Iranian Journal of Science & Technology, Transaction A, Vol. 34, No. A4
Printed in the Islamic Republic of Iran, 2010

"Research Note"

TWO-PHASE SAMPLE SIZE ESTIMATION WITH PRE-ASSIGNED VARIANCE UNDER NORMALITY ASSUMPTION*

M. SALEHI[1, 2**], P. S. LEVY[3] AND J. RAO[4]

[1]Department of Mathematics, Statistics and Physics, Qatar University, Doha, Qatar
[2]Department of Mathematical Sciences, Isfahan University of Technology, Isfahan, I. R. of Iran
[3]RTI International, Research Triangle Park, North Carolina, USA
[4]School of Mathematics and Statistics, Carleton University, Ottawa, Canada
Email: salehi@qu.edu.qa, salehi_m@cc.iut.ac.ir

Abstract – We develop a two phase sampling procedure to determine the sample size necessary to estimate the population mean of a normally distributed random variable and show that the resulting estimator has pre-assigned variance and is unbiased under a regular condition. We present a necessary and sufficient condition under which the final sample mean is an unbiased estimator for the population mean.

Keywords – Population mean, sample size determination, two phase sampling

1. INTRODUCTION

We consider here the question of how large a sample size is needed to estimate the mean of a variable of interest, X, that is assumed to be normally distributed with mean μ and variance σ^2. One common formulation often used translates loosely into a statement that "we wish to take a sample size large enough to obtain $100\times(1-\alpha)\%$ confidence intervals that are within ε of the true mean, μ." This leads to the following basic formula:

$$n = \frac{z_{1-\alpha/2}^2 \sigma^2}{\varepsilon^2}, \tag{1}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution.

Another perhaps less common formulation is to choose a sample size large enough such that the sample mean has either a pre-assigned variance v_0 or a pre-assigned limit of error (e.g., $a\sigma$, where a is some constant). If $a = z_{1-\alpha/2}/\sqrt{n}$, then the formulation becomes equivalent to (1). We assume that either σ^2 is known or it is estimated, but it is very rare that we would actually know the variance, σ^2. We consider here the latter formulation that σ^2 is not known and must be estimated. Three possible ways of estimating σ^2 for sample size determination are: 1) pure guessing; 2) from previous studies on the same or a similar population or as a result of a pilot survey; and 3) by taking the sample in two phases. With respect to the two-phase option, at the first phase, one selects a random sample of size n_1 and obtains an estimator S_1^2 of σ^2, and then uses S_1^2 to obtain an estimator of the required sample size, $\hat{n}$. If $\hat{n} > n_1$, an independent random sample of size $\hat{n} - n_1$ is taken and combined with the first phase sample of fixed

*Received by the editor May 20, 2010 and in final revised form November 29, 2010
**Corresponding author

size n_1; for a similar discussion in a finite population context, [1]. We consider here the two phase method of estimating the required sample size $\hat{n}$.

The mean, $\bar{X}$, of the combined two-phase sample is not generally an unbiased estimator of the population mean, μ. Cox [2], however, showed that its bias is of order $O(n_1^{-2})$ under some regularity conditions. In the same article, he provided, using a Taylor expansion, an approximation to $\hat{n}$ for which the variance of $\bar{X}$ is equal to $v_0\{1+O(n_1^{-2})\}$ under some regularity conditions. In this note, we show that $\bar{X}$ is an unbiased estimator of μ if one assumes that the random variable X follows a normal distribution. We also provide a formula for the exact sample size, $\hat{n}$, for which the variance of $\bar{X}$ is exactly equal to the pre-assigned variance v_0 under the condition $\hat{n} > n_1$, which is equivalent to Cox's regularity condition (iv) (bottom of page 218 in the paper cited above).

