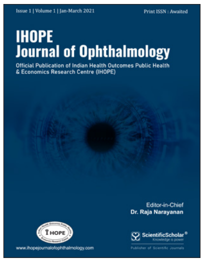
ScientificScholar®
Knowledge is power
Publisher of Scientific Journals

www.ihopejournalofophthalmology.com

IHOPE Journal of Ophthalmology



Review Article

Sample size and its evolution in research

Sai Prashanti Gumpili[1,2], Anthony Vipin Das[1,2]

Departments of [1]EyeSmart EMR and AEye, [2]Indian Health Outcomes, Public Health and Economics Research Center, L V Prasad Eye Institute, Hyderabad, Telangana, India.

***Corresponding author:**
Anthony Vipin Das,
Department of eyeSmart
EMR and AEye L V Prasad
Eye Institute, Hyderabad,
Telangana, India.

vipin@lvpei.org

Received : 14 July 2021
Accepted : 28 October 2021
Published : 07 January 2022

DOI
10.25259/IHOPEJO_3_2021

Quick Response Code:

ABSTRACT

Objective: Sample size is one of the crucial and basic steps involved in planning any study. This article aims to study the evolution of sample size across the years from hundreds to thousands to millions and to a trillion in the near future (H-K-M-B-T). It also aims to understand the importance of sampling in the era of big data.

Study Design - Primary Outcome measure, Methods, Results, and Interpretation: A sample size which is too small will not be a true representation of the population whereas a large sample size will involve putting more individuals at risk. An optimum sample size needs to be employed to identify statistically significant differences if they exist and obtain scientifically valid results.

The design of the study, the primary outcome, sampling method used, dropout rate, effect size, power, level of significance, and standard deviation are some of the multiple factors which affect the sample size. All these factors need to be taken into account while calculating the sample size. Many sources are available for calculating sample size. Discretion needs to be used while choosing the right source. The large volumes of data and the corresponding number of data points being analyzed is redefining many industries including healthcare. The larger the sample size, the more insightful information, identification of rare side effects, lesser margin of error, higher confidence level, and models with more accuracy. Advances in the digital era have ensured that we do not face most of the obstacles faced traditionally with regards to statistical sampling, yet it has its own set of challenges. Hence, considerable efforts and time should be invested in selecting sampling techniques which are appropriate and reducing sampling bias and errors. This will ensure the reliability and reproducibility in the results obtained. Along with a large sample size, the focus should be on getting to know the data better, the sample frame and the context in which it was collected. We need to focus on creation of good quality data and structured systems to capture the sample. Good data quality management makes sure that the data are structured appropriately.

Keywords: Sample size, Power of study, Effect size, Study design, Big data

Sample size and its importance in research

In statistics, the term "population" is defined as an entire group of events or items which is of interest to our research question. Since it is not feasible to study the entire population, a subset of the population is chosen to adequately represent the same. This subset is defined as the sample. Every individual in the chosen population should have an equal chance of being selected. Sample size which is typically denoted by signifies the total number of observations or participants included in a study.

One of the key steps involved in a clinical study is calculation of the sample size. If the sample size is very small, the sample is not a true representation of our population and the results obtained cannot be extrapolated to the entire population. In addition, the differences between the groups will not be detected if the sample size is too small as they state that "absence of evidence is no evidence of absence." If our sample size is larger than required, it involves putting more individuals at risk of that particular intervention, which is highly unethical. Small differences manifest into clinically significant differences which can potentially be misguiding and may lead to grave consequences like failure in making the right decision about treatments.[1] It is also a huge waste of finances, human resources, and time. Further, saturation is defined as the point after which collection of any data will no longer yield any new results.[2] Saturation is dependent on various factors such as homogeneity or heterogeneity of the population being studied, criteria used for selection, financial resources available, and timelines set. All these factors need to be carefully considered before the start of any study.

The central limit theorem states that irrespective of the distribution of the population, as the sample size increases the distribution of sample approximate a normal distribution ("bell curve").[3] Therefore, as sample size increases, mean and standard deviation of the sample will be closer in value to the mean (μ) and standard deviation (σ) of the population.

