Algorithm for Sample Availability Prediction in a Hospital based Epidemiological Study: Spreadsheet-based Sample Availability Calculator

Amrit Sudershan
University of Jammu
Parvinder Kumar (parvinderkb2003@yahoo.co.in)
University of Jammu
kanak Mahajan
University of Jammu

Research Article

Keywords: probability, statistical power, sample size, quantitative sampling, population

Posted Date: June 30th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-664438/v1

Version of Record: A version of this preprint was published at Scientific Reports on February 3rd, 2022.
See the published version at https://doi.org/10.1038/s41598-021-03399-1.

Algorithm for sample availability prediction in a hospital based Epidemiological study: Spreadsheet-based Sample Availability calculator

Amrit Sudershan[1], Kanak Mahajan[1], Parvinder Kumar[*]

[1]Institute of Human Genetics, Jammu University, Jammu

Abstract

Looking at population's behavior by taking samples is quite uncertain due to its big and dynamic structure and unimaginable variability. All quantitative sampling approaches aim to draw representative sample from the population so that the results of the studying samples can then be generalized back to the population. The probability of detecting a true effect of a study largely depends on the sample size and if taking small samples will give lowers statistical power, thus having higher risk of missing a meaningful underlying difference. There have lot of online sample size calculators which are based on population size, allele frequency which tell us about the number of samples required for the study but none will help in setting a threshold for the availability of the sample from a single hospital in a particular period. This study aims to provide an efficient calculation method for setting a threshold for the availability of samples during a specific period of a research study which is an important question to be answered during the research study design. So we have designed a spreadsheet-based sample availability calculator tool implemented in MS-Excel 2007.

Keywords: probability, statistical power, sample size, quantitative sampling, population

Introduction

The transmission of genetic information from one generation to the next generation is a law of probability and population genetics take this concern to an entire population (**Relethford, 2012**). It makes us understand human variation, their origin, and impacts of these on population by linking medical and evolutionary themes (**Conrad & Hurles, 2007**).

In Epidemiological study, apart from the "Clinical investigation" we pile up the facts related to the diseases by collecting history (assessment questionnaires) from individuals and establish the cause of a disease (**Zaccai, 2004**;**Attal et al., 2018**) then estimates the individual risk of diseases and gives the chance of avoiding its risk of disease. The epidemiological study is categorized under two different types i.e., Observational study and Experimental study where the former one is further divided into three different classes including case-control study, cohort study, and cross-sectional study (**Mann, 2003).** These retrospective study design help in determining whether exposure is associated with an outcome or not in a population by comparing two groups of matched cases and controls (Case-control study design) **(Lewallen et al., 1998)** and establish the risk factor of the diseases.

Population data are analyzed by different arms of science (**Bhopal, 2016**) and used different terms to define the population. As per "biologist", the number of all the organisms of the same group or species capable of interbreeding in a particular geographical area is called the population (**Krieger., 2012**). In this article we are strictly restricted to statistics, therefore, a population is an entire pool of people or events (hospital visits, small strata), from where fraction or percentage of a group is drawn which represents the statistical sample (**Figure 1A**) (**Krieger et al., 2012**).

Population, a big and dynamic structure with unimaginable variability, so looking at the population's behavior by taking the whole population as a sample is quite uncertain and this is because of the restricted amount of time, ethical irrelevant, and money limitation. The quantitative sampling approach "quantifying the difference in effect, but unable to answer the question of *how it affects"* (***Kim*, 2015; Noyes et al., 2019)** draws a representative sample through a random sampling approach from the considered population. The probability of success

of a research study depends on the sufficient study sample size to produce clinically relevant difference (**Pourhoseingholi et al., 2013; Biswas & Charan, 2016**) but sometimes, not having a well-designed research study design tend to recruit a small sample size which increases the chance of assuming as true a false premise (**Faber & Fonseca, 2014**; **Greenland et al., 2016; Noordzij et al., 2010**). Having too large a sample size will become more expensive than necessary and also much time-consuming (**Atkinson& Columb, 2016**). Studying with Sample size calculations relates to the probability of a study correctly detecting a true effect **(Goodall et al., 2009)** to specify estimated parameters of the study design (**Machin et al., 2018**).

