Outbreak, Surveillance, Investigation & Response (OSIR) Journal
Field Epidemiology Training Program, Division of Epidemiology
Department of Disease Control, Ministry of Public Health, Thailand
Tel: +6625903894, Fax: +6625903845, Email: osireditor@osirjournal.net, http://www.osirjournal.net

The Grammar of Science: How Many Participants Do I Need in My Survey?

Jaranit Kaewkungwal*

Mahidol University, Thailand

*Corresponding author email: jaranitk@biophics.org

Received: 4 Nov 2023; Revised: 24 Nov 2023; Accepted: 17 Dec 2023
https://doi.org/10.59096/osir.v16i4.266224

An adequate sample size is a prerequisite for any research study.[1] Sample size justification is required to be considered from both scientific and ethical points of view. A smaller sample size than required may result in "underpowered study" with questionable reliability or reproducibility, and not detecting results that are in fact important. A larger sample size than required consumes unnecessary resources and may also lead to questionable statistically significant result for a tiny different deviation even though it is not practically important.[1–3]

In this paper, we will focus on sample size calculation for a descriptive survey. Survey is generally a descriptive cross-sectional study design to describe certain phenomena in population of interest; (e.g., prevalence of back pain among office workers, mean depression score among cancer patients). However, survey can be an analytic cross-sectional study design with the specific objective to assess the association among variables or to compare a variable between groups (e.g., association between smoking and lung cancer, association between depress score and quality of life score).

Sample Size for Descriptive Survey

Main objective of a descriptive cross-sectional study is to "estimate" population characteristics, so-called "parameter". The classic formula for sample size calculation of a descriptive study depends on the types of parameter to be estimated, categorical data (prevalence, proportion, percentage) or continuous data (mean/variability).

For categorical data

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

For continuous data

$$n = z^2_{1-\frac{\alpha}{2}} \frac{\sigma^2}{d^2}$$

There are three basic elements in both formula that are required.

Expected Value of the Parameter

In sample size calculation, you need to input the expected value of the parameter that you want to estimate. The expected value is a proportion (p) for categorical data or mean and standard deviation (μ, σ) for continuous data. This expected value is typically based on "priori information" which could be obtained from previously published studies, pilot study, or expert opinion.

Confidence Level / Interval (CI)

CI refers to the percentage of all possible samples that can be expected to include the true value of the population parameter that you attempt to estimate. CI is typically set up as 95% or higher. The CI relates to the value of the area under curve ($Z_{\alpha/2}$). When setting up CI=95%, the $Z_{\alpha/2}$=1.960; and when CI=99%, the $Z_{\alpha/2}$=2.576. The higher the CI, the larger $Z_{\alpha/2}$ and consequently the larger the sample size.

Precision or Margin of Error (d)

As different samples from the same population would give different estimates of the true value, thus the estimate of the true value is generally inferred with some margin of error or precision. The terms, "precision" and "margin of error", are opposite to one another. Margin of error expresses the maximum expected difference between the true population parameter and a sample estimate of that parameter.[4] Setting up the lower the margin of error would lead to the higher the precision (reliability) of the estimate. The higher precision (the lower margin of error), the larger sample size. It is a challenging issue to select the precision level which may be quite subjective, depending on the objective and nature of the survey.[1,5,6] The precision can be set up as "absolute precision" or

"relative precision". Absolute precision is simply set by specifying the exact value of the margin of error or the absolute uncertainty of the estimated parameter. For example, based on a study elsewhere the prevalence of tuberculosis (TB) is 20%; and in your survey, you simply set up 10% as the absolute margin of error, that means the you expect the prevalence is to be estimated with an uncertainty of 10% on either side of the estimate (between 10–30%). Relative precision is set corresponding to the priori information of the estimate. For example, based on a study elsewhere the prevalence of TB is 20% and you set up a relative margin of error for the current survey as 10% of the previous estimate (10% of 20%=2%), that means you expect the prevalence is to be estimated with an uncertainty of 2% on either side of the estimate (between 18–22%).

