Determination of sample sizes when testing parameters of the populations taken from discrete distributions

Latif ÖZTÜRK
Faculty of Economics and Administrative Sciences
Kırıkkale University, Kırıkkale, Turkey
Habip KOÇAK
Faculty of Economics and Administrative Sciences
Marmara University, İstanbul, Turkey

Abstract: *This study aims to investigate which discrete distribution requires less sample size when estimating parameters. For this reason databases are formed from different discrete distributions with different parameter values using simulation. Samples are taken from the formed databases with Simple Random Sampling (SRS) without replacement and with Sequential Analysis (SA). The parameters are calculated from the databases constructed before and are compared with the values calculated from the sample taken. The minimum differences between parameter value and sample value are determined. In addition, these sample sizes between SRS and SA are compared and which sampling method that requires less sample size when estimating the parameters are determined. Finally, which sampling method that would be used is suggested, when the population distribution fit different discrete distributions with different parameter values.*

Key Words: Discrete distributions; Databases; Sample size; Simulation

1. Introduction

Statistical studies are always better when they are carefully planned. Samples must be selected from the appropriate population and reliable instruments should be used to obtain measurements. Finally the sample size must be of adequate size, relative to the aims of the study. Some process of obtaining samples is often expensive, involved, and time consuming. Sample size determination is an important step in planning statistical research. If the sample size is too large, we waste resources. If the sample size is too small, we cannot draw inferences with the desired precision. Thus, we need to calculate the sample size of a study before proceeding in order to determine the best trade-off between precision and resource use.

Often, a study has a limited budget, and that in turn determines the sample size. Another common situation is that a researcher may have established some convention regarding how much data is "enough". Not all the sample size problem is the same, nor is the sample size equally important in all studies.[1] The sample size is not only an essential element in every statistical procedure but it is also an item of great economic importance.[2] While the sample size of study is important, the way in which the sample has been collected is even more important. Quantitative research is often based on the assumption that the findings for a sample of people can be generalized to the larger population. We can not assume that the findings for the sample can be generalized to the larger population if the procedures to select the sample of research are not planned well.[3] Sampling should allow sufficiently reliable information about the particular population

[1] Russel, V. Lenth, "Some Practical Guidelines for Effective Sample Size Determination", The American Statistician , August 2001, Vol.55, No:3 pp. 187-193.

[2] Saeed, N., Pervaiz, M. K., Shahbaz, M. Q., "Determination of Sample Size", European Journal of Scientific Resaerch, 2006, Vol:14, No:3, pp:319-325.

[3] http://www.cyfc.umn.edu/policy/issues/briefings/savvyresearch.pdf "Savvy Use of Research: Tips for Policy Makers", Children, Youth&Family Consortium, 2004, University of Minnesota.

under investigation. When estimating population parameters from sample statistics, the sample size is important; larger sample sizes usually result in greater statistical reliability. [4] However, optimum sample size is a balance between statistical and practical considerations. In this study, the parameters of statistical distributions are estimated, taking samples from the databases generated from discrete distribution by means of simulation. Minimum error and minimum sample size criterion are considered when deciding the use of SRS or SA sampling methods.

2. Theoretical Background

Sequential Probability Ratio Test (SPRT) or SA is a specific sequential hypothesis test, first developed by Abraham Wald.[5] While originally developed for use in quality control studies in the realm of manufacturing, SA has been formulated for use in the computerized testing of human examinees as a termination criterion (Ferguson, 1969; Reckase, 1983; Eggen, 1999). This test is one statistical model available for making mastery decisions during computer-based criterion referenced tests. [6]

In SA, decisions about sample size and the type of data to be collected are made and modified as the study proceeds, incorporating information learned at earlier stages.[7] Thus a conclusion may sometimes be reached at a much earlier stage than would be possible with SRS, at consequently lower financial cost. SA may be used for either continuous or discrete distributions.[8] Samples are assumed to be identically independently distributed and drawn randomly. If SA is used in a cluster sampling it is applicable within the cluster, not to whole sample. Sample size of SA must be less than the predetermined fixed sample size of SRS. It is a necessity.