There are a lot of online sample size calculators which are based on population size (https://www.calculator.net/sample-size; https://www.surveymonkey.com/mp/sample-size-calculator/; https://www.surveysystem.com/sscalc.htm; http://www.raosoft.com/samplesize.html; https://www.qualtrics.com/blog/calculating sample-size/), prevalence based (http://sampsize.sourceforge.net/iface/) and also on allele frequency (http://osse.bii.a-star.edu.sg/calculation1.php) tell us about the number of samples required for the study but none will help in setting a threshold for the availability of the sample from a single hospital in a particular period. Different from population-based studies is a hospital-based study (**Li et al., 2011**) which provides strata from where the patients were identified regardless of the population from which they arise (**Lunet et al., 2009**).

In this article, an algorithm is designed which is useful for calculating the probability of availability sample number per year which will help to meet your required sample size to detect absolute power. It will help in setting a threshold for the availability of the sample from a single hospital and may inform about the exigency, which tells the researcher whether sampling needs to be done from the more than one hospital or region.

Material method

In this research article, we are tried to solve a problem that most of the researcher's faces during their pre-study design "estimating how much time will take to cover the required sample size". This sample availability calculator based on the "probability" will set a threshold for the availability of the sample from a single hospital and is implemented in MS-Excel (**Figure 2**) and can run on MS-Excel 2000-2007 on MS-Windows 2000, XP, Vista, and Windows 7 beta.

Algorithm

An essential tool in statistics is the probability which measures, "how much chance that a given event will occur" **(Paola et al., 2018) and which** have been significantly evolved for last decades.

To solve the problem, this mathematical model which is an algorithm-based (set of steps to solve the problem) expressed in the formula (symbolically to construct a relationship between given quantities) helps to link every value of a variable to the probability.

First, we will sort out the number of patients from the given population using equation I where we use the previous knowledge of prevalence of a disease (figure 1A-1B). In this article we use a simulated diseases study where we use simulated prevalence.

$$\boldsymbol{S = n \times \frac{Pr}{Pc}}$$

Where "**S**" representing the sample availability per year from the total population, "**n**" is the total number of population/ population or size of the population, "**Pr**" prevalence of the diseases, and "**Pc"** is the percentage (100%) represents the whole population **(figure 1A)**. Once we find out the number of diseased individual we do a Uniform distribution (U), where samples are equally distributed to the **default** number of the hospital. Hospital numbers can vary with different population size (positive correlation) therefore act as a continuous variable and can be changed (Δ) from population to population, therefore this is managed by setting a threshold for hospital number (constant number) 100 (X=100).

For equal distribution to "X", equation – II is used

$$\boldsymbol{U = S \times \frac{1}{\Delta X}}$$

Where “U” is the uniform distribution of samples to a variable representing as “X” and “S” is sample availability per year from the total population from equation-I.

Introduction of variable: It is very important to introduce variables that may affect the sample numbers available and thus increasing the probability and overcome the bias that may be created during stratified (hospital) sampling. There are several reasons to introduce variables which may reduce the variability which includes the number of cases who did not seek medical care or may some cases are seen elsewhere geographically, there may be death or remission of patients before diagnosis, etc. **(Taber et al., 2015) (Figure 1C).** Therefore, excluding these cases which may be responsible for creating bias and may affect the result, we equally distribute the stratified population into 5 different variables (X'). Thus dividing into smaller groups reduces variance and completes the sampling process.