An optimum sample size is the minimum number of individuals needed to identify any statistically significant difference if it truly exists and a means by which we attain results which are valid scientifically. A fine balance needs to be maintained and an optimum sample size needs to be arrived at. Therefore, the sample size lies at the heart of any study.

FACTORS AFFECTING SAMPLE SIZE CALCULATION

Sample size impacts the precision of our estimates and the power of our study to arrive at any conclusion. The sample size for any study depends on the design of the study, the primary outcome that is being studied (continuous or binary), one-tailed or two-tailed test, sampling method used, dropout rate and the measures of outcome such as effect size, power, level of significance, and standard deviation.[4-7] Descriptive studies such as surveys, case-series, case-reports and questionnaires require a larger sample size when compared to analytical studies. The sample size in methods of qualitative research is often smaller than that used in quantitative research.[8] Observational studies need more samples than experimental studies.[9]

As the effect of the size which has to be detected decreases, the sample size increases and vice versa. If the population being studied is more homogenous, it implies lesser standard deviation, and hence smaller sample size. More heterogeneous population entails a large sample size to get accurate results.

Before the start of the study, we set an acceptable value of level of significance (*P*-value). *P* = 0.05 indicates that there is a 5% probability that the observed results are due to chance and not due to the intervention (False positive result, Type I error). In other words, 5 out of 100 times we accept that there is a difference when in fact there is none. As the level of significance decreases, the sample size increases. In an exactly converse situation, there are chances of failing to detect the difference even when it is actually present (False negative result and Type II error). The probability of committing a type II error is called beta (β). (1-β) is called power, which is defined as probability of failing to detect a difference even though it is there. As the desired power value increases, the sample size also increases. The two most applicable type of power analyses are a priori and *post hoc* power analysis. As the name suggests a priori analysis is performed before the experiment is conducted as a part of the process of research design.[10] *Post hoc* analysis is performed after the study has been conducted.

During the calculation of sample size, we need to accept the risk of a false negative or a false positive result, if not we would need a sample size which is infinitely large. It was observed that for most trials whose results were negative the sample size was not large enough. Hence, reporting of statistical power and sample size need to be improved.[11,12] Calculation of sample size can be guided by pilot studies undertaken, previous literature, and past clinical experiences. Sample size calculations require careful judgment and a compromise between strict criteria and practicality of access. [Table 1] summarizes the effect of factors on the required sample size.

SOFTWARE FOR ESTIMATING SAMPLE SIZE

Sample size calculating software has made it easier and simpler to estimate sample size. The software for estimating sample size varies with the type of study design. Existing statistical software such as Statistical Package for the Social Sciences, SAS, Stata, and R has the methods for determining sample size incorporated into them. Exclusive software such as PASS, G*Power, Power and Precision, Russ Lenths power, Minitab, and SampSize is also available for calculating sample size.[9,13]

Table 1: Impact of factors on sample size.

Factor	Magnitude/type	Required sample size
Effect of size to be detected	Small	Large
Population	Homogenous	Small
Level of significance (P-value)	Small	Large
Power	Small	Small

Most of the software used for estimation of sample size have limited validity as usually they use a single formula. Any error can further mislead the researcher and the results of the study and hence it is essential that these errors are controlled. A review by Abbassi *et al*. which studied the accuracy of online sample size calculators showed that most of the sites merely calculated the sample size for estimating proportions and considered 50% as a fixed value in the formula for calculation.[14] The results were not accurate for the examples which were considered. Discretion needs to be exercised while using online calculators and the researcher should be well aware of the research design, the outcome, common errors, and the method and parameters being used to estimate the sample size.

HYPOTHESIS RESEARCH AND NON-HYPOTHESIS RESEARCH AND ITS RELATION WITH SAMPLE SIZE

Most of scientific research is driven by hypothesis, which is described as an educated guess based on observations made and prior knowledge on the same. The null and alternative hypotheses are two mutually exclusive statements about a population. The null hypothesis (H_0), states the opposite of what is expected by the researcher and the alternative hypothesis (H_a) states the results which are expected by the researcher. Hypothesis testing uses data to determine whether to accept or reject the null hypothesis. Inability to reject the null hypothesis may also mean that the evidence required to reject it is not sufficient.