A vast literature in substantial detail of SA can be found in Lai (2001). The paper constitutes of a comprehensive review of recent developments in SA and some challenges and opportunities ahead. The review focuses on several classical problems and new horizons which highlight the interdisciplinary nature of subject.[9]

3. Generation of Random Variables From Discrete Distributions

In this part first a short description about the discrete distributions which are used to generate data by using random numbers is given. Then the algorithm of SRS and SA are given. To generate random numbers from a non-uniform distribution is usually done by applying a transformation to $U(0,1)$ random numbers. Assume that $F_X(x)$ is cumulative distribution function of a continuous random variable X, than it is not difficult to show that the random variable[10]

$$U = F_X(x) \quad (1)$$

has a distribution. If the cumulative distribution function is strictly increasing this equation can be rewritten in the equivalent form

$$X = F_{X-1}(U) \quad (2)$$

This process is called the inverse transform method. The inverse transform method is applicable to discrete distributions, but the inverse of the cumulative distribution function cannot be taken. However, the generalized inverse may be taken defined by

$$F_{X-1}(u) = \min\{x | u \le F_X(x)\} \quad (3)$$

Algorithm of Binomial Distribution

[4] http://www.utas.edu.au/sciencelinks/exdesign/S3.HTM 24.08.2008

[5] Wald, Abraham , *Sequential Tests of Statistical Hypotheses*, The Annals of Mathematical Statistics 16 (2): 117186., 1945.

[6] Welch, R. Edwin, Frick, Theodore W., *Computerized Adaptive Testing in Instructional Settings*, Educational Technology Research and Development, 41(3), 47-62, 1993.

[7] http://www.statistics.com/resources/glossary/s/seqan.php 31.08.2008

[8] Maxfield, M.W. & Barton-Dobenin, J.,.*A Sequential Sampling Plan For Determining Market Boundaries, Journal of Small Business Management*, VII(3), 25-59., 1980.

[9] Lai, T. L. , *"Sequential analysis: Some classical problems and new challenges"*, Statistica Sinica, 11, 303-351., 2001.

[10] http://ima.epfl.ch/cmos/Pmmi/interactive/rng7.htm 01.09.2008

[11] Kadry, S., Kouta, R., Smaili, K., "The Inverse of Transformation Method in Stochastic Mechanical Structure", Journal of Applied Sciences Research, 3(12): 1819-1824, 2007.

Bernoulli random variable with parameter p, $B(1, p)$ can be generated from a $U(0,1)$ random variable by the following algorithm:[12] If $U < p$, then deliver 1, otherwise deliver 0. For generating a binomial random variable with parameters N and p the fact is used that it is the sum of N independent $B(n, p)$ Random variables:[13]
The QBX program code for binomial distribution is;

```
for i=1 to n,
u=rnd
 if u<p then b(i)=1 else if u>=p then b(i)=0
x=x+b(i)
next i
```

Algorithm of Geometric Distribution

A geometric random variable gives the time until the first success in a sequence of Bernoulli trials.[14]
The QBX program code for geometric distribution is;

```
x=int(log(u)/log(1-p))+1
```

Algorithm of Poisson Distribution

The method we use is based on the following fact. If events occur randomly in time at a certain rate *r* and *X* is the number of events that occurs in a time period *t*, then *X* is a Poisson random variable with parameter $\lambda = rt$. Therefore we set[15]

$$X = \min\{n : U_1 U_2 \ldots U_n < \exp(-\lambda)\} - 1$$

or equivalently

$$X = \max\{n : U_1 U_2 \ldots U_n \geq \exp(-\lambda)\}.$$

The QBX program code for poisson distribution is;

```
u=rnd
p=exp(-lamda)
s=p
if u<s then stop
while u>s
x=x+1
p=(lamda*p)/x
s=s+p
wend
```

After generating the data bases from the random variable, samples are taken with SRS and the results are saved. Then samples are taken with SA and the results are compared with the result of SRS by means of sample sizes.

4. Empirical Results

The goal is to test population parameters by using the statistics obtained from samples. For mean and variance of the population the hypothesis are respectively, $H_0 : \mu = \mu_0$ $H_0 : \sigma^2 = \sigma_0^2$ versus $H_1 : \mu > \mu_0$, $H_1 : \sigma^2 > \sigma_0^2$, $H_1 : \mu < \mu_0$, $H_1 : \sigma^2 < \sigma_0^2$ and $H_1 : \mu \neq \mu_0$, $H_1 : \sigma^2 \neq \sigma_0^2$. First the fixed sample size for SRS must be determined. Then samples are taken from the databases constructed before and the parameters are estimated and hypotheses are tested. Finally the same process applied by using SA instead of SRS. All these are made by using a simulation program written before. The sample sizes are compared when achieving same result in SRS and SA.[16] The parameters given initially, the estimated parameters, sample sizes, hypothesis and results are shown in tables.