$$S' = U \times \frac{1}{X'} \quad \textbf{Eq}$$

Where **S'** represents “sample available for sampling” after sub stratification in 5 different layers. Each variable representing cases who did not seek medical care (X'1), cases are seen elsewhere geographically (X'2), cases misdiagnosed, death or remission before diagnosis (X'3), unknown variable (X'4), and cases available for the case-control study (X'5) therefore X'= 5 **(Figure1C).**
Selecting an unknown variable is to remove the biases created by a variable that can’t be defined but may have an effect on our experimental data (**confounding variable**). The reason for the equal distribution is that the chance of distribution and selecting samples will remain the same for all if we chose uneven distribution then we cannot say its probability because it will become “definitely/ surely”. We need the “chance of outcome” not the “definitely it will be the outcome” because it is not applicable for so big a population which dynamic and changeable.
By combining all the equations (1, 2, and 3) we get,

$$A = n \times \frac{Pr}{Pc}\left(\frac{1}{\Delta X}\right)\left(\frac{1}{X'}\right)$$

Where “A” represents the availability of the sample per year at a particular hospital [after equally distributed to each variable (X)'].

As we are sampling from the real world situation where there is a limitation of works, time, patients, etc. so to overcome all these real-world situations we did some tricky calculation for the

time and the day calculation which is important to reduce the chance of bias and so increasing the probability. For being with the smallest probability we chose 1-hour representation, which means only we have limited accesses to the patients, and also we exclude Sunday from the week because the hospitals are not open on Sunday.

Simulated example (Graph) (Figure 3)

More patients more will come and thus the chance of getting patients will also increase with each day (percentage) 1 hour being the lowest probability and maximum probability with maximum hours of the day.

$$\boldsymbol{R = \frac{A}{Yd} \times \left(\frac{1}{Hd} \times dM \times dY\right)}$$

Where "**R**" is the value after refining, "**A**" is the availability of the sample per year at a particular hospital, "**Yd**" is the days in the year (365 days), "**Hd**" is hours in the days (24), "**dM**" is days in the month [excluding Sunday (26), and "**dY**" is now a total month in a year (12).

The probability of sample availability per year is calculated by using equation VI,

$$\boldsymbol{P = \frac{R}{A}\left(\frac{Pc}{100}\right)}$$

Where P is the probability of availability of sample per year, R is the data value after refining, A is the availability of the sample per year at a particular hospital, and Pc is the percentage (100%). **Probability for 1 day with time managed (1 hrs to 10).** It is important to note that if we increase the sampling time for a specific day the chance of availability of sample will **increase (Figure 4).**

Discussion

In epidemiology studies, case-control studies help to determine if an exposure is associated with an outcome (i.e., disease or condition of interest) or not, and (**Lewallen et al., 1998**) are one of the most frequently used retrospectives. In a population-based case-control study, cases are ascertained from a diseases registry or from hospital networks from a specific geographical area within a specified period (**Schlesselman, 1982)** to study the associate risk factor and estimate the effect of exposure on the risk of diseases.

But the question is how many numbers of samples from the population are required to draw out the meaningful difference? and the probability of detecting a true effect of a study for a population that is very dynamic with unimaginable variability largely depends on the sample size. If we take a small sample size that will give lowers statistical power, higher risk of missing a meaningful underlying difference. Here biomedical statistics have come under increased scrutiny (**Biswas & Charan, 2016).**

There are a lot of online sample size calculators which are based on population size (https://www.calculator.net/sample-size; https://www.surveymonkey.com/mp/sample-size-calculator/; https://www.surveysystem.com/sscalc.htm; http://www.raosoft.com/samplesize.html; https://www.qualtrics.com/blog/calculating sample-size/), prevalence based (http://sampsize.sourceforge.net/iface/) and also on allele frequency (http://osse.bii.a-star.edu.sg/calculation1.php) tell us about the number of samples required for the study but none will help in setting a threshold for the availability of the sample from a single hospital in a particular period.

The designed algorithm is useful for calculating the probability of availability sample number per year which will help to meet your required sample size to detect absolute power. It will help in setting a threshold for the availability of the sample from a single hospital and may inform about the exigency, which does to do researcher whether sampling needs to be done from the more than one hospital or region.