Conventionally, research in many sectors was pursued through hypothesis-driven investigations. Lately, research which is not driven by hypothesis is gaining momentum. Research not driven by hypothesis may include model and database development, high throughput genomics, engineering, and biology.[15] This way of research allows data to lead the way and we can embark on bolder journeys which are not constrained by our existing knowledge. High performance computing and large sample size are an important catalyst which will fuel the journey of hypothesis free research and open new avenues.

EVOLUTION OF SAMPLE SIZE IN DIFFERENT SECTORS

The data volumes are exploding, more data have been created in the past 2 years than in the entire previous history of the human race. Big data is encompassing all sectors right from healthcare, governance, finance, psychiatry, remote sensing, manufacturing, education, etc.

Companies across various sectors of industry are leveraging big data to ensure data driven decision making. The large volumes of data and the corresponding number of data points being analyzed are redefining many industries.

A review by Button *et al*. indicated how small sample size undermines reliability in the field of neuroscience. Their results indicated that the median statistical power in neuroscience is only 21%. When a study which is underpowered discovers a true effect, it is likely that the estimate of the size of the effect provided will be inflated. This is called as the winner's curse. Hence, if a study with small sample size is the only source of evidence it is difficult to have confidence in that evidence. In spite of scientists pursuing smaller effects the average sample size has not changed over time in the field of neuroscience. The advances in the analysis techniques and time taken have not been reflected in the aspects of study design and research in the field.[16,17] Unreliable research is wasteful and inefficient and hence there is an ethical dimension to low power.

On the contrary, in the field of empirical finance, vast majority of the studies use large sample sizes and use the conventional thresholds for statistical significance which may lead to a large sample bias.[18] Therefore, suitable thresholds for statistical significance have to be employed for a given sample size.

Estimates suggest that by better integrating big data, healthcare could save as much as $300 billion a year. The health-care industry is rapidly following suit given the advent of electronic medical record systems which capture the data in a structured format.[19-21] This is a highly welcome change and will ensure the reliability and reproducibility in the field of healthcare.

IMPORTANCE OF SAMPLING IN THE ERA OF BIG DATA

Big data refers to datasets that are too large or complex for traditional data processing applications. At present, the pace at which data are being generated in all fields on a day to day basis is rapid. Due to this advancement in the current digital era and decrease in the costs of collection of data and processing, we have overcome few obstacles which were faced traditionally with respect to statistical sampling.

Large datasets have their fair share of advantages. The data can be used to study rare events given its volume. If any outlier is present in our sample, large sample size refrains us from making any statistically misguided decisions. Margin of error is a measure which indicates to what extent our results will differ from the actual value of the population. The relationship between margin of error and sample size is inverse. The larger the sample size, the smaller the margin of error. Lower margin of error also signifies a higher confidence level in the obtained results. Increasing the sample size after a certain point provides a diminishing return as the increase in accuracy turns out to be negligible.[22] Bringing down the margin of error below a certain threshold is rarely beneficial.

In fact, it would be ideal to spend the existing resources on reducing sources which are responsible for bias. [Figure 1] illustrates the relationship between sample size and margin of error.

LARGE SAMPLE SIZE AND POTENTIAL FOR BIAS: HOW TO TREAD WITH CAUTION

While advances in the digital era have ensured that we do not face most of the obstacles that we faced traditionally with regard to statistical sampling, it has its own share of challenges. In spite of the sample size being large, our data might yet be a representative of only a part of the population and not the whole. Sampling bias is one of the major factor which affects the performance of our model and the obtained results.[23] It should be ensured that the training and test datasets which are used to train and test our model should mirror the same distribution which is reflected in the entire dataset.

