AN EMPIRICAL STUDY OF THE BOOTSTRAP AND THE JACKKNIFE METHODS APPLIED IN DOUBLE SAMPLING

Jing Wang and Ferry Butar Butar, Sam Houston State University
Ferry Butar Butar, Sam Houston State University, Huntsville, Texas 77341

Abstract: When the variable of interest is relatively expensive to measure and a correlated auxiliary variable can be measured easily then it is efficient to employ a double sampling. The primary purpose of this paper is to compare the efficiency between the Jackknife method and the bootstrap method when applied to double sampling. The variance of the estimated total or the estimated mean will be sought. The estimated variance will be compared by using the jackknife and the bootstrap methods, respectively. Using a simulation study, we evaluate the efficiency of double sampling by comparing the bootstrap and the jackknife to simple random sampling.

Key Words: Ratio Estimation, Regression Estimation, Difference Estimation, Double Sampling, Bootstrap, Jackknife

1. INTRODUCTION

Double sampling, proposed for the first time by Neyman (1938), now is widely applied in sample surveys for various reasons. Usually, this technique is used for situations when response data are too expensive to obtain but some correlated data can be easily measured. In the first phase, the correlated variables are measured for every unit and thus the phase I sample is generally relatively large. Then, samples are randomly selected from the phase I sample population by a probability scheme and taken as phase II sample. After the phase II sample is obtained, the variables of interest are measured for each unit in the sub-sample. Since the phase I sample is treated as the population from which the phase II sample is drawn, the auxiliary information gathered in the phase I might be used to design the phase II sample.

2. VARIANCES IN DOUBLE SAMPLING

Let y_i be the variable of interest for the *i*th unit, and the auxiliary variable be x_i. Let n' be the number of the units in the first sample S' and *n* be the number of the units in the second sample. For each unit in the second sample, both y_i and x_i are observed, while for the rest of the units in the first sample, only x_i is observed. The first sample S' is selected from the whole population of N units, where *x*-values are observed by random sampling without replacement. The estimated variance of a ratio estimation of $\bar{Y}_r$ is:

$$\hat{V}(\hat{\bar{Y}}_r) \approx \frac{(N-n')}{N\,n'} s_y^2 + \frac{(n'-n)}{n'n} s_r^2, \qquad (2.1)$$

where $s_y^2 = (n-1)^{-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$, and

$s_r^2 = (n-1)^{-1}\sum_{i=1}^{n}(y_i - rx_i)^2$, where *r* is the ratio.

The estimated variance of a regression estimation is:

$$\hat{V}(\hat{\bar{Y}}_{lr}) = \left[\frac{1}{n} - \rho^2\left(\frac{1}{n} - \frac{1}{n'}\right)\right] s_y^2, \qquad (2.2)$$

The estimated variance of a difference estimation is

$$\hat{V}(\hat{\bar{Y}}_d) = \left(\frac{1}{n'} - \frac{1}{N}\right) s_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right) s_d^2, \qquad (2.3)$$

where $s_d^2 = (n-1)^{-1}\sum_{i=1}^{n}[y_i - \bar{y} - k(x_i - \bar{x})]^2$, where k is a contant.

3. THE BOOTSTRAP METHOD

3.1 Percentile Bootstrap Confidence Intervals

Suppose that $Y = (Y_1, Y_2, \ldots, Y_n)$ is a random sample of size *n* and $\hat{\theta} = \hat{\theta}(y)$ is a point estimator of θ. A (1-α)100% confidence interval for θ is $(\hat{\theta}_L, \hat{\theta}_U)$, where $\hat{\theta}_L = \hat{\theta} - z_{1-\alpha/2}\delta_{\hat{\theta}}$ and $\hat{\theta}_U = \hat{\theta} - z_{\alpha/2}\delta_{\hat{\theta}}$, where $\delta_{\hat{\theta}}$ is standard error of $\hat{\theta}$. Then $P(\hat{\theta}^* \le \hat{\theta}_L) = \alpha/2$ and $P(\hat{\theta}^* \le \hat{\theta}_U) = 1 - \alpha/2$ (see Hogg et al. (2005)). Let $\hat{\theta}^*_{(1)} \le \hat{\theta}^*_{(2)} \le \ldots \le \hat{\theta}^*_{(B)}$ denote the ordered values $\hat{\theta}^*_{(1)}, \ldots, \hat{\theta}^*_{(B)}$. Let *m*= $[(\alpha/2)B]$, then $(\hat{\theta}^*_{(m)}, \hat{\theta}^*_{(B+1-m)})$ is (1-α)100% confidence

interval for θ, which is called the percentile bootstrap confidence interval for θ.
The standard error of the bootstrap estimate is

$$\delta_{\hat{\theta}}=\sqrt{\frac{1}{B-1}\sum_{i=1}^{B}(\theta_i^*-\theta_{avg}^*)^2}\ , \tag{3.1}$$

where θ_i^* is the bootstrap value of a quantity and θ_{avg}^* is the average of the bootstrap values of the quantity.

4. THE JACKKNIFE METHOD

Let $\hat{\theta}$ denote an estimator of the unknown parameter θ based on the sample of size $n=gh$ where g is the number of group and h is the size of the sample in each group and $\hat{\theta}_{-i}$ be the estimator of θ based on the sample by deleting the *i*th group, then the estimated variance is

$$\hat{V}_{jack}(\hat{\theta})=\sqrt{\frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{-i}-\hat{\theta}_{avg})^2}\ , \tag{4.1}$$

where $\hat{\theta}_{avg}$ is the average of all estimator $\hat{\theta}_{-i}$ s' values (see Quenouille (1956)) .