A well-designed spreadsheet in MS-Excel 2000-2007 will help in the calculation which is set accordingly to the algorithms that are stated above. It can run on MS-Excel 2000-2007 on MS-Windows 2000, XP, Vista, and Windows 7 beta. We just have to enter the total population size, the prevalence, the total hospital will remain to the defaults if want to change its editable, all these will provide the exactly equally distributed samples accordingly to the time mentioned. The sample availability tool in MS-Excel is readily available to any researcher and wishes to use it for non-commercial purposes without any restriction.

Conclusion

This sample availability calculation tool will help in finding the number of samples that are available during the specific period of your research study and thus meet your required sample size to detect absolute power. This sample availability calculation is well-designed in an excel spreadsheet (MS-Excel 2000-2007) (**Figure 2**) which can run on MS-Excel 2000-2007 on MS-

Windows 2000, XP, Vista, and Windows 7 beta and will use it for non-commercial purposes without any restriction.

Reference

1. **Attal N, Bouhassira D, & Baron R (2018).** Diagnosis and assessment of neuropathic pain through questionnaires. **The Lancet Neurology, 17(5): 456-466.**
2. **Bhopal RS (2016)**. Concepts of epidemiology: integrating the ideas, theories, principles, and methods of epidemiology. **Oxford University Press.**
3. **Charan J & Biswas T (2013).** How to calculate sample size for different study designs in medical research?. **Indian journal of psychological medicine, 35(2): 121–126.**
4. **Columb M O, & Atkinson MS (2016).** Statistical analysis: sample size and power estimations. **Bja Education, 16(5): 159-161.**
5. **Conrad DF, & Hurles ME (2007).** The population genetics of structural variation. **Nature genetics, 39(7): S30-S36.**
6. **Di Paola G, Bertani A, De Monte L and Tuzzolino F (2018).** A brief introduction to probability. **Journal of thoracic disease, 10(2): 1129.**
7. **Faber J, & Fonseca LM (2014).** How sample size influences research outcomes. **Dental press journal of orthodontics, 19(4): 27–29.**
8. **Goodall EA, Moore J, & Moore T (2009).** The estimation of approximate sample size requirements necessary for clinical and epidemiological studies in vision sciences**. Eye, 23(7): 1589-1597.**
9. **Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, and Altman DG (2016).** Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations**. European journal of epidemiology, 31(4): 337-350.**
10. **Kim, H.Y., 2015.** Statistical notes for clinical researchers: Type I and type II errors in the statistical decision. **Restorative dentistry & endodontics, 40(3): 249-252.**
11. **Krieger N, (2012)**. Who and what is a "population"? Historical debates, current controversies, and implications for understanding "population health" and rectifying health inequities. **The Milbank Quarterly, 90(4):.634-681.**

12. **Krieger N. (2012).** Who and what is a "population"? Historical debates, current controversies, and implications for understanding "population health" and rectifying health inequities. **The Milbank quarterly, 90(4): 634–681.**
13. **Lewallen S and Courtright P (1998).** Epidemiology in practice: case-control studies. **Community Eye Health, 11(28): 57.**
14. **Lunet N and Azevedo A (2009).** On the comparability of population-based and hospital-based case-control studies. **Gaceta sanitaria, (23): 564-564.**
15. **Machin D Campbell MJ Tan SB and Tan SH (2018).** Sample sizes for clinical, laboratory and epidemiology studies. **John Wiley & Sons.**
16. **Mann CJ (2003).** Observational research methods. Research design II: cohort, cross sectional, and case-control studies. **Emergency medicine journal, 20(1): 54-60.**
17. **Noordzij M, Tripepi G, Dekker FW, Zoccali C, Tanck M W, & Jager K J (2010).** Sample size calculations: basic principles and common pitfalls. **Nephrology dialysis transplantation, 25(5): 1388-1393.**
18. **Noyes J, Booth A, Moore G, Flemming K, Tunçalp Ö and Shakibazadeh E (2019)**. Synthesising quantitative and qualitative evidence to inform guidelines on complex interventions: clarifying the purposes, designs and outlining some methods. **BMJ global health, 4**
19. **Pourhoseingholi MA, Vahedi M, & Rahimzadeh M (2013).** Sample size calculation in medical studies. **Gastroenterology and hepatology from bed to bench, 6(1): 14–17.**
20. **Relethford J H (2012).** Human population genetics (Vol. 7). **John Wiley & Sons.**
21. **Ruano-Ravina A, Pérez-Ríos M and Barros-Dios JM (2008).** Population-based versus hospital-based controls: are they comparable?. **Gaceta sanitaria, 22(6): 609-613.**
22. **Schlesselman JJ (1982).** Case-control studies: design, conduct, analysis. **Oxford University Press.**
23. **Song JW and Chung KC (2010).** Observational studies: cohort and case-control studies. **Plastic and reconstructive surgery, 126(6): 2234.**
24. **Suresh K, Thomas SV and Suresh G (2011).** Design, data analysis and sampling techniques for clinical research. **Annals of Indian Academy of Neurology, 14(4): 287**.

25. **Taber JM, Leyva B and Persoskie A (2015).** Why do people avoid medical care? A qualitative study using national data. Journal of general internal medicine**, 30(3): 290-297.**
26. **Zaccai JH (2004).** How to assess epidemiological studies. **Postgraduate medical journal, 80(941):140-147**.
27. **Li L, Zhang M, & Holman D (2011).** Population versus hospital controls for case-control studies on cancers in Chinese hospitals. **BMC medical research methodology 11, 167.**

Websites

1. https://www.calculator.net/sample-size-calculator.html?type=1&cl=95&ci=5&pp=50&ps=&x=95&y=24
2. https://www.surveymonkey.com/mp/sample-size-calculator/
3. https://www.surveysystem.com/sscalc.htm
4. http://www.raosoft.com/samplesize.html
5. https://www.qualtrics.com/blog/calculating-sample-size/
6. http://sampsize.sourceforge.net/iface/
7. http://osse.bii.a-star.edu.sg/calculation1.php

Figures

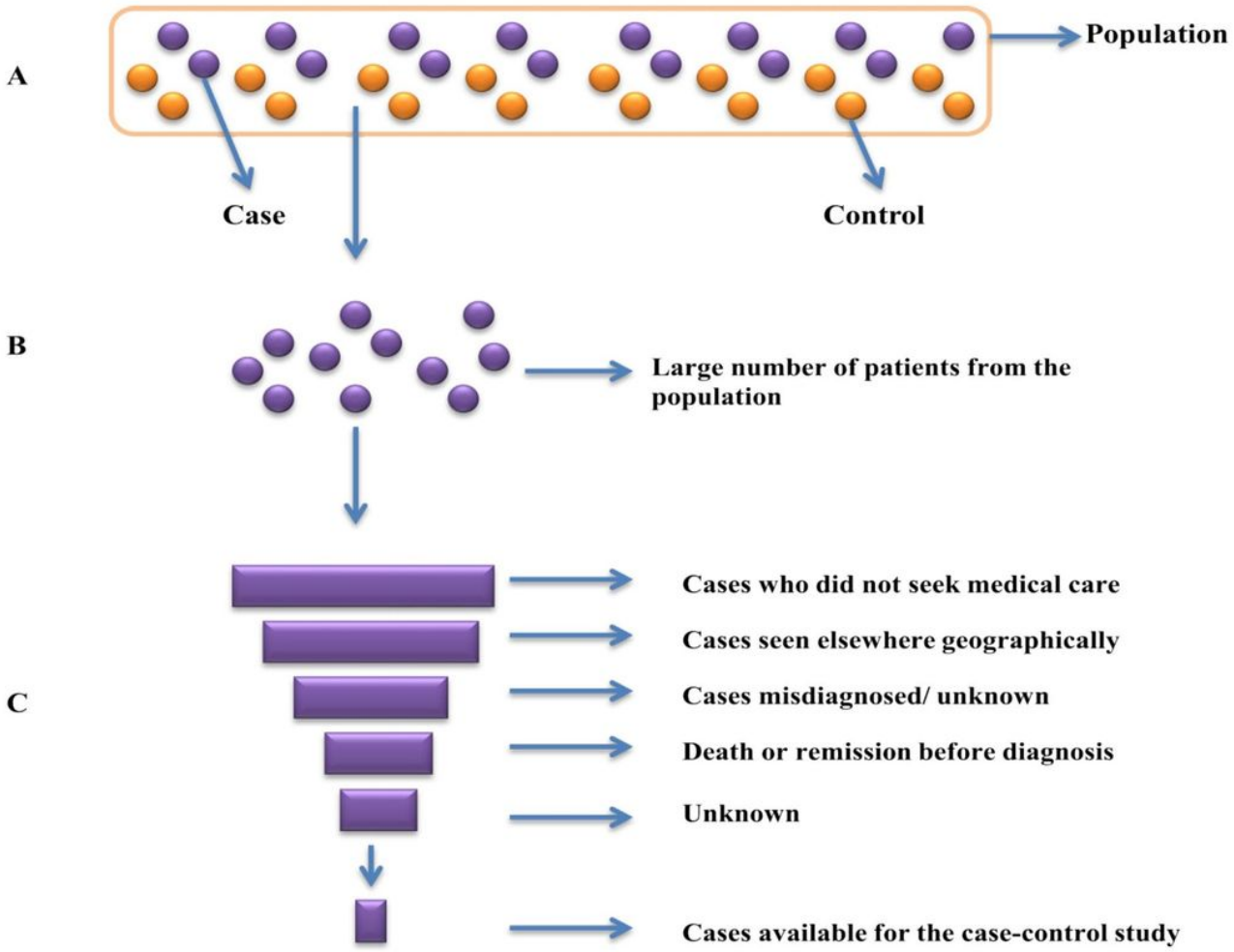

Figure 1

(A) Picture depicts a population having two kind of individual normal and diseased. (B) The number of patients is sorted out from the given population and then distributed into five different categories (C) Cases are distributed to five different categories including those case who did not seek medical care, some case are from different geographical, some cases may be misdiagnosed, some may died and some may locate under unknown. Only small numbers of cases are available for research study.