The simulations are performed for three discrete distributions described above. When performing simulations, 10.000 random numbers are generated with different parameter values which are distributed respectively Binomial, Geometric and Poisson. Each simulation is run for 100 times. For SRS fixed sample size is used and calculated as;[17]

$$n = \frac{N}{1 + N(e)^2} = \frac{10000}{1 + 1000(0{,}05)^2} \cong 385 \quad (4)$$

---

[12] http://ima.epfl.ch/cmos/Pmmi/interactive/rng8.htm 01.09.2008

[13] Moloy De, Samindranath Sengupta, "Bernoulli Sequential Estimation of the Size of a Finite Population", Sequential Anal., 22, No.1-2, 95-106, 2003

[14] Knopfmacher, A., Prodinger, H., "The first descent in samples of geometric random variables and permutations" Discrete Mathematics and Theoretical Computer Science, Vol:8, pp.215-234, 2006.

[15] Atkinson, A.C., "The Computer Generation of Poisson Random Variables", Applied Statistics, 28, No.1.pp.29-35, 1979.

[16] Kepner James L., Chang Myron N., "Samples of exact k-stage group sequential designs for Phase II and Pilot studies", Controlled Clinical Trials, 25, pp.326-333, 2004.

[17] Israel, Glenn D. "sampling The Evidence Of Extension Program Impact", Program evaluation and Organizational Development, IFAS, University of Florida. PEOD-5. October, 1992.

where e is the error rate and N is the finite population size.

For testing parameters only the hypothesis $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$ and $H_0 : \sigma^2 = \sigma_0^2$, $H_1 : \sigma^2 \neq \sigma_0^2$ are used. That is why the sample means and sample variances are calculated.

Table 1.The Simulation Result of Binomial Distribution5 $BINOM(N, p)$

Parameter Values Initially	Expected Parameters		Observed Parameters Sampling From Database				Calculated t Values		Smp. Size for SA	Result
	μ (np)	σ^2 (npq)	$\bar{X}_{SRS}$	$\bar{X}_{SA}$	σ^2_{SRS}	σ^2_{SA}	SRS	SA		
$B(5, 0.5)$	2.5	1,25	2.53	2.54	1.23	1.15	0.45	0.58	19	Accept
$B(5, 0.2)$	1.0	0,80	1.01	1.03	0.81	0.79	0.26	0.71	20	Accept
$B(5, 0.8)$	4.0	0,80	4.01	4.04	0.80	0.72	0.33	0.91	20	Accept
$B(10, 0.5)$	5	2,5	5.01	5.02	2.43	2.37	0.10	0.22	20	Accept
$B(10, 0.2)$	2	1,6	1.99	2.04	1.63	1.47	-0.07	0.57	19	Accept
$B(10, 0.8)$	8	1,6	7.99	8.06	1.63	1.47	-0.10	0.94	20	Accept

Table 2.The Result of Geometric Distribution $GEOM(p)$

P.V. Initially	Expected Parameters		Observed Parameters Sampling From Database				Calculated t Values		Smp. Size for SA	Result
	μ (1/p)	σ^2 (q/p^2)	$\bar{X}_{SRS}$	$\bar{X}_{SA}$	σ^2_{SRS}	σ^2_{SA}	SRS	SA		
$G(0.1)$	10	90	10.12	10.04	88.47	80.00	0.25	0.08	19	Accept
$G(0.2)$	5	20	5.87	5.66	20.47	19.41	0.38	0.30	20	Accept
$G(0.4)$	2.5	3.75	2.49	2.51	3.65	3.54	-0.06	0.13	20	Accept
$G(0.5)$	2	2	2.00	2.02	1.98	1.77	-0.01	0.23	19	Accept
$G(0.8)$	1.25	0.3125	1.26	1.28	0.32	0.33	0.24	0.96	20	Accept
$G(0.9)$	1/9	10/81	1.11	1.09	0.11	0.09	-0.39	-1.04	19	Accept