2. ESTIMATION

Estimation of the mean μ consists of two steps. In the first step, we take a random sample of size n_1, say $\Omega_1 = \{X_1, X_2, \ldots, X_{n_1}\}$ and obtain the sample mean

$$\bar{X}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i$$

and the sample variance

$$S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i - \bar{X}_1)^2,$$

based on Ω_1. If n_1 is large enough, S_1^2 will be a reliable estimator of σ^2. In the second step, we use $\hat{n}$ given by (2) below:

$$\hat{n} = \begin{cases} n_1 & \text{if} \quad ks_1^2/v_0 \le n_1 \\ ks_1^2/v_0 & \text{if} \quad ks_1^2/v_0 \le n_1 \end{cases} \tag{2}$$

where

$$k = 1 + \frac{2}{n_1 - 3} = \frac{n_1 - 1}{n_1 - 3}$$

we can regard k as the effect of not knowing σ^2 and using a two-phase sample. In essence, if $kS_1^2/v_0 \le n_1$, the required sample size is equal to n_1. On the other hand, if $kS_1^2/v_0 > n_1$, we take an independent random sample of size $n_2 = \hat{n} - n_1$ and append it to Ω_1, ending up with the final random sample $X_1, X_2, \ldots, X_{\hat{n}}$. Otherwise, we consider Ω_1 as the final sample.

The proposed estimator of $\hat{\mu}$ is given by

$$\hat{\mu} = \begin{cases} \frac{1}{\hat{n}}\sum_{i=1}^{\hat{n}} X_i = \bar{X} & \text{if} \quad \hat{n} > n_1 \\ \frac{1}{n_1}\sum_{i=1}^{n_1} X_i = \bar{X}_1 & \text{if} \quad \hat{n} \le n_1 \end{cases} \tag{3}$$

we will now show, using lemmas 1 and 2 below, that the estimator $\hat{\mu}$ is unbiased for the mean μ and has variance equal to the pre-assigned value v_0.

Lemma 1. $\hat{\mu}$ is an unbiased estimator of μ.

Proof:

$$E[\hat{\mu}] = (1 - p)E_1E_2[\bar{X} \mid \Omega_1] + pE_1[\bar{X}_1], \tag{4}$$

where $p = \Pr(\hat{n} \le n_1)$ and E_1, E_2 denote the expectations over the first phase and the second phase, respectively. We have,

$$E_2[\bar{X} \mid \Omega_1] = E_2[\frac{n_1\bar{X}_1 + (\hat{n} - n_1)\bar{X}_2}{\hat{n}} \mid \Omega_1]$$
$$= \mu + \frac{n_1}{\hat{n}}(\bar{X}_1 - \mu)$$

where $\bar{X}_2$ is the sample mean of the second phase sample. Since $\bar{X}_1$ and S_1^2 are independent under the normal distribution assumption, we have

$$\begin{aligned} \mathrm{E_1E_2}[\overline{\mathrm{X}} \mid \Omega_1] &= \mathrm{E_1}[\mu + \frac{\mathrm{n_1}}{\hat{\mathrm{n}}}(\overline{\mathrm{X}}_1 - \mu)] \\ &= \mu + \mathrm{n_1E_1}[\frac{(\overline{\mathrm{X}}_1 - \mu)}{\mathrm{S_1^2}/\nu_0}] \\ &= \mu + \mathrm{n_1E_1}[\frac{1}{\mathrm{S_1^2}/\nu_0}]\mathrm{E_1}[\overline{\mathrm{X}}_1 - \mu] \end{aligned}$$

since $E_1[\bar{X}_1] = \mu$, we have $\mathrm{E_1E_2}[\overline{\mathrm{X}} \mid \Omega_1] = \mu$ and hence $E[\hat{\mu}] = \mu$ from (4).

Corollary. A necessary and sufficient condition under which $\bar{X}$ is an unbiased estimator of the mean μ is

$$E_1[\frac{(\bar{X}_1 - \mu)}{S_1^2}] = 0.$$

Lemma 2. Under normality and the regularity condition $\hat{n} > n_1$, the required sample size to achieve the pre-assigned variance υ_0 is given by

$$\hat{n} = (1 + \frac{2}{n_1 - 3})\frac{S_1^2}{\nu_0}$$

Proof: To evaluate the variance inflation caused by using S_1^2 instead of σ^2, we compute $Var[\bar{X}]$ where the sample size is equal to $\hat{n} = S_1^2/\upsilon_0$ (ignoring momentarily for reasons of simplicity the factor *k*, the effect of using the two-phase design that was shown in (2)). We have

$$\mathrm{var}[\bar{X}] = V_1E_2[\bar{X} \mid \Omega_1] + E_1V_2[\bar{X} \mid \Omega_1], \tag{5}$$

where V_1 and V_2 denote variances over the first and second phases, respectively. The first term of the right hand side of (5) is

$$V_1E_2[\bar{X} \mid \Omega_1] = V_1[\mu + \frac{n_1}{\hat{n}}(\bar{X}_1 - \mu)]$$
$$= n_1^2V_1[\frac{\bar{X}_1 - \mu}{\hat{n}}] = n_1E_1[\frac{1}{\hat{n}^2}]\sigma^2.$$