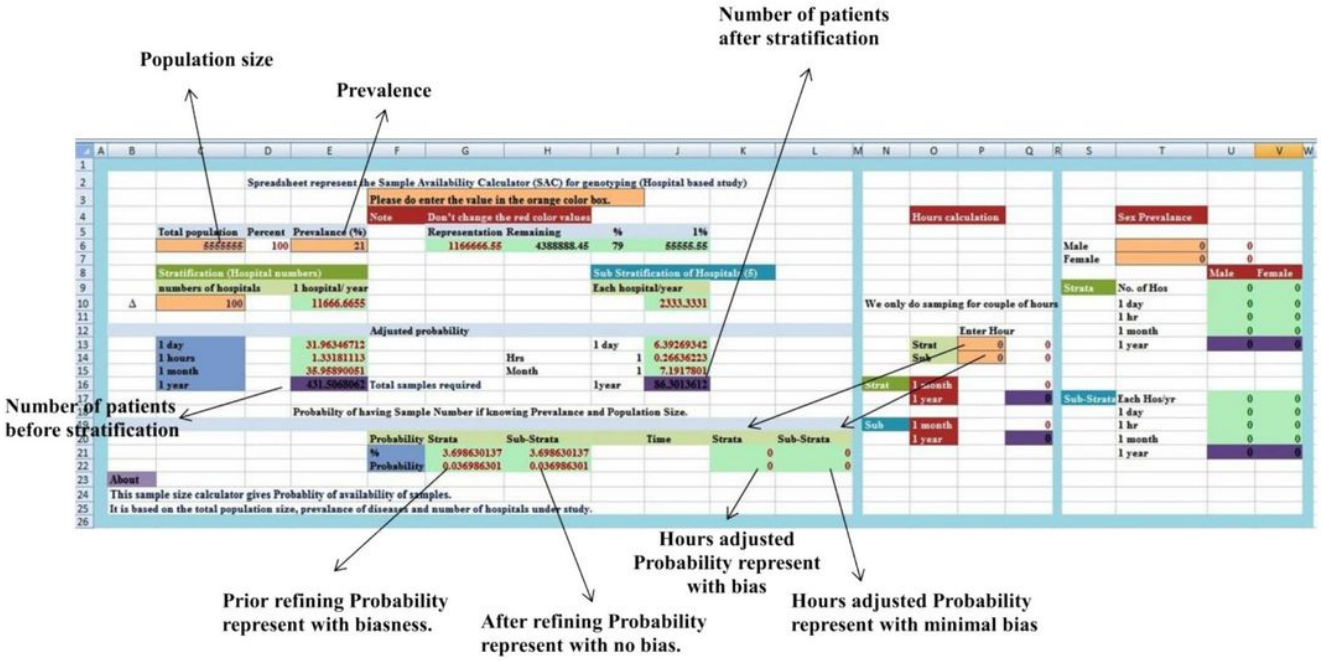

Figure 2

Main menu Sample availability calculator tool in MS-Excel 2002-2007

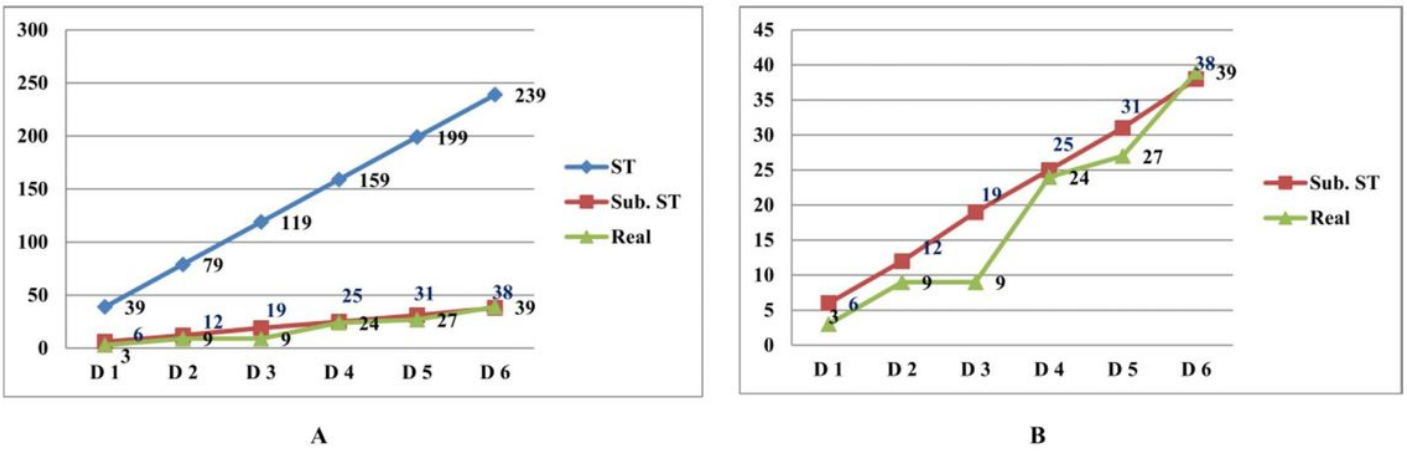

Figure 3

(A). In the real situation, we will never get so many samples and deviate from the calculated samples as shown in the graphs by stratified sampling method so representing biasness (Cluster-based sampling method). (B) if we will do equal distribution of cases in five different categories then sampling ends up with minimal bias so it's important to do refining with equal distribution.