In a certain situation when samples are drawn from a finite (limited and small size) population, the formula for sample size calculation can be adjusted by taking into consideration the size of the population under survey.[6] For example, if you want to get a certain estimate form the 100 specimens archived in the hospital laboratory, the population size in this case is limited to only N=100 specimens, not the "population at large" or infinite population. The formulas for sample size of finite population are:

For categorical data

$$n = \frac{Np(1-p)z^2_{1-\frac{\alpha}{2}}}{d^2(N-1)+p(1-p)z^2_{1-\frac{\alpha}{2}}}$$

For continuous data

$$n = \frac{N\sigma^2 z^2_{1-\frac{\alpha}{2}}}{d^2(N-1)+\sigma^2 z^2_{1-\frac{\alpha}{2}}}$$

Non-responses in Survey

The calculated sample size is the minimum number that you should have at the end of the study in order to obtain the parameter estimate with the precision and CI that you proposed. However, when conducting a survey, you will unlikely get the responses back from all potential participants that you attempt to recruit, and responses from some participants may be incomplete. Such non-response is a potential source of bias. Securing a high response rate to a survey can be hard to control, particularly in a postal survey, but still difficult for a face-to-face or telephone interview.[7] The non-response rate is usually unknown and unpredictable; it may be based on previous experience or a pilot study.[3] It is suggested in literature that achievable and acceptable rate for a survey study should be around 75% for interviews and 65% for self-completion postal questionnaires.[7] If p is the proportion of non-responses, the number of sample size must be increased by a factor of (1 – p).

$$n_{adjusted} = n / (1 - p)$$

Sampling Techniques

Simple Random Sampling vs. Cluster Sampling

Sampling refers to the process of choosing samples from a total population. We can classify sampling methods into 2 types: probability vs. non-probability sampling.[4] Probability sampling includes methods that are based on two concepts: (1) equal probability of selection, everybody in the population has equal chance of being selected, and (2) proportionate to size, the proportions of the sample subgroups reflect the proportions within the population subgroups. Non-probability sampling is based on the concept of relevancy or representativeness of the samples to the population of interest. Survey sampling is usually based on probability sampling technique.[4] The goal of a probability sampling technique is to minimize the sampling error of the parameter to be estimated.[8]

Simple Random Sampling (SRS)

SRS is a probability sampling with equal probability of selection approach which can be done with or without replacement. Usually, the SRS is conducted without replacement because it is more convenient and gives more precise results.[8] The sample size formula presented above are for SRS survey. With a large enough sample size, SRS has high external validity as it represents the larger targeted population.[9,10] If you want to estimate the parameter of interest in different subgroups or strata (say by gender, age, geography, etc.), given using the same priori information and level of precision for all stratum, you can simply multiply the sample size calculated from SRS by the number of strata.[5]

Cluster Sampling (CS)

Sometimes it is too expensive to draw samples that spread out over a large geographic area. It may be much more practical and reduce costs to conduct a survey employing CS that the participants will be randomly selected within only the selected areas—so-called "clusters".[11] CS divides the population into clusters. A number of clusters are then selected randomly to represent the total population, and all eligible participants within the selected clusters are included in the survey. If not all, but some participants are randomly selected within the selected clusters, it is called multi-stage sampling technique.[8] Figure 1 presents the differences between simple random sampling and cluster sampling.

