generally chosen by convention, the underlying expected event rate (P_1) (in the standard or control group) must be established by other means, usually from previous studies including observational cohorts. These often provide the best information available but may overestimate event rates, as they can be from a different time or place and thus subject to the changing and differing background practices. Additionally, trial participants are often “healthy volunteers”, or at least people with stable conditions without other comorbidities which may further

erode the study event rate compared with observed rates in the population.

The effect of the treatment (P_2) in a trial can be expressed as an absolute difference, i. e. the difference between the rate of the event in the control group and the rate in the intervention group, or as a relative reduction, i. e. the proportional change in the event rate with the treatment. If the rate in the control group is 6.3% and the rate in the intervention arm is 4.2%, the absolute difference is 2.1%; the relative reduction with intervention is 2.1% / 6.3%, or 33%. Investigators should take into consideration any cost or logistical advantages or disadvantages of the interventional treatment compared with a standard care.

The number required to participate in each group is given in the equation below:

$$n = \frac{(P_1(1-P_1) + P_2(1-P_2))}{(P_2 - P_1)^2} * f(\alpha, \beta)$$

The values of $f(\alpha, \beta)$ for various values of α and β are given in the table below.

Table of values of $f(\alpha, \beta)$

		POWER (1-Beta)			
		0.05	0.1	0.2	0.5
	0.1	10.8222	8.5638	6.1826	2.7055
	0.05	12.9947	10.5074	7.8489	3.8415
	0.02	15.7704	13.0169	10.0360	5.4119
Alpha (2- sided)	0.01	17.8142	14.8794	11.6790	6.6349

Fleiss (12) suggests that the above formula may be improved by making a continuity correction giving n' in each group, where

$$n' = \frac{n}{4} \left[1 + \sqrt{1 + \frac{4}{n|P_2 - P_1|}} \right]^2$$

Many adjustments have been made to improve the properties of power calculations for binary outcomes, and different packages may give slightly different results.

In the example discussed above, P_1 and P_2 represent percentage successions in the two groups, where $P_1 = 90\%$, $P_2 = 95\%$, $\alpha = 0.05$ (2-tail), $\beta = 0.1$

$$578 = \frac{(0.90(0.10) + 0.95(0.05))}{(0.95 - 0.90)^2} * (1.96 + 1.28)^2$$

Hence, 578 patients would be required in each treatment group.

Calculating the sample size for comparing a continuous outcome between two groups

The number required to participate in each group is given in the equation below:

$$n = 2 \left\{ \frac{(z_\alpha + z_\beta) \sigma}{\delta^*} \right\}^2$$

$$n = 2 \cdot \left\{ \frac{f(\alpha, \beta) \sigma^2}{\delta^2} \right\}$$

We will power our trial for a 5% 2-sided test with 80% power, i. e. $z_\alpha = 1.96$ and $z_\beta = 0.84$. Thus, the required number per group is:

$$n = 2 \left\{ \frac{(1.96 + 0.84) 3}{1} \right\}^2 = 141.12 \text{ per group.}$$

In the reported trial, only 59/69 = 85.5% had data on social functioning (though this probably refers to social functioning at 18 months). So, anticipating 15% to drop out gives $141.12 / 85 = 166.02$ per group, i.e. about 170 per group, i.e. 340 in total.

Conclusions

Epidemiologists Kenneth F Schulz and David A wrote in "Lancet" one year ago that if the scientific world insisted solely on large trials, many unanswered questions in medicine would languish unanswered. Some shift of emphasis from a fixation on sample size to a focus on methodological quality would yield more trials with less bias. Unbiased trials with imprecise results trump no results at all [19]. We believe that our considerations will be interesting for medical professionals.

References

1. Cochran WG. Sampling Techniques. 2nd ed. New York: John Wiley and Sons, Inc., 1963.
2. Yamane T. Statistics, An Introductory Analysis. 2nd ed. New York: Harper and Row, 1967.
3. Miaoulis G, Michener RD. An Introduction to Sampling. Dubuque, Iowa: Kendall/Hunt Publishing Company, 1976.

4. Eng J. Sample size estimation: how many individuals should be studied? *Radiology*. 2003; 227: 309–13.
5. Daniel W. *Biostatistics: a foundation for analysis in the health sciences*. 7th ed. New York, NY: Wiley, 1999; 180–5, 268–70.
6. Israel GD. The Evidence of Sampling Extension Program Impact. Program Evaluation and Organizational Development, IFAS, University of Florida. PEOD-5. 1992.
7. Dupont WD. Power calculations for matched case-control studies. *Biometrics*. 1988; 44: 1157–68.
8. Dupont WD. Power and sample size calculations. *Controlled Clinical Trials*. 1990; 11: 116–28.
9. Casagrande JT, Pike MC, Smith PG. An improved approximate formula for calculating sample sizes for comparing two binomial distributions. *Biometrics*. 1978; 34: 483–6.
10. Schlesselman JJ. *Case-Control Studies*. New York: Oxford University Press, 1982.
11. Breslow NE, Day NE. *Statistical Methods in Cancer Research: The Analysis of Case-Control Studies*. Lyon: International Agency for Research on Cancer, 1980.
12. Fleiss JL. *Statistical Methods for Rates and Proportions*. 2nd ed. New York: Wiley, 1981.
13. Meinert CL. *Clinical Trials: Design, Conduct and Analysis*. New York: Oxford University Press, 1986.
14. Moher D, Schulz KF, Altman D. The CONSORT statement: revised recommendations for improving the quality of reports of parallel group trials. *Lancet*. 2001; 357: 1191–94.
15. Altman DG, Schulz KF, Moher D et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med*. 2001; 134: 663–94.
16. Kirby A, Gebiski V, Keech AC. Determining the sample size in a clinical trial. *MJA*. 2002; 177(5): 256–7.
17. Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. *JAMA*. 1994; 272: 122–4.
18. Goodman SN, Berlin JA. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Ann Intern Med*. 1994; 121: 200–6.
19. Schulz KF, Grimes DA. Sample size calculations in randomized trials: mandatory and mystical. *Lancet*. 2005; 365: 1348–53.

*Received 16 June, 2006,
accepted 11 September, 2006*

IMTIES DYDŽIO NUSTATYMAS EPIDEMIOLOGINĖSE STUDIJOSE

V. Kasiulevičius¹, V. Šapoka¹, R. Filipavičiūtė²

¹ Vilniaus universitetas

² Vilniaus universiteto Eksperimentinės ir klinikinės medicinos institutas

Santrauka

Imties nustatymas dažnai yra svarbus numatant epidemiologinį tyrimą. Yra keletas metodikų, padedančių nustatyti imties

dydį. Tai priklauso nuo tyrimo rūšies. Aprašomųjų, stebimųjų ir atsitiktinės atrankos būdu kontroliuojamų tiriamųjų imtims nustatyti naudojamos skirtingos formulės. Šiame straipsnyje aptariame tyrime vartojamas formules imties dydžiui nustatyti, pateikiame keletą pavyzdžių, padėsiančių geriau suvokti problemos aktualumą.

Raktažodžiai: imties dydžio nustatymas