Alqalam Journal of Medical and Applied Sciences. 2026;9(9):2646-2650
https://doi.org/10.54361/ajmas.269902

Original article

Framework for Calculating Sample Size to Estimate Disease Prevalence

Elmabrok Masaoud, Maamar Mohmmed, Abeer Alhajaji, Nagat Alforje

Faculty of Veterinary Medicine and Agricultural Sciences, University of Zawia, Libya
Corresponding Email. e.masaoud@zu.edu.ly

Abstract
Determining an appropriate sample size is a critical component of medical and veterinary research. In cross-sectional studies, in particular, an adequate sample size is an essential part for accurately estimating baseline prevalence of diseases or other health factors in human or animal populations. Proper sample size calculation ensures that the study includes enough humans or animals to estimate the prevalence of a disease, condition, parasite, pathogen, or to establish a baseline with acceptable precision. When sample sizes are too small, studies become underpowered and may fail to identify true effects or accurately characterize disease prevalence. In this paper, we provide a practical and accessible framework for calculating sample size for prevalence in medical and veterinary cross-sectional research studies.
Keywords. Sample size, Prevalence, Medical and Veterinary Research.

Introduction

Determining an appropriate sample size is very important when designing a cross-sectional study aimed at estimating the prevalence (proportion) of a population [1]. These cross-sectional studies in human medicine or veterinary science that aim to determine the prevalence of a disease, condition, parasite, pathogen, or to establish a baseline at a given point in time are called prevalence studies [2-4]. Prevalence [4] refers to the proportion of cases of disease, condition, exposure, pathogen, or other health factor present in a population at a specific point in time or over a specified period [5,6]. A precise estimation of unknown parameters, such as the population prevalence, requires drawing a (random) sample of an adequately large sample size [7]. Larger samples produce more precise prevalence estimates by narrowing the associated confidence intervals. If the sample size is insufficient, the resulting estimate may lack the precision needed to identify a prevalence level of practical relevance.

A small sample size can substantially affect the precision of a prevalence estimate. When the number of observations is limited, the estimate becomes more susceptible to random variation, which reduces statistical stability. This increase in uncertainty is reflected in a wider confidence interval, indicating that the true population prevalence could plausibly fall across a broader range of values. In other words, insufficient sample size diminishes the reliability of the estimate by weakening the degree of confidence with which researchers can generalize their findings to the target population.

A small sample size can also influence the level of significance associated with a prevalence estimate. When the number of observations is limited, statistical tests may have reduced power, meaning they are less capable of detecting true differences or associations. As a result, even when a real effect exists, the analysis may fail to reach the conventional threshold for statistical significance. This occurs because small samples produce greater variability and less stable estimates, which in turn inflate standard errors. Consequently, the probability of obtaining a statistically significant result decreases, not because the effect is absent, but because the study lacks sufficient data to demonstrate it with confidence.

Despite the large body of literature on sample size determination [8-18], many authors overlook or simply omit the calculation of appropriate sample size regardless of its importance; see, for example, [19-22]. Prevalence estimates play a critical role in informing public health policy, guiding the allocation of resources, and supporting clinical planning. However, when these estimates are derived from inadequately small sample sizes, they become vulnerable to substantial statistical error. Underpowered samples increase the likelihood that the calculated prevalence will deviate from the true population value, potentially leading decision-makers to rely on inaccurate or misleading information. Such biases can have deep and expansive impacts, including misdirected funding, inappropriate service planning, and ineffective public health interventions. Therefore, this study aims to provide a practical and accessible framework for calculating sample size to estimate prevalence in medical and veterinary research. It reviews the statistical methodologies employed in medical and veterinary research. It presents simplified procedures for their application, thereby supporting researchers in designing survey programs that are both scientifically rigorous and economically efficient.