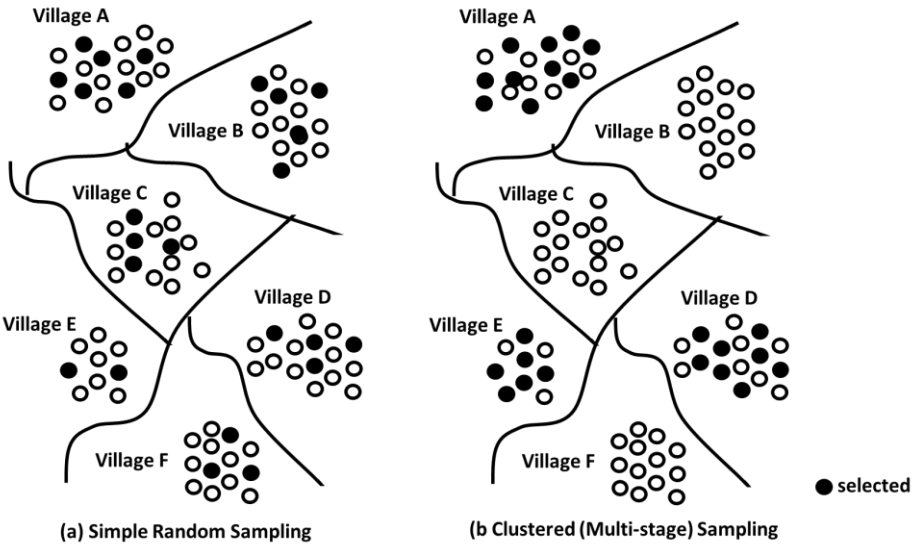
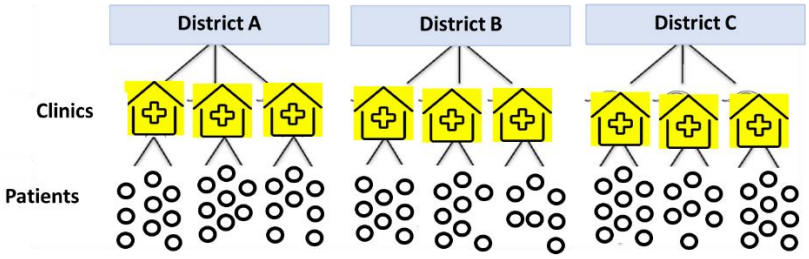


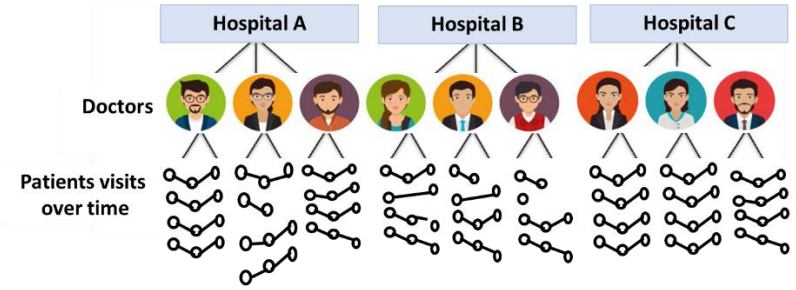
Figure 1. Simple random sampling vs. Cluster sampling

Effect of Clustered Samples in Cluster Sampling Survey

Clustered samples generally refer to surveyed participants who are physically grouped in geographical locations (villages, districts, provinces), but they may also refer to a group of participants within a shared relationship such as within a clinic/hospital. There might be more than one level of clustering. For example, a group of villagers is clustered within (i.e., get accessed to) a sub-district health facility, and a set of sub-district facilities are clustered in the same district, while different districts may possess different types of resources or surrounding conditions. Or, patients are seen by (clustered within) the same doctor, while doctors are working the same hospital, and different hospitals may have different resources and quality (Figure 2).



(a) patients clustered within a sub-district clinic which are clustered within a district



(b) patient visits (repeated measures) clustered within a doctor who are clustered within each hospital

Figure 2. Cluster levels

Cluster samples violate the SRS assumption of independence of observations as they reside in and share the same environment or settings. The observations within the same cluster may potentially be similar to one another than observations across clusters.[12] Similarities among participants in a cluster can reduce the variability of observations. This leads to a statistical concept called—the intra-cluster correlation (ICC), ρ *("rho")*—such that there might be variation across clusters more than variation within clusters. That is, ICC reflects a measure of the relatedness of clustered samples. It accounts for the relatedness of clustered samples by comparing the variation of observations (variance) within clusters (S^2_w) with the variance between

clusters (S^2_b).[11,12] Similar to the process of comparing the between and within group variances in analysis of variance (ANOVA), the ICC is calculated by dividing the between-cluster variation by the total variation of the variable to be estimated.[13,14]

$$ICC\ (\rho) = S^2_b / (S^2_b + S^2_w)$$

Like other correlations, ICC value ranges between 0 to 1. Figure 3 shows theoretical quantity where ICC=1 when all observations (Ys) within a cluster are identical (Figure 3(a)). ICC will be smaller when the variance within cluster S^2_w is much greater than the variance between clusters S^2_b. When there is no correlation of observations within a cluster, ICC=0 (Figure 3(b)).[11,14]