“Big data Hubris” is the notion that big data analytics can be used as a substitute rather than a supplement to traditional means of analytics. Google flu trend and poll of Literary Digest of the 1936 United States Presidential election are classic examples of this. Before the 1936 election, the poll by the literary digest magazine had always correctly predicted the winner. In 1936, the poll concluded that the Republican candidate, Governor Landon was likely to win by a majority against the incumbent President Franklin D. Roosevelt. On the day of the results, Roosevelt won the elections by a landslide. The magazine had polled a sample of over 2 million people based on car and telephone registrations. However, there was a problem with the sample frame. That was the time of the Depression and not everyone could afford a car or a telephone. In spite of the large sample size, there was a discrepancy in the sample frame. In 2008, Google launched Google Flu Trends (GFT) to predict the spread of influenza across the US. GFT consistently overestimated visits related to flu and was highly inaccurate during the peak flu season when it could have proven to be most useful. This reiterates the fact that a sample frame which is incorrect could potentially destroy a study irrespective of the sample size. Hence, considerable efforts and time should be invested in selecting sampling techniques which are appropriate rather than amassing data of the whole population who are accessible.

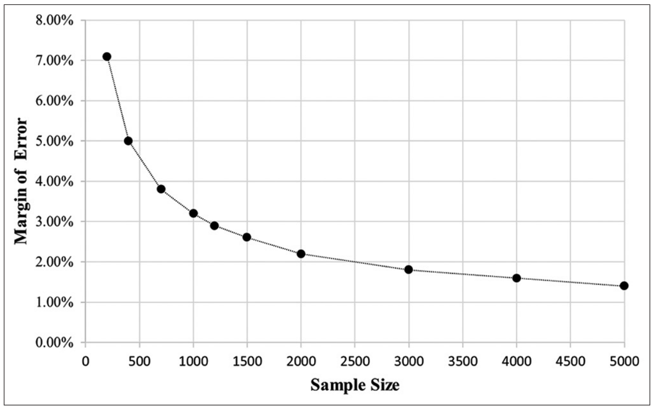


Figure 1: Relationship between sample size and margin of error.

When the sample size is small, it is easier to check and control the quality of data and spot errors if any. In case of a larger sample size greater effort and time should be spent to check the accuracy and quality of data and identifying outliers or missing values before we perform any further analysis.

Studies involving large sample size can identify effects which are significant but may be inconsequential.[23] If we were to compare two trials one with a smaller sample size and another with a larger sample size and we assume the level of significance to be 0.05 in both the cases, the effect size in case of the study with a smaller sample size would be significantly larger than the one with the larger sample size to achieve the same level of significance. Even though studies with larger sample sizes have many advantages, we should remain cautious of the fact that the effect of the treatment can be quite modest. Larger sample size will never be able to compensate for the other challenges which are faced in analytics. Hence, our main focus apart from increasing the sample size should also involve reduction in sampling bias, and other errors.[23]

As with any type of data, the data captured in Electronic Medical Records are only as good as the information captured by the systems. Even though the sample size is large, the system capturing the data needs to be robust and should ensure that the data is captured in a uniform format with structured forms and databases. In the future, all kinds of data pertaining to behavior, environment, and other important aspects need to be captured to expand the scope of variables which can be included to study their effect on the outcomes.

The larger the sample size it implies that there is more insightful information, lesser margin of error, higher confidence level and models with more accuracy with respect to how they have been used. Identification of rare side effects due to medication and outcomes in people with diseases which are rare will also greatly benefit from a larger sample size.

We need to focus on creation of good quality data and structured systems to capture the sample. Good data quality management makes sure that the data are structured appropriately. Maintaining high levels of data quality enables organizations to reduce the cost of identifying and fixing bad data in their systems. It also helps prioritize and ensure the best use of resources.

Along with having a large sample size, the focus should be on getting to know the data better, the sample frame and the context in which it was collected. Exploratory data analysis as an initial step will help in unearthing all that the data have to reveal and also identify the outliers and missing values.

We have witnessed the evolution of sample size from hundreds to thousands to millions and it will continue evolve to trillion and beyond (H-K-M-B-T) with rapid growth of data and exponential growth in technology. We are hopeful that the generation of new knowledge and data will open up the frontiers of research, development and growth.

Sources of literature review include peer-reviewed articles, books, and conference papers.

