Journal of Data Analysis and Information Processing, 2015, 3, 163-167
Published Online November 2015 in SciRes. http://www.scirp.org/journal/jdaip
http://dx.doi.org/10.4236/jdaip.2015.34016

A Proposed Method for Choice of Sample Size without Pre-Defining Error

Loc Nguyen[1], Hang Ho[2]

[1]Huong Duong Company, Ho Chi Minh, Vietnam
[2]Vinh Long General Hospital, Vinh Long, Vietnam
Email: ng_phloc@yahoo.com, bshangvl2000@yahoo.com

Received 28 October 2015; accepted 20 November 2015; published 23 November 2015

Abstract

Sample size is very important in statistical research because it is not too small or too large. Given significant level α, the sample size is calculated based on the z-value and pre-defined error. Such error is defined based on the previous experiment or other study or it can be determined subjectively by specialist, which may cause incorrect estimation. Therefore, this research proposes an objective method to estimate the sample size without pre-defining the error. Given an available sample $\mathcal{X} = \{X_1, X_2, \cdots, X_n\}$, the error is calculated via the iterative process in which sample X is re-sampled many times. Moreover, after the sample size is estimated completely, it can be used to collect a new sample in order to estimate new sample size and so on.

Keywords

Sample Size, Choice of Sample Size, Pre-Defined Error

1. Introduction

Given a sample of size n, $\mathcal{X} = \{X_1, X_2, \cdots, X_n\}$ from a normal distribution with theoretical unknown mean μ and known variance σ^2, it implies that the sample mean

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is also normally distributed with mean μ and known variance σ^2/n. Given a confident level $100(1-\alpha)$ percentage, the confident interval [1] [2] of theoretical unknown mean μ is:

How to cite this paper: Nguyen, L. and Ho, H. (2015) A Proposed Method for Choice of Sample Size without Pre-Defining Error. *Journal of Data Analysis and Information Processing*, **3**, 163-167. http://dx.doi.org/10.4236/jdaip.2015.34016

$$\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

where $Z_{\alpha/2}$ which is the z-value at significant level α is the upper $100\alpha/2$ percentage point of standard normal distribution. Let E is the absolute deviation between the sample mean $\bar{X}$ and the theoretical mean μ, we have:

$$E = \left|\bar{X} - \mu\right|$$

The value E is also called estimated error, which is always less than or equal to $\frac{Z_{\alpha/2}}{\sqrt{n}}$.

$$E = \left|\bar{X} - \mu\right| \le Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

There is a requirement that how to estimate the sample size n so as to the deviation $\left|\bar{X} - \mu\right|$ is less than or equal to the pre-defined error E at given a $100(1-\alpha)$ % confident level. This is the choice of sample size. Following formula [1] indicates that $\left|\bar{X} - \mu\right|$ does not exceed the error E if the sample size n is:

$$n = \left(Z_{\alpha/2}\frac{\sigma}{E}\right)^2$$

(Readers can refer to [1] with regard to pages 252 - 253, 293 - 297, 304 - 305, 309 - 310, 312 - 314, 331 - 333, 344 - 345, 359, 364 - 365 for more details about choice of sample size). There is an issued problem that how to define the error E. Normally, E is defined based on the previous experiment or other study or it can be determined subjectively by specialist. Therefore, this research proposes an objective method to calculate the error E given an available sample $\mathcal{X} = \{X_1, X_2, \cdots, X_n\}$. This is an iterative method which is described in next section.

2. Proposed Method to Choose Sample Size

The formula of choice of sample size is re-written:

$$n = \left(Z_{\alpha/2}\right)^2\frac{\sigma^2}{\left|\bar{X} - \mu\right|^2}$$

The z-value $Z_{\alpha/2}$ is totally determined and so what we need to do is to calculate the variance σ^2 and the error $E = \left|\bar{X} - \mu\right|^2$. We will use a novel method when n is considered as a variable. The formula of sample size above is reduced as below:

$$n \propto h(i) = \frac{\sigma(i)^2}{\left|\bar{X} - \mu(i)\right|^2}$$

where $\bar{X}$ is sample mean and $h(i)$ is proportional to sample size n and the notation $\propto$ denotes the proportion.

Fixing variance $\sigma(i)^2$ and mean $\mu(i)$, we have:

$$h = \frac{\sigma^2}{\left|\bar{X} - \mu\right|^2} \Rightarrow \frac{1}{h} = \frac{\left|\bar{X} - \mu\right|^2}{\sigma^2}$$

Suppose there is an available $\mathcal{X} = \{X_1, X_2, \cdots, X_n\}$ and given m iteration times, for example, $m = 100$ and a new sample $\mathcal{Y}_i$ containing n elements is sampled randomly from $\mathcal{X}$ at the i^{th} iteration. This is a form of bootstrap sampling with replacement.

$$\mathcal{Y}_i = \left(Y_{i1}, Y_{i2}, \cdots, Y_{in}\right) \quad \text{where } Y_{ij} \in \mathcal{X}$$

Note that Y_{ij} (s) are taken randomly from with replacement.