ID	Cluster	Y		
1	1	10	Mean within cluster =	10.00
2	1	10	Variance within Cluster =	0.00
3	1	10		
4	2	17	Mean within cluster =	17.00
5	2	17	Variance within Cluster =	0.00
6	2	17		
7	3	21	Mean within cluster =	21.00
8	3	21	Variance within Cluster =	0.00
9	3	21		
Mean Total =16.00			ICC = 1	
Variance Total =23.25				

(a) variances between clusters, no variances within clusters

ID	Cluster	Y		
1	1	10	Mean within cluster =	16.00
2	1	17	Variance within Cluster =	31.00
3	1	21		
4	2	10	Mean within cluster =	16.00
5	2	17	Variance within Cluster =	31.00
6	2	21		
7	3	10	Mean within cluster =	16.00
8	3	17	Variance within Cluster =	31.00
9	3	21		
Mean Total =16.00			ICC = 0	
Variance Total =23.25				

(b) no variances between clusters; variances within clusters

Figure 3. Intra-cluster correlation (ICC) with hypothetical scenarios

In other words, ICC tells you the degree of similarity between participants belonging to the same cluster; If it is 0, there is no evidence of clustering effects in the observations. If the ICC is approaching 1, then there is a clustering effect. Failure to take into consideration of ICC when designing CS survey might result in an under-powered study. ICC and cluster size (average number of observations to be sampled within a cluster) are thus used as part of "design effect" in sample size calculation for CS survey.[12] The difficulty is what should be the level of ICC to be input in the sample size calculation. We can calculate ICC by a post-hoc examination of the study results from previous studies that have used CS, but this has been rarely reported in the published works.[12] ICC generally varies corresponding to the parameters being estimated and the type of clustering, However, it was suggested in literature that ICC value are commonly set at ranges between 0.005–0.30.[12–14]

Design Effect in Descriptive Survey

Design effect (D) is a measure for the relative efficiency of the estimated parameter under a sampling technique employed in the survey. In other words, D is a constant used to correct for the effect (i.e., sampling error (SE)) of clustering and stratification on the estimated parameter.[15] In this case, the SE can be defined as the difference between study result and population value due to random selection of sample. It should be noted that SE is not the "bias" of the study because it can be predicted, calculated, and accounted for; and SE is influenced by sample size and sampling technique employed in the survey.[15] Thus, when taking into consideration of clustered observations, the D should be applied to adjust for the efficient sample size, particularly in CS survey.

$$n_{adjusted} = n \times D$$

In CS survey, you start with sample size calculation using the formula for SRS method and follow by accounting for the D.[16] The sample size will increase or decrease by D. When D=1, it means no effect of sample design on SE. If D >1 then sample design inflates the SE of the estimate while D <1 reduces the SE.[13] As suggested in literature, there are several ways in determining the design effect.[5,14,17]

Variance (Standard Error) Difference

D is defined as a ratio between the variance of the parameter to be estimated under CS method vs. the variance of the same estimate under SRS method. In this approach you need to have the known variances (S^2) which may be obtained from previous survey experiences. If not known, as a rule of thumb, D is typically set at 1.5, 2 or 2.5.

$$D = \frac{S^2 \text{ (Variance or Standard Error) of the designed sampling method}}{S^2 \text{ (Variance or Standard Error) of simple random sampling}}$$

Example

Adapted from a cross-sectional survey that was designed to determine prevalence of HIV among people who use drugs.[18] In the sample size calculation, the following elements were set: (1)

population size—the estimated people who used drugs (PWUD) in the study areas, N=13,000; (2) priori information—the estimated HIV prevalence of 3.5%, p=0.035, based on national report in recent years; (3) absolute margin error of 1.5%, d=0.015; (4) CI=95%. With the formula for sample size calculation for finite population SRS sampling survey, the sample size=553. For sample size adjustment, additional elements were set: (1) the design effect of variance difference, D=2.0, based on previous surveys, and (2) the expected non-response rate: 20%. The minimum sample size required in the survey=(553 x 2 /0.8)=1,383. According to the sampling plan, the total number of sample size was stratified by study sites in 21 locations, making it roughly 15% of the estimated PWUD in each site.