Target population

The population [8] refers to the entire group of human individuals or animals in a certain geographical area at a given point in time that share similar characteristics to which the study findings will be generalized and applied. A population is called finite if its items, units, individuals, or animals can be counted. Sometimes it is not possible to count the population's items, units, individuals, or animals; therefore, it is called an infinite or uncountable population.

Received: *25-06-2026 -* ***Accepted:*** *23-08-2026 -* ***Published:*** *01-09-2026*

https://doi.org/10.54361/ajmas.269902

Accessible population

The accessible (study) population [8] refers to the entire group of human individuals, or animals selected based on pre-defined inclusion and exclusion criteria and is accessible at the time of the study. To generalize the study findings and conclusions to the target population, the accessible population should represent the target population closely. Moreover, it is the population from which the sample will be selected.

Sample

The sample [8] refers to a specific group of items, human individuals, or animals that represent the accessible population and will be included in the study. The sample items should be selected in a way that represents the study population closely. The size of the sample is always less than the size of the accessible population.

Sample size

The sample size refers to the minimum or sufficient number of units, human individuals or animals to be included in the study and is usually represented by n. The main purpose of the sample size calculation is to determine the minimum number of observations that are sufficient to provide an estimate of prevalence (p) with certain precision associated with a level of confidence.

Factors that affect the sample size

Level of significance

The level of significance is denoted by alpha (α) and refers to a threshold that represents the random sampling variation in the results of the study. It refers to the acceptable level of risk that the findings of the study are due to random variation only. For example, a common value of α=0.05 indicates a 5% random sampling variation, i.e., the risk that the study concludes a difference exists when there is no actual difference.

Confidence level

The confidence level refers to the probability that the sample estimate of the population prevalence falls within certain specified limits and is usually denoted by "1- α". A common 95% value is widely used.

Confidence interval

It is an estimated interval computed from the sample data based on pre-defined statistical methodology. This interval is associated with a level of confidence (commonly 95%) that a specified proportion of these intervals contain the true value of the unknown population prevalence in repeated sampling. It is usually denoted by CI.

Expected prevalence

The expected prevalence refers to the historical prevalence that can be obtained from previous studies with similar study design or a pilot study conducted or based on the recommendation of a subject matter expert. If we have a range of values of historical prevalence, then it is better to use the value that is close to 0.5 If no information is available, then we recommend using 0.5, as it leads to a larger (i.e., more conservative) sample size.

Desired precision of an estimate

The desired precision (also called the margin of error, allowable error, or acceptable error in the estimate) refers to the clinically meaningful (acceptable) difference between estimated prevalence and the population's prevalence (which is still an unknown parameter) that we want to estimate. It is an estimated value (commonly 5%) usually set at half the width of the desired confidence interval of the population's prevalence and usually denoted by d. For example, for the desired width of the confidence interval of 0.1, the value of the precision is +/- 0.05 or 5%. Based on our experience, we suggest using a precision of 5% if the expected prevalence of the condition of interest is common (i.e., between 10% and 90%). In the case of rare condition (or diseases), when the expected prevalence is below 10%, then we suggest using a precision value that is half of the expected prevalence. In the case of a very common condition (or diseases), when the expected prevalence is more than 90%, we suggest using a precision value that can be {0.5(1- expected prevalence)}. For example, if the expected prevalence is 0.04, you may choose a precision of =0.02, and if the expected prevalence is 0.98, then you may use a precision of 0.01.

The formula for sample size calculation for prevalence estimation

- Infinite population

In the case of an infinite or very large population, the minimum required sample size n can be obtained using the following formula:

$$n = \frac{\left(Z_{\frac{\alpha}{2}}\right)^2 \times p(1-p)}{d^2} \qquad (1)$$

Received: *25-06-2026 -* ***Accepted:*** *23-08-2026 -* ***Published:*** *01-09-2026*

where n is the required sample size, d is the precision and p is the expected prevalence, $Z_{\frac{\alpha}{2}}$ is the critical value of the normal distribution at α/2 (e.g. for a confidence level of 95%, α is 0.05 and the critical value is 1.96).