Let $M(i)$ is the sample mean of $\mathcal{Y}_i$, we have:

$$M(i) = \frac{1}{n}\sum_{j=1}^{n} Y_{ij}$$

where,

$$Y_{ij} \in \mathcal{Y}_i$$

We assume that the sample mean $M(i)$ is approximated to the sample mean $\bar{X}$ and is random variable with theoretical mean μ. We have:

$$M(i) = \bar{X}$$

$$\mu = \frac{1}{m}\sum_{i=1}^{m} M(i)$$

$$\frac{1}{h} = \frac{\left| M(i) - \frac{1}{m}\sum_{i=1}^{m} M(i) \right|^2}{\sigma^2}$$

Note that $\bar{X}$ is approximated by $M(i)$.

Summing accumulatively $\frac{1}{h}$ through m iterations corresponding to m sample $\mathcal{Y}_i$ (s), we have:

$$\frac{m}{h} = \frac{\sum_{i=1}^{m}\left(M(i) - \mu\right)^2}{\sigma^2} = \frac{\sum_{i=1}^{m}\left(M(i) - \frac{1}{m}\sum_{j=1}^{m} M(i)\right)^2}{\sigma^2}$$

Dividing both sides of formula above by $\frac{1}{m-1}$, we have:

$$\frac{m}{(m-1)h} = \frac{\sum_{i=1}^{m}\left(M(i) - \frac{1}{m}\sum_{j=1}^{m} M(i)\right)^2}{(m-1)}\frac{1}{\sigma^2}$$

Let

$$\Delta^2 = \frac{\sum_{i=1}^{m}\left(M(i) - \frac{1}{m}\sum_{j=1}^{m} M(i)\right)^2}{(m-1)}$$

It is easy to infer that Δ^2 is sample variance of the set of sample means $M(i)$ (s).

$$\frac{m}{(m-1)h} = \frac{\Delta^2}{\sigma^2}$$

Therefore, the formula for calculating variable h with fixed variance σ^2 is:

$$h = \frac{m}{(m-1)}\frac{\sigma^2}{\Delta^2}$$

Because the theoretical variance σ^2 is not defined, it is approximated by sample variance s^2 of sample $\mathcal{X} = \{X_1, X_2, \cdots, X_n\}$.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

where $\bar{X}$ is sample mean.

Substituting s^2 into the formula for calculating variable h, we have:

$$h = \frac{m}{(m-1)}\frac{s^2}{\Delta^2}$$

Finally, the sample size n is calculated by following formula:

$$n = \left(Z_{\alpha/2}\right)^2 \frac{m}{(m-1)} \frac{s^2}{\Delta^2}$$

It is necessary to have an example for illustrating the proposed formula to calculate sample size without pre-defined error. Given 10-element sample $\mathcal{X}$ = {X_1 = 8.05, X_2 = 9.60, X_3 = 2.98, X_4 = –20.26, X_5 = –6.52, X_6 = –10.85, X_7 = 8.14, X_8 = 26.48, X_9 = 10.57, X_{10} = 2.26}, we will estimate the optimal size of the next sample based on $\mathcal{X}$. Suppose there are 10 iteration times (m = 10), we have 10 new sample $\mathcal{Y}_i$ (s) is sampled randomly from $\mathcal{X}$ with replacement. **Table 1** shows such 10 new sample $\mathcal{Y}_i$ (s) and their means $M(i)$ (s).

The sample variance Δ^2 of sample means $M(i)$ (s) is:

$$\Delta^2 = \frac{\sum_{i=1}^{10}\left(M(i) - \frac{1}{10}\sum_{j=1}^{10} M(i)\right)^2}{(10-1)} \approx 16.78$$

The mean $\bar{X}$ of sample $\mathcal{X}$ is:

$$\bar{X} = \frac{1}{10}\sum_{j=1}^{10} X_i \approx 3.04$$

The sample variance s^2 of sample $\mathcal{X}$ is:

$$s^2 = \frac{1}{10-1}\sum_{i=1}^{10}\left(X_i - \bar{X}\right)^2 = 169.79$$

Given the confident level 95% (α = 0.05), it is easy to calculate the optimal sample size as follows:

$$n = \left(Z_{\alpha/2}\right)^2 \frac{m}{(m-1)} \frac{s^2}{\Delta^2} \approx (-1.96)^2 \frac{10}{9} \frac{169.79}{16.78} \approx 43$$

According to results from many experiments, if the origin sample (previous sample $\mathcal{X}$) conforms normal distribution, the optimal sample size is 4 - 5 times larger than the size of such origin sample so that it is possible to gain better experiment (testing, analysis, estimation, etc.) because the origin sample that conforms normality is itself good sample. We can implement the proposed formula of sample size choice by *R* language [3] so that it is convenient to do many experiments on such formula. Following is *R* language code for implementing the proposed formula.