5. DATA ANALYSIS

The example in this paper is based on an experiment, through which some conclusions can be obtained. In this experiment, the dataset "audit" (Lohr, 1998) is adopted to compute the estimated variances of the population mean and the population total when double sampling is applied with three estimation schemes: ratio estimation, difference estimation, and regression estimation. For each estimation scheme, the bootstrap and the jackknife sampling techniques are applied, respectively, to compute the estimated values. We use only two variables from the dataset, one is bookval (book value account) treated as the auxiliary variable *x*, and cumbv (cumulative book value) taken as the response variable *y*. From the scatter plot (not shown here), it appears that there is a linear relationship between the variable *x* and the response *y*. Table 1 summarizes some necessary information for this experiment.

Note that three different sample sizes (16, 32, 64) for the second phase are tried in this experiment. The values of k and repeat are just assumed by the experimenter and can be changed though programming interface. The other parameters can be observed by running the main function.

Table 1. Parameter information of experiment

N	n'	n	$\bar{x}'$	k (difference)	Repeat (bootstrap)
87	79	16, 32, 48	7400.86	2	20

N: the size of the population; n': the size of the first sample; *n*: the size of the second sample; $\bar{x}'$: the average value of the auxiliary variable in the first sample; *k*: the coefficient in the difference estimation equation.; repeat: the number of bootstrap samples generated in this experiment.
The following table summarizes the means for different sample sizes.

Table 2. Summary of the estimated mean for three sample sizes

	Ratio	Difference	Regression
n=16			
Simple	61526	44685	42437
Boot	44808	35908	43097
Jack	59106	43192	42330
n=32			
Simple	72473	27964	27856
Boot	50589	22582	28408
Jack	71149	27533	27850
n=48			
Simple	20897	18570	18552
Boot	16980	15495	18777
Jack	20694	18396	18548

We produce 6 figures from the SAS output.

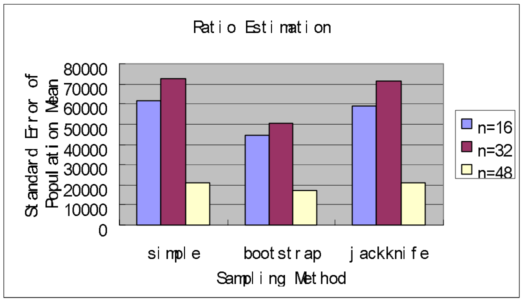

Figure 1: Ratio estimation

6. CONCLUSIONS AND DISCUSSIONS

Following four conclusions are obtained from the figures:
1. Bootstrap method makes the best performance when given the estimation scheme and the size of the second sample in double sampling. Jackknife method performs not much different from simple method.

2. The performances of estimation schemes are relative to the size of the second sample in double sampling. The larger the sample size is, the better the performance is.

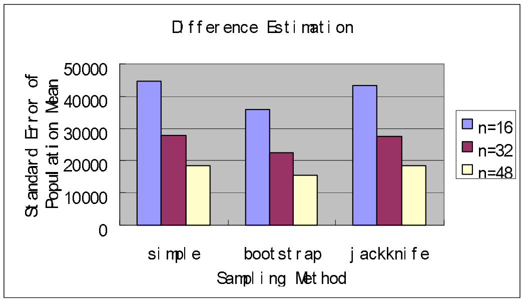

Figure 2: Difference estimation

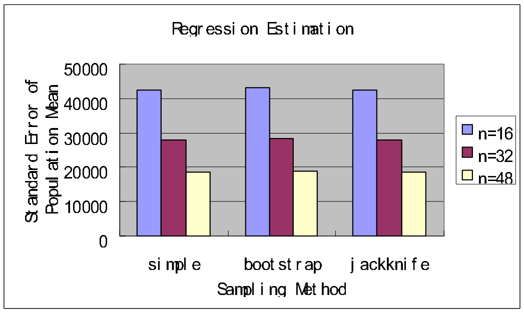

Figure 3: Regression estimation

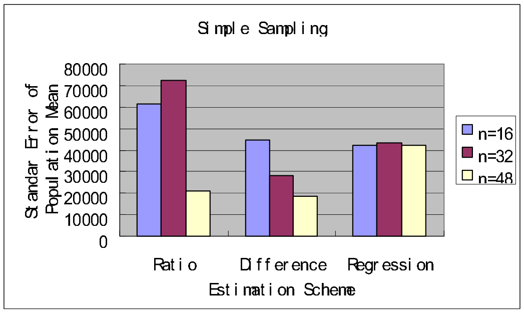

Figure 4: Simple sampling performance

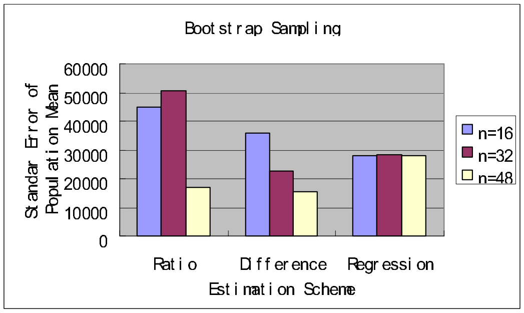

Figure 5: Bