Published online in http://ijam.co.in ISSN No: 0976-5921

International Journal of Ayurvedic Medicine, Vol 12 (2), 267-269

Sample Size determination for Censored Data

Research Article

Naseem Asghar[1*], Umair Khalil[2], Dost Muhammad Khan[2], Zardad Khan[3], Iftikhar Ud Din[4]

1. Lecturer, Department of Mathematics & Statistics, University of Swat, Pakistan.
2. Associate Professor, Department of Statistics, Abdul Wali Khan University Mardan, Pakistan.
3. Assistant Professor, Department of Statistics, Abdul Wali Khan University Mardan, Pakistan.
4. Assistant Professor, Public Health, Bacha Khan Medical College, Mardan, Pakistan.

Abstract

This study aims to describe sample size determination procedure in survival analysis using a real-world example. In this method simulation is used for sample size and precision calculations with censored data that concentrates on various sample sizes involved in carrying out the estimates and precision calculation. The Kaplan-Meier (K-M) estimator is chosen as a point estimator, and the precision measurement focuses on the mean square error, standard error, and confidence limits. Information obtained on the recovery time, in days, of patients from the population are compared with results taken from the sample group. Results showed a cutoff point of sample of size 675 on the basis of mean square error, standard error and confidence limit.

Key Words: Censored data, Kaplan-Meier estimator, Sample size, Sampling with replacement, Simulation.

Introduction

Researchers involved in clinical research, have the desire to publish their results to generalize findings to the population. This starts with the first step of deciding the topic to be studied, the subjects, and the type of study design. In this context, the researcher must determine how many subjects would be required for the proposed study (1). The size of the sample is not only an essential element in every statistical procedure but is also a great economic concern. In a sample survey, a statistician must determine the sample size. Statistical studies are always better when they are carefully planned. Good planning has many aspects. The study must be of adequate size, relative to the goals of the study. The sample size is important for economic reasons: An under-sized study can be a waste of resources for not having the capability to produce useful results, while an over-sized one uses more resources than are necessary (2).

Sample size estimation is very important in all types of research studies. The appropriate sample size formula depends on the study design, objective of the study, variable type, number of groups in the study, statistical analysis planned, and sampling technique to be used. The number of individuals to be included in a study depends on the precision of the estimated value, power of the study, level of significance or confidence level, and other constraints such as manpower, availability of subjects, money, feasibility in general, and time in particular (3).

Statisticians figure out that, the necessary sample size is dependent on the purpose of the study, the degree of certainty, and the degree of accuracy. The analytic formula of sample size is one strategy to decide the sample size and the alternate method to assess the sample size is, to use simulation technique (4). The simulation procedure accommodates many complex statistical designs. The Monte Carlo simulation can manage uncertainty and it endeavours to imitate the system tests gathered from the population, which underpins its utilization in sample size and parameter estimation (5).

Statistical studies need to be carefully planned. As it is hard to consider the whole population, decisions regarding population using sample data are without a problem. The problem should be carefully defined and operationalized. Sample units should be chosen randomly from the population of interest. The study size must be adequate relative to the objectives of the study. That is, it should be “adequately large” to detect statistical significance (6).

Not all sample size issues are the same, nor is sample size important in all studies. Sample size issues are generally more significant when it requires some investment to gather the information. An under-size study opens the subjects to possibly hurtful treatment without having the ability to create helpful outcomes, while an over-size study opens subjects to conceivably unsafe medicines that utilize a bigger number of assets than is needed (7).

There are a few ways to deal with sample size. There is sample size to accomplish a predefined

*** Corresponding Author:**
Naseem Asghar
Lecturer,
Department of Mathematics & Statistics,
University of Swat,
Pakistan.
Email Id: nsmsghr@gmail.com

standard error and sample size to accomplish a predetermined probability of obtaining statistical significance that is one can indicate the necessary width of a confidence interval and decide the sample size that accomplishes that objective (8).

The main objective of this study is to determine the optimum sample size for survival analysis, to get maximum efficiency through simulation by example. It will also be of interest that of which sample size we can arrive at a specific desired accuracy, that we can get an improvement of the predetermined estimator.