Table 3.The Result of Poisson Distribution $POISSON(\lambda)$

Parameter Values Initially	Expected Parameters		Observed Parameters Sampling From Database				Calculated t Values		Smp. Size for SA	Result
	μ (λ)	σ^2 (λ)	$\bar{X}_{SRS}$	$\bar{X}_{SA}$	σ^2_{SRS}	σ^2_{SA}	SRS	SA		
Poisson(1)	1	1	1.02	1.03	1.00	0.99	0.30	0.56	20	Accept
Poisson(2)	2	2	2.02	2.00	2.06	1.86	0.30	0.02	20	Accept
Poisson(3)	3	3	3.03	3.01	2.83	2.61	0.39	0.06	19	Accept
Poisson(4)	4	4	4.01	3.99	3.92	3.88	0.08	-0.06	20	Accept
Poisson(5)	5	5	5.03	5.04	4.93	4.57	0.28	0.37	19	Accept
Poisson(10)	10	10	10.05	10.07	9.88	9.67	0.34	0.43	20	Accept

5. Conclusion

As it is seen on the tables above the test result of SRS and SA are the same. However the sample sizes are different. In SRS fixed sample size is used and standard errors are calculated, in SA samples are taken one by one and about 20 samples the same standard error is reached. In addition this result, the sampling means and sampling variances are almost the same when we use fixed sample size about 380 for SRS and sequential sample about 20 for SA. As a result, we could conclude that, SA should be used instead of SRS when the researcher has knowledge about the distribution of populations. In sample size points of view the three discrete distributions mentioned above require the same sample size when SA is used.

References

1. Russel, V. Lenth, "Some Practical Guidelines for Effective Sample Size Determination", The American Statistician , August 2001, Vol.55, No:3 pp. 187-193.
2. Saeed, N., Pervaiz, M. K., Shahbaz, M. Q., "Determination of Sample Size", European Journal of Scientific Resaerch, 2006, Vol:14, No:3, pp:319-325.
3. http://www.cyfc.umn.edu/policy/issues/briefings/savvyresearch.pdf "Savvy Use of Research: Tips for Policy Makers", Children, Youth&Family Consortium, 2004, University of Minnesota.
4. http://www.utas.edu.au/sciencelinks/exdesign/S3.HTM 24.08.2008
5. Wald, Abraham , Sequential Tests of Statistical Hypotheses, The Annals of Mathematical Statistics 16 (2): 117186., 1945.
6. Welch, R. Edwin, Frick, Theodore W., Computerized Adaptive Testing in Instructional Settings, Educational Technology Research and Development, 41(3), 47-62, 1993.
7. Http://www.statistics.com/resources/glossary/s/seqan.php 31.08.2008
8. Maxfield, M.W. & Barton-Dobenin, J.,.A Sequential Sampling Plan For Determining Market Boundaries, Journal of Small Business Management, VII(3), 25-59., 1980.
9. Lai, T. L. , "Sequential analysis: Some classical problems and new challenges", Statistica Sinica, 11, 303-351., 2001.
10. http://ima.epfl.ch/cmos/Pmmi/interactive/rng7.htm 01.09.2008
11. Kadry, S., Kouta, R., Smaili, K., "The Inverse of Transformation Method in Stochastic Mechanical Structure", Journal of Applied Sciences Research, 3(12): 1819-1824, 2007.
12. http://ima.epfl.ch/cmos/Pmmi/interactive/rng8.htm 01.09.2008
13. Moloy De, Samindranath Sengupta, "Bernoulli Sequential Estimation of the Size of a Finite Population", Sequential Anal., 22, No.1-2, 95-106, 2003
14. Knopfmacher, A., Prodinger, H., "The first descent in samples of geometric random variables and permutations" Discrete Mathematics and Theoretical Computer Science, Vol:8, pp.215-234, 2006.
15. Atkinson, A.C., "The Computer Generation of Poisson Random Variables", Applied Statistics, 28, No.1.pp.29-35, 1979.
16. Kepner James L., Chang Myron N., "Samples of exact k-stage group sequential designs for Phase II and Pilot studies", Controlled Clinical Trials, 25, pp.326-333, 2004.
17. Israel, Glenn D. "Sampling The Evidence Of Extension Program Impact", Program evaluation and Organizational Development, IFAS, University of Florida. PEOD-5. October, 1992.