Author contributions

Concept, A.V.D.; Design, A.V.D., G.S.P.; Literature search, G.S.P.; Manuscript preparation, G.S.P., A.V.D.; Manuscript editing, A.V.D., G.S.P.; Manuscript review, A.V.D. G.S.P.

Declaration of patient consent

Patient's consent not required as there are no patients in this study.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

REFERENCES

1. Faber J, Fonseca LM. How sample size influences research outcomes. Dental Press J Orthod 2014;19:27-9.
2. Faulkner SL, Trotter SP. Data saturation. In: The International Encyclopedia of Communication Research Methods. New Jersey, United States: Wiley; 2017. p. 1-2.
3. Trotter HF. An elementary proof of the central limit theorem. Arch Math 1959;10:226-34.
4. Charan J, Biswas T. How to calculate sample size for different study designs in medical research? Indian J Psychol Med 2013;35:121-6.
5. Noordzij M, Dekker FW, Zoccali C, Jager KJ. Sample size calculations. Nephron Clin Pract 2011;118:c319-23.
6. Kirby A, Gebski V, Keech AC. Determining the sample size in a clinical trial. Med J Aust 2002;177:256-7.
7. Phillips A, Campbell M. Using aspects of study design in sample size estimation. J Biopharm Stat 1997;7:215-26.
8. Dworkin SL. Sample size policy for qualitative studies using in-depth interviews. Arch Sex Behav 2012;41:1319-20.
9. Chander N. Sample size estimation. J Indian Prosthodont Soc 2017;17:217-8.
10. Farrokhyar F, Reddy D, Poolman RW, Bhandari M. Why perform a priori sample size calculation? Can J Surg 2013;56:207-13.
11. Abdulatif M, Mukhtar A, Obayah G. Pitfalls in reporting sample size calculation in randomized controlled trials published in leading anaesthesia journals: A systematic review. Br J Anaesth 2015;115:699-707.
12. Noordzij M, Tripepi G, Dekker FW, Zoccali C, Tanck MW, Jager KJ. Sample size calculations: Basic principles and common pitfalls. Nephrol Dial Transplant 2010;25:1388-93.
13. Dattalo P. A review of software for sample size determination. Eval Health Prof 2009;32:229-48.
14. Abbassi M, Emamzadeh-Fard S, Yoosefi-Khanghah S, Mohammadi-Vajari M-A, Taee F, Meysamie A. Sample size calculation on web, can we rely on the results? J Med Stat Inform 2014;2:3.
15. Rigoutsos I, Floratos A. Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. Bioinformatics 1998;14:55-67.
16. Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, *et al*. Power failure: Why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 2013;14:365-76.
17. Szucs D, Ioannidis JP. Sample size evolution in neuroimaging research: An evaluation of highly-cited studies (1990-2012) and of latest practices (2017-2018) in high-impact journals. Neuroimage 2020;221:117164.
18. Michaelides M. Large sample size bias in empirical finance. Finance Res Lett 2021;41:S1544612320316494.
19. Das AV, Kammari P, Vadapalli R, Basu S. Big data and the eyeSmart electronic medical record system-an 8-year experience from a three-tier eye care network in India. Indian J Ophthalmol 2020;68:427-32.
20. Donthineni PR, Kammari P, Shanbhag SS, Singh V, Das AV, Basu S. Incidence, demographics, types and risk factors of dry eye disease in India: Electronic medical records driven big data analytics report I. Ocul Surf 2019;17:250-6.
21. Das AV, Podila S, Prashanthi GS, Basu S. Clinical profile of pterygium in patients seeking eye care in India: Electronic medical records-driven big data analytics report III. Int Ophthalmol 2020;40:1553-63.
22. Horowitz I. The diminishing returns to sample information in the beta-binominal process. Z Nationalökonomie 1972;32:493-500.
23. Kaplan RM, Chambers DA, Glasgow RE. Big data and large sample size: A cautionary note on the potential for bias. Clin Transl Sci 2014;7:342-6

How to cite this article: How to cite this article: Gumpili SP, Das AV. Sample size and its evolution in research. IHOPE J Ophthalmol 2022;1:9-13.