- Finite population

In the case of a finite or small population, the sample size n can be corrected as follows:

$$\text{Adjusted sample size} = \frac{(N \times n)}{(N+n-1)}$$

Where N is the number of items, units, human individuals, or animals in the study population, and n is the calculated sample size

Assumption of the normal approximation

The above formula (1) is based on a common rule of thumb that the normal approximation for the binomial works well when np > 5 and n(1-p) > 5. Therefore, a small sample size might not fulfill this assumption. Also, the suggested precision (d) as half of the expected prevalence for a rare condition (or disease) or {0.5(1- expected prevalence)} in case of a very common condition (or disease), will ensure this assumption is met.

Design effect

The above formula (1) is based on simple random sampling where the probabilities of selection are known. However, in the case of collecting sample data using other types of sampling methods such as stratified or multistage sampling, etc. Then the sample size needs to be adjusted for that. One way is to multiply the sample size by a design effect, which is the ratio of the variance of the estimated prevalence based on current sampling selection (different from simple random sampling) to the expected variance of the estimated prevalence had the sample been collected using simple random sampling. One way, we may suggest, is to use a bootstrapping method to estimate the variance. The bootstrap method is a statistical resampling technique used to draw large numbers of samples of the same size, with replacement, from the original sample dataset. Then, the empirical variance across bootstrap samples is used as an estimate of the variance.

Clusters

The above formula (1) is based on the assumption that sampling units are independent. However, in some cases, this assumption is violated, such as when human individuals or animals are clustered within some (physical) clusters or groups such as provinces, hospitals, herds, etc. Therefore, the sample size needs to be adjusted according to the number of clusters.

Software for calculating the required sample size for estimating prevalence

For sample size calculation for estimating prevalence, we recommend using EpiInfo, a software program developed by the United States Centers for Disease Control and Prevention (CDC) and freely available via the link https://restoredcdc.org/www.cdc.gov/epiinfo/pc.html. Once the software is installed, follow the next steps to use this software:

- Open the software,

Figure (1): Epiinfo main menu

Select "STATCALC" for sample size calculation

Received: *25-06-2026 -* ***Accepted:*** *23-08-2026 -* ***Published:*** *01-09-2026*