Materials and methods

In this study the survival/recovery estimates were obtained using the Kaplan-Meier product-limit estimator (9). We tried to attain estimation accuracy and budget efficiency at a certain level of precision by finding out a possible threshold sample size value (10). To deal with uncertainty, Monte Carlo simulation was used in sample size and parameter estimation (5), (11). The measurements used for efficiency evaluations are the mean square error, standard errors, and confidence limits.

Data was obtained from the office of Additional Director General Health, Khyber Pakhtunkhwa (KP) via ref: 252-55/ADGHS/PH/2019-20. The study protocol received ethical approval from the Ethics Committee of Khyber Medical University Peshawar, KP ref: KMU/ IPH/20200705. The data of all 7296 COVID-19 positive patients admitted in different hospitals of KP during the first wave was used in this study. The simulation and analysis were implemented in R.

Results

There was a total of 7296 patients admitted during the first wave of the COVID-19 pandemic at different hospitals of Khyber Pakhtunkhwa (KP), Pakistan. **Table 1** describes the average (median) of a patient's stay at the hospital. 6501(89.10%) patients recovered while 795 (10.89%) died. Died patients were considered as censored. The censoring rate was 11 percent.

Table 1: Population median of K-M recovery time, confidence limits, standard error, bias square, and mean square error with censoring percentage

n	event	median	0.95LCL	0.95UCL	St.err	bias^2	MSE	Censoring (%)
7296	6501	23	23	23	0.0249	0.0898	0.0904	11

The recovery estimates of K-M and their standard errors, mean square errors, confidence limits, and bias squares are calculated from varying sample sizes via simulation. In every simulation, the associated median time, confidence limits, standard errors, and corresponding estimates are stored. The bias square and mean square error were obtained from these estimates. The desired outcome (n= 25, 75, 125, 175, 225, 275, 325, 375, 425, 474, 525, 575, 625, 675, 725, 775, 825, 875, 925, 975) of population data from sampling with replacement, is shown in **Table 2**. A sample size of n=675 showed close results to the population results that is the median time is the same, confidence limits are the shortest, bias square, and mean square error is minimum.

Table 2: Samples medians of K-M recovery time, confidence limits, standard error, bias square, and mean square error with censoring percentage

n	event	Median	0.95LCL	0.95UCL	St.err	bias^2	MSE	Censoring (%)
25	23	31	22	46	0.2186	0.0005	0.0483	8
75	69	27	24	39	0.1371	0.0048	0.0236	8
125	117	22	20	32	0.1087	0.0065	0.0183	6
175	153	25	21	28	0.0996	0.0106	0.0205	13
225	198	22	20	25	0.1023	0.0302	0.0406	12
275	243	23	21	26	0.0905	0.0276	0.0358	12
325	280	23	22	26	0.0819	0.0233	0.0300	14
375	330	22	21	25	0.08	0.0316	0.0380	12
425	379	22	22	25	0.0744	0.0314	0.0370	11
475	419	23	22	25	0.0815	0.0550	0.0616	12
525	468	23	22	26	0.0689	0.0358	0.0405	11
575	516	22	22	25	0.0733	0.0548	0.0602	10
625	557	23	22	26	0.0625	0.0410	0.0449	11
675	600	23	22	24	0.0578	0.0280	0.0313	11
725	637	23	23	26	0.0632	0.0472	0.0512	12
775	699	25	23	26	0.059	0.0438	0.0472	10
825	750	25	23	26	0.0566	0.0425	0.0457	9
875	775	23	22	25	0.0577	0.0484	0.0518	11
925	816	23	22	25	0.0567	0.0501	0.0533	12
975	864	23	22	25	0.0553	0.0506	0.0536	11

Figure 1 compares estimates of patients of COVID-19 population and sample group of size n=675. The time from randomization to recovery is the response variable of interest.

Red line shows information obtained on the recovery time, in days, of patients who are in the population group. This information is compared with results taken from the sample group. The Kaplan–Meier estimate of the survivor function derived from both of the data sets is shown as the step function in figure 1.