Variance Inflation Factor

In this approach, the design effect gives the increase in the variance arising from the clustering size (average cluster size, m) and the mean (μ) and standard deviation (sd) of the parameter to be estimated across clusters. The mean and sd may be based on previous studies/experiences. The coefficient of inflation variation is the ratio of the sd to the mean across all possible clusters, $\alpha=sd/\mu$. As the result, the D will be big if there are large clusters (big m), the clusters are very different (big α) and/or the parameter to be estimate is high (big μ, high prevalence).

$$D = 1 + m\alpha^2\mu$$

Example

Adapted from a cross-sectional survey that was designed to determine TB prevalence at national level.[19] In sample size calculation, the following elements were set: (1) priori information based on previous reports—the expected prevalence of TB=483/100,000, p=0.00483; (2) relative precision of 25%=0,00483 x 0.25, d=0.0012075; (3) CI=95%. With the formula for SRS sampling survey, the sample size=12,675. With the plan for cluster sampling, the design effect was based on variance inflation formula, $D = 1 + m\alpha^2\mu$. With the targeted 42 clusters across the country, the cluster size, m=12,675/42 ≈ 302. The sd or variation of TB prevalence across clusters was assumed to be ±40% of the average value, thus α ≈ 0.4. With the prevalence of TB, μ=0.00483. then $D = 1 + 302 \times 0.4^2 \times 0.00483=1.23$ (the design effect increased the variance by 23%), The sample size required for this survey was approximately 12,675 x 1.23=15,590.

Intraclass Correlation Approach

Another popular variance inflation factor, D is accounted for cluster size (average number of participants per cluster, m) and ICC (ρ).

$$D = 1 + (m-1)\,\rho$$

Example

Adapted from a cross-sectional study with the purpose to determine the use of physical restraints in nursing home.[20] The primary sampling unit of study were 103 nursing homes. In sample size calculation, the following elements were set: (1) priori information based on previous studies as well as consensus among 5 experts in the field—prevalence of physical restraints of 25%, p=0.25; (2) absolute margin of error of 6%, d=0.06; (3) CI=95%. With the formula for SRS sampling survey, the sample size=200. The sample size was adjusted for design effect by assuming: (1) ICC=0.08 according to the result from a similar study, and (2) mean cluster size (#participants/home)=68. The design effect, D= + (m−1) x ICC=1 + (68−1) x 0.08=6.36. With cluster design effect, the required sample size=6.36 x 200=1,272 residents. With cluster size of 68, overall 19 (1,272/68) out of 103 nursing homes were randomly selected to participate in the study.

Conclusion

To answer **"***How many participants do I need in my survey?"* or *"What should be my sample size?"* depends on your study objective, type of parameter, sampling technique, design effect, and non-response rate. Plus two more important issues that are not discussed here —the costs/resources (man & money) and logistics (management) to conduct your survey.

Suggested Citation

Kaewkungwal J. The grammar of science: how many participants do I need in my survey? OSIR. 2023 Dec;16(4):224–9. doi:10.59096/osir.v16i4.266224.

References

1. Nundy S, Kakar A, Bhutta ZA. How to practice academic medicine and publish from developing countries? [Internet]. Singapore: Springer: 2021 Oct 23 [cited 2023 Nov 4]. 465 p. <https://link.springer.com/content/pdf/10.1007/978-981-16-5248-6.pdf>. doi:10.1007/978-981-16-5248-6.
2. Althubaiti A. Sample size determination: A practical guide for health researchers. J Gen Fam Med. 2022 Dec 14;24(2):72–8. doi:10.1002/jgf2.600.