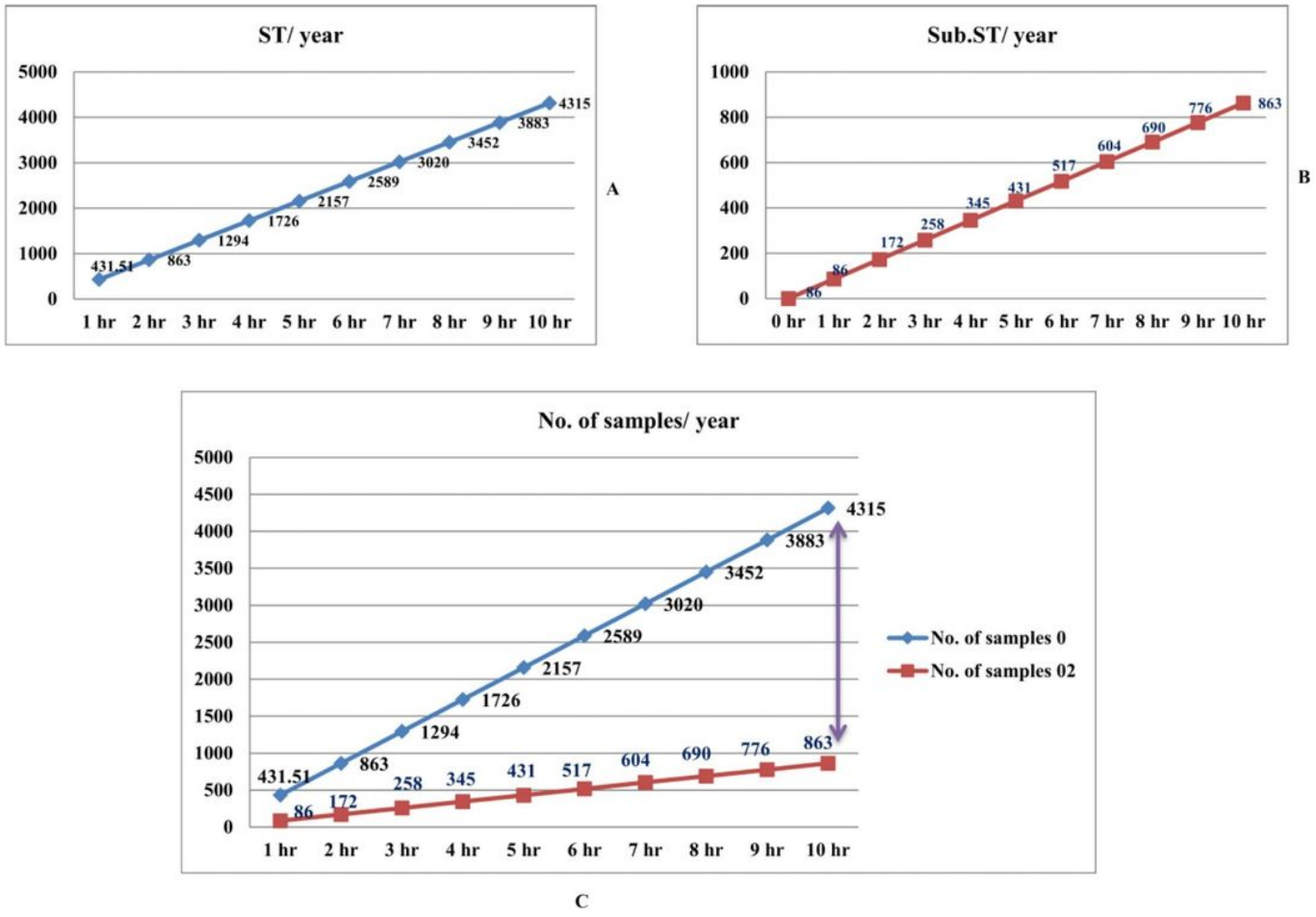

Figure 4

(A). The graph depicts, as we increase the period of sample collection, the chances of getting individuals have also increased therefore the probability of getting a sample to become increased. (B) as much as defined So if we increasing the time the chance of getting individuals also increased (C). Difference between these two representing the refining of samples and removing biases.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- sampleavailabilitycalculatorSAVlockedfinal.xlsx