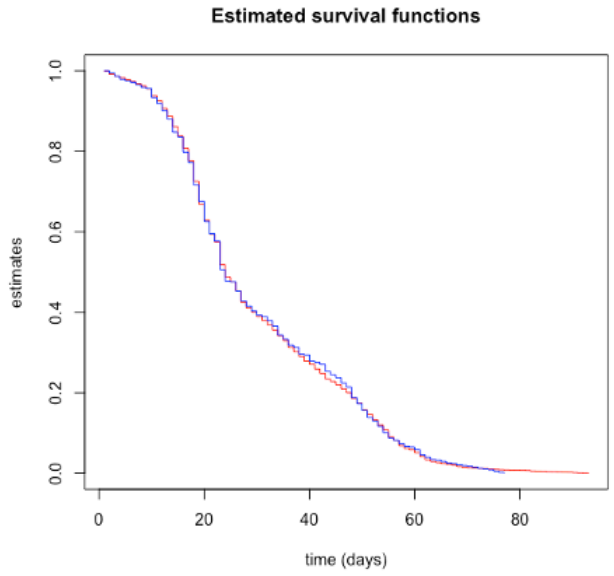

Figure 1: Estimated survivor functions for individuals from population(red) and from sample size 675 (blue)

Discussion

The recovery estimates of 7296 COVID-19 patients were investigated using the Kaplan-Meier product-limit estimator. The medians, standard errors, mean square errors, and confidence limits were calculated from population data as well as for samples of different sizes. Random index of size n was generated from sampling with replacement. Twenty samples of sizes (n= 25, 75, 125, 175, 225, 275, 325, 375, 425, 474, 525, 575, 625, 675, 725, 775, 825, 875, 925, 975) were obtained from population. Median estimates were stored. These estimates were recorded on average after 5000 repetitions. The bias square and mean square error were calculated for each median estimate. The results generated via simulation are different up to sample size n=575 and then started repeating till sample size 975.

The flexibility of simulation assists researchers to estimate the sample size required in the study design as it is common to examine the treatment effect in clinical trials (12). Also, with the simulation technique, one can determine the sample size needed for detecting interaction effect at some level of significance. Simulation necessitates statisticians to clearly express the analysis procedures, which encourages researchers to be more accurate and more careful about estimation and modelling (13).

The present study has several limitations. First, mean square error and shortest confidence limits were considered as measurements of precision. Second, due to the real data case, the results are applicable to this COVID-19 study only.

Conclusion

In this study, an attempt is made to find a cut off sample size point using a practical COVID-19 case in Pakistan. The Kaplan-Meier estimator was taken as the point estimator and the precision was made on basis of standard errors, shortest confidence intervals, and mean square errors. Twenty samples with replacement were generated randomly through simulation. Median estimates of different samples were compared with the population median estimate. A cut-off point of sample size of 675 was obtained on the basis of mean square error and shortest confidence limits.

References

1. Rao UK. Concepts in sample size determination. Indian Journal of Dental Research 2012;23(5):660.
2. Saeed N, Pervaiz MK, Shahbaz MQ. Determination of sample size. Editorial Advisory Board e 2005;4(3):319-25.
3. Binu VS, Mayya SS, Dhar M. Some basic aspects of statistical methods and sample size determination in health science research. Ayu 2014;35(2):119.
4. Che H. Cutoff sample size estimation for survival data: a simulation study 2014.
5. Preacher KJ, Selig JP. Advantages of Monte Carlo confidence intervals for indirect effects. Communication Methods and Measures 2012;6(2):77-98.
6. Lipsey MW, Aiken LS. Design sensitivity: Statistical power for experimental research. Sage 1990.
7. Shuster JJ. CRC handbook of sample size guidelines for clinical trials. CRC Press 1990.
8. Odeh RE, Fox M. Sample size choice: Charts for experiments with linear models. CRC Press 2020.
9. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. Journal of the American statistical association. 1958;53(282):457-81.
10. Dattalo P. Determining sample size: Balancing power, precision, and practicality. Oxford university press 2008.
11. Teare MD, Dimairo M, Shephard N, Hayman A, Whitehead A, Walters SJ. Sample size requirements to estimate key design parameters from external pilot randomised controlled trials: a simulation study. Trials 2014;15(1):1-3.
12. Arnold BF, Hogan DR, Colford JM, Hubbard AE. Simulation methods to estimate design power: an overview for applied research. BMC medical research methodology 2011;11(1):1-0.
13. Landau S, Stahl D. Sample size and power calculations for medical studies bysimulation when closed form expressions are not available. Statistical methods in medical research 2013;22(3):324-45.

*****