3. Cornish RP. Statistics: An introduction to sample size calculations. [Seattle]: Mathematics Learning Support Center; 2006 [cited 2023 Nov 4]. 5 p. <https://www.semanticscholar.org/paper/Statistics-%3A-An-introduction-to-sample-size-Cornish/bec3b6bfdbb7672032933573ea29bb57c9ef9a39#citing-papers>

4. Berman HB. Survey sampling [Internet]. [place unknown]: Stat Trek; [cited 2023 Oct 28]. <https://stattrek.com/survey-research/survey-sampling>

5. World Health Organization. Calculating the sample size for surveys for the prevalence of TB. Geneva: World Health Organization; [cited 2023 Nov 4]. 13 p. <https://cdn.who.int/media/docs/default-source/hq-tuberculosis/global-task-force-on-tb-impact-measurement/meetings/2008-03/p05_sample_size_design.pdf>

6. Wayne WD. Biostatistics: A foundation of analysis in the health sciences. 6th ed. Chichester (UK): John Wiley & Sons, Inc.; 1995.

7. Kelley K, Clark B, Brown V, Sitzia J. Good practice in the conduct and reporting of survey research. Int J Qual Health Care. 2003 Jun; Jun;15(3):261–6. doi:10.1093/intqhc/mzg031.

8. Statistics Canada. Probability sampling [Internet]. Ontario: Authority of the Minister responsible for Statistics Canada; 2021 Sep 2 [cited 2023 Nov 4] <https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch13/prob/5214899-eng.htm>

9. Fleetwood D. Simple random sampling: definition and examples [Internet]. Seattle: Question Pro; [cited 2023 Nov 4]. <https://www.questionpro.com/blog/simple-random-sampling/>

10. Thomas L. Simple random sampling: definition, steps & examples [Internet]. Scribbr; 2020 Aug 28 [updated 2023 Jun 22, cited 2023 Nov 4]. <https://www.scribbr.com/methodology/simple-random-sampling/>

11. Killip S, MD, Mahfoud Z, Pearce K. What is an intracluster correlation coefficient? Crucial concepts for primary care researchers. Ann Fam Med. 2004 May–Jun;2(3):204–8. doi:10.1370/afm.141.

12. Knox SA, Chondros P. Observed intra-cluster correlation coefficients in a cluster survey sample of patient encounters in general practice in Australia. BMC Med Res Methodol. 2004 Dec 22;4(1):30. doi:10.1186/1471-2288-4-30.

13. Holodinsky JK, Austin PC, Williamson TS. An introduction to clustered data and multilevel analyses. Fam Pract. 2020 Oct 19;37(5):719–22. doi:10.1093/fampra/cmaa017.

14. Newsom JT. Intraclass correlation coefficient [Internet]. Portland (OR): Portland State University; 2019 [cited 2023 Nov 4]. 2 p. <https://web.pdx.edu/~newsomj/mlrclass/ho_icc.pdf>

15. Rafferty A. Session 1: Introduction to complex survey design [Internet]. Manchester: University of Manchester; [cited 2023 Nov 4]. 58 p. <https://dam.ukdataservice.ac.uk/media/440347/rafferty.pdf>

16. IMMPaCt Program, Division of Nutrition, Physical Activity, and Obesity Centers for Disease Control and Prevention. Micronutrient Survey Manual & Toolkit. [Internet]. Atlanta: Nutrition International, Centers for Disease Control and Prevention (US); [updated 2022 Mar 31]. Module 5: Calculation of sample size for a single cross-sectional cluster survey; [cited 2023 Nov 4]; [about 8 p}. <https://mnsurvey.nutritionintl.org/categories/13>

17. Hsieh FY, Lavori PW, Cohen HJ, Feussner JR. An overview of variance inflation factors for sample-size calculation. Eval Health Prof. 2003 Sep;26(3):239–57. doi:10.1177/0163278703255230.

18. Tuot S, Mburu G, Mun P, Chhoun P, Chann N, Prem K, et al. Prevalence and correlates of HIV infection among people who use drugs in Cambodia: a cross-sectional survey using respondent driven sampling method. BMC Infect Dis. 2019 Jun 11;19(1):515. doi:10.1186/s12879-019-4154-5.

19. Williams B, Gopi PG, Borgdorff MW, Yamada N, Dye C. The design effect and cluster samples: optimising tuberculosis prevalence surveys. Int J Tuberc Lung Dis. 2008 Oct; 12(10):1110–5.

20. Hofmann H, Schorro E, Haastert B, Meyer G. Use of physical restraints in nursing homes: a multicentre cross-sectional study. BMC Geriatr. 2015 Oct 21;15:129. doi:10.1186/s12877-015-0125-x.