Table 1. Ten new samples and their means.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	$M(i)$
$\mathcal{Y}_1$	2.26	10.57	2.26	2.98	26.48	10.57	2.98	–20.26	26.48	–10.85	5.35
$\mathcal{Y}_2$	10.57	–20.26	26.48	8.05	2.98	26.48	–20.26	9.6	2.26	8.05	5.4
$\mathcal{Y}_3$	–10.85	–20.26	10.57	26.48	–10.85	2.26	–10.85	–20.26	–10.85	9.60	–3.5
$\mathcal{Y}_4$	–6.52	2.98	9.60	–6.52	2.98	26.48	9.60	–20.26	10.57	2.98	3.19
$\mathcal{Y}_5$	10.57	–20.26	2.26	2.98	–6.52	2.98	–6.52	–6.52	8.14	–20.26	–3.31
$\mathcal{Y}_6$	10.57	–6.52	26.48	2.98	2.26	8.05	9.6	8.14	–6.52	8.05	6.31
$\mathcal{Y}_7$	9.60	–6.52	–10.85	9.60	–20.26	–20.26	2.98	8.14	2.26	8.05	–1.73
$\mathcal{Y}_8$	9.60	26.48	2.98	–20.26	26.48	8.14	8.14	8.05	10.57	–6.52	7.37
$\mathcal{Y}_9$	8.05	–10.85	8.14	8.05	–20.26	10.57	9.60	10.57	2.98	26.48	5.33
$\mathcal{Y}_{10}$	8.14	10.57	2.98	26.48	–20.26	8.14	8.05	–20.26	2.26	–6.52	1.96

```
sample.size.choice<– function(X, m=10, conf.level=0.95)
{
    M <–vector(length = m);
    for(i in 1:m)
    {
        Y <– sample(X, length(X), replace=TRUE);
        M[i]<– mean(Y);
    }
    delta.square<–var(M);
    s.square<–var(X);
    z <–qnorm((1-conf.level)/2);
    z*z*(m/(m–1))*s.square/delta.square;
}
```

3. Conclusions

I invent this method when discussing with the co-author Dr. Hang Ho about choice of sample size. At that time, I make the simile that the ideology of this method is similar to the problem "hen and egg". Regardless that hen exists before or egg exists before, you feed hen to lay new egg and incubate such egg to hatch new hen. Therefore, given an available random sample is used to estimate the sample size and such sample size is applied to collect new random sample; after that new sample size is estimated based on the new random sample and so on. Now, we analyze the formula for estimating sample size:

$$n = \left(Z_{\alpha/2}\right)^2 \frac{m}{(m-1)} \frac{s^2}{\Delta^2}$$

The variance s^2 in numerator expresses the coherent variation of data and the value Δ^2 in denominator specifies the variation of disturbed data (data is disturbed for many times). It means that Δ^2 specifies the *variation of change* (or variation of variation). The smaller the value Δ^2 is, the more precise the variance s^2 is and so the sample size is much proportional to s^2. In other words, the small Δ^2 makes an increase in sample size. Ratio

$$\frac{m}{m-1}$$

approaches 1 when m approaches $+\infty$ and so, the larger the number of iterations is, the more precise the sample size is. If m is small, the sample size tendentiously increases, but the balance is established because Δ^2 will increase if m is small, and as known the large Δ^2 makes decrease in sample size. But why the small m makes an increase in Δ^2 and otherwise? As known the number of iterations m specifies the variation of disturbed data. The larger the number m is, the much more the data is disturbed and so it is easier for the tendency that data is reverted in equilibrium, which causes the decrease in Δ^2. In other words, the small m makes increase in Δ^2.

References

[1] Montgomery, D.C. and Runger, G.C. (2003) Applied Statistics and Probability for Engineers. 3rd Edition, John Wiley & Son, Inc., Hoboken.

[2] Walpole, R.E., Myers, R.H., Myers, S.L. and Ye, K. (2007) Probability & Statistics for Engineers & Scientists. 9th Edition, Pearson Education, Inc.

[3] R Development Core Team (2010) R: A Language and Environment for Statistical Computing. R Foundation for Statistical computing, Vienna, Austria. http://www.R-project.org