The second term on the right hand side of (5) is

$$E_1V_2[\bar{X} \mid \Omega_1] = E_1V_2[\frac{n_1\bar{X}_1 + (\hat{n} - n_1)\bar{X}_2}{\hat{n}} \mid \Omega_1]$$
$$= E_1[(\frac{\hat{n} - n_1}{\hat{n}})^2 \frac{\sigma^2}{\hat{n} - n_1}] = \left(E_1[\frac{1}{\hat{n}}] - E_1[\frac{n_1}{\hat{n}^2}] \right)\sigma^2 .$$

Adding the two terms in (5), we have

$$\text{var}[\bar{X}] = E_1[\frac{1}{\hat{n}}]\sigma^2 = \upsilon_0\sigma^2 E_1[\frac{1}{S_1^2}]$$

the random variable $(n_1 - 1)S_1^2/\sigma^2$ has a chi-squared distribution with $(n_1 - 1)$ degrees of freedom. Hence,

$$E_1[1/S_1^2] = (n_1 - 1)/(\sigma^2(n_1 - 3))$$

and

$$\text{var}[\bar{X}] = \upsilon_0\left(1 + \frac{2}{n_1 - 3}\right),$$

noting that

$$E[Y^r] = 2^r\Gamma(\frac{k}{2} + r)\Big/\Gamma(\frac{k}{2})$$

if $X \sim \chi^2$ with k degrees of freedom. Hence, by letting $\hat{n} = [1 + 2/(n_1 - 3)]S_1^2/v_0$ rather than $\hat{n} = S_1^2/v_0$ we achieve the exact pre-assigned variance, v_0.

3. ILLUSTRATIVE EXAMPLE

We generated 10,000 random normal variable with mean 29.03 and variance 8.86., representing the age when first employed of an artificial population of 10,000 employees at a factory. We wish to take a sample of those employees to estimate the mean with pre-assigned variance, v_0, equal to 0.04 years. From a first stage sample of $n_1 = 30$ observations, we obtained s_1^2=12.81 and $\hat{n} = (s_1^2/0.04) \times (30 - 1)/(30 - 3) =$ $343.97 \approx 344$. If we were to assume that the true population variance is 12.81 (the value of s_1^2 obtained from the first phase sample), the estimated sample size would be obtained from equation (1) with $z_{1-\alpha/2}$=1, σ^2=12.81, and $\varepsilon = \sqrt{0.04}$ =0.2, and the required sample size would be equal to 321. At a cost of 23 additional observations (7% more than what would have been obtained from using equation (1)), one has a sample size that is based on a procedure in which the value of σ^2 used to estimate the required sample size is based on actual data, and which ensures that the estimator $\hat{\mu}$ has the pre-assigned variance v_0.

4. CONCLUSIONS

We revisited a well-known two-phase sampling procedure in which it is desired to estimate a population mean with pre-assigned variance, and the required sample size is determined from data collected at the first phase. If a second phase is necessary, the data obtained from the two phases are combined to form the required estimator. Otherwise, the estimator is obtained from the phase 1 sample. Under the assumption of normality, we derived an estimator that is unbiased for the mean. Under a regularity condition that the

final sample size, $\hat{n}$, determined from this procedure, will always have greater than the phase 1 sample size, n_1, (equivalent to saying that the sampling procedure will always have two phases), then the variance of the resulting estimated mean will be equal to the pre-assigned variance, v_0.

This method could be useful in scenarios where one does not wish to assume a known variance for the distribution of the variable of interest and the first phase sample size is not large. Its main disadvantage is that the unknown final sample size adds difficulty to budgeting and planning the logistics of the study.

REFERENCES

1. Cochran, W. G. (1977). *Sampling Techniques.* 2nd ed. New York, Wiley & Sons.
2. Cox, D. R. (1952). Estimation by double sampling. *Biometrika, 39*, 217-27.