https://doi.org/10.54361/ajmas.269902

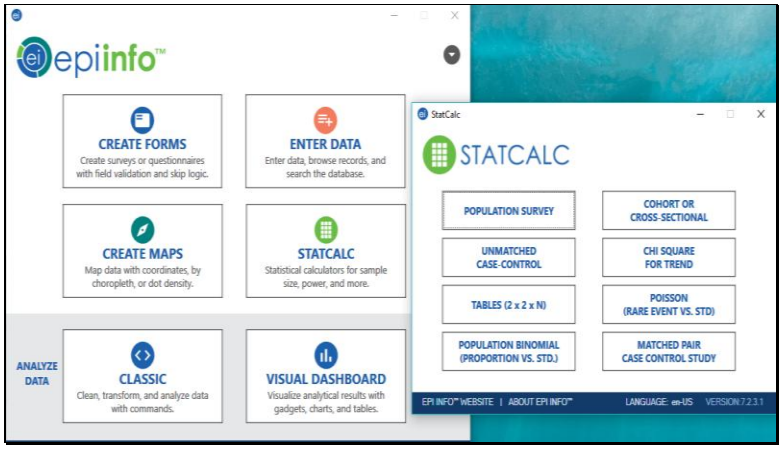

Figure (2): Epiinfo - STATCALC menu

- Select “population survey”,

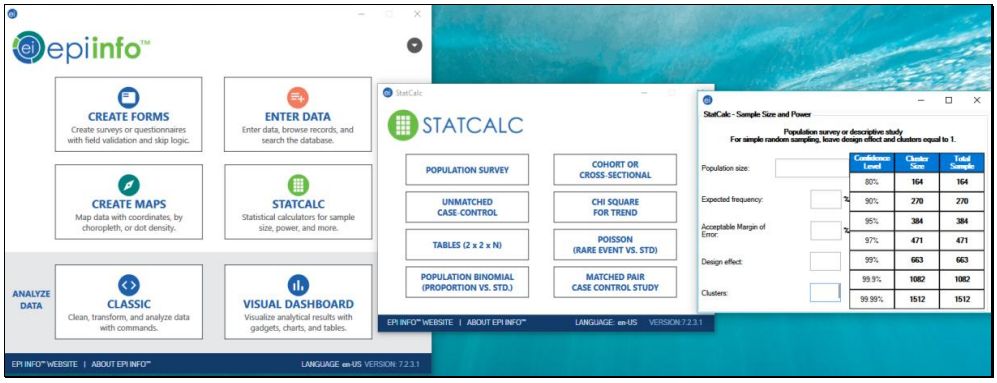

Figure (3): Epiinfo - STATCALC - Inputs

- Submit your data,
  - Population size: It refers to how many people there are in the population from which you are sampling. If the size of the population is not known, enter 1,000,000 as the sample size should not change much for a very large population,
  - Expected frequency: it refers to the expected prevalence
  - Acceptable margin of error: it refers to the precision value
  - Design effect: can be considered only if the selection is not based on simple random sampling
  - Clusters: refers to the number of clusters that may exist
- Check the calculation results.

Example

Let us assume that we are planning a baseline study to estimate the prevalence of a certain disease. From the literature, we found a similar study that showed a prevalence estimate of 30%. We choose a confidence interval of 95%, a precision value of 5%, and the population is very large. Moreover, the sample is to be collected based on simple random sampling.

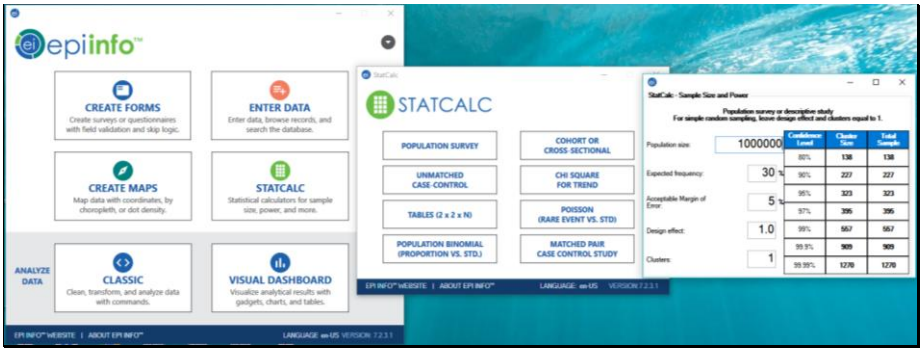

Figure (4): Epiinfo - STATCALC – results

Then, as depicted in figure 4, the results showed the minimum required sample size is 323.

Received: *25-06-2026 -* ***Accepted:*** *23-08-2026 -* ***Published:*** *01-09-2026*

Summary
The calculation of adequate sample size is an important part of any prevalence study and is required by ethics committees for study approval. Yet the process is challenging and depends on assumptions that are rarely precise. The calculated sample size should not be treated as a definitive answer; the final choice must balance statistical guidance with practical constraints, including time, cost, and clinical judgment. Therefore, consulting a professional statistician during study planning is strongly recommended. He/she may assist you to ensure the correct application of formula, considering the assumptions, design effects, and finite population correction.

Conflict of interest
The authors declare no conflicts of interest.

Funding
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

1. Last JM. A dictionary of epidemiology. 4th ed. Oxford: Oxford University Press; 2001.
2. Thrusfield MV, Christley R. Veterinary epidemiology. 4th ed. Wiley-Blackwell; 2018.
3. Friis RH, Sellers TA. Epidemiology for public health practice. 7th ed. Jones & Bartlett Learning; 2021.
4. Buitrago-Garcia D, Salanti G, Low N. Studies of prevalence: how a basic epidemiology concept has gained recognition in the COVID-19 pandemic. BMJ Open. 2022; 12(10).e061497
5. Rothman KJ, Greenland S, Lash TL. Measures of occurrence. In: Rothman KJ, Greenland S, editors. Modern epidemiology. Boston (MA): Lippincott Williams & Wilkins; 2020. p. 32–50.
6. Porta M. A dictionary of epidemiology. 6th ed. Oxford: Oxford University Press; 2014. Available from: https://www.oxfordreference.com/view/10.1093/acref/9780199976720.001.0001/acref-9780199976720
7. Pourhoseingholi MA, Vahedi M, Rahimzadeh M. Sample size calculation in medical studies. Gastroenterol Hepatol Bed Bench. 2013;6(1):14–7.
8. Khaled Fahim N, Negida A. Sample size calculation guide – Part 1: how to calculate the sample size based on the prevalence rate. Adv J Emerg Med. 2018;2(4).
9. Naing L, Winn T, Rusli BN. Practical issues in calculating the sample size for prevalence studies. Arch Orofac Sci. 2006;1:9–14.
10. Naing L, Nordin RB, Abdul Rahman H, Naing YT. Sample size calculation for prevalence studies using ScaleX and ScalaR calculators. BMC Med Res Methodol. 2022;22(1):209
11. Dupont WD, Plummer WD. Power and sample size calculations: a review and computer program. Control Clin Trials. 1990; 11(2):116–28.
12. Dattalo P. A review of software for sample size determination. Eval Health Prof. 2009;32(3):229–48.
13. Kumar A, Dogra S, Kaur A, Modi M, Thakur A, Saluja S. Approach to sample size calculation in medical research. Curr Med Res Pract. 2014;4(2):87–92.
14. Lenth RV. Some practical guidelines for effective sample size determination. Am Stat. 2001;55(3):1–11.
15. Jones SR, Carley S, Harrison M. An introduction to power and sample size estimation. Emerg Med J. 2003; 20(5):453–8.
16. Zrineh A, Al-Usta M, Alwawi A. Sampling methods and sample size determination in clinical research: an educational review. J Gen Fam Med. 2026; 27(1):e70096.
17. Althubaiti A. Sample size determination: a practical guide for health researchers. J Gen Fam Med. 2022; 24(2):72–8.
18. Said A, Abuagela M, Elyounsi N, Dakhil A, Elansari A. Prevalence of hepatitis B and C infection among patients in Tripoli Central Hospital. Alq J Med App Sci. 2022; 5(2):406–10.
19. Abdulwahed E, Rhoma H, Sultan M. Prevalence of anemia in a sample of elderly subjects from Tripoli, Libya. Alq J Med App Sci. 2021; 4(2):238–44.
20. Murshid M, Zahmoul T, Masaoud E, Gamag E, Almabrouk A. Prevalence of hydatidosis in sheep, goats, cattle, and camels slaughtered in selected abattoirs in Sabratha City, Libya. Alq J Med App Sci. 2026; 9(3): 613–621.
21. Aljedaemi A, Hwas R, Himeer S, Ghabasa A. Prevalence of oral habits (finger sucking) among preschool children in Zliten, Libya. Alq J Med App Sci. 2025; 8(4):2761–4.
22. Mohmmed S, Ekhnefer A, Fadel A, Sharif M. Prevalence of some parasitic infestations in Lagocephalus sceleratus in Eastern Libya. Alq J Med App Sci. 2023; 6(2):608–16.

Received: *25-06-2026 - **Accepted:** 23-08-2026 - **Published:** 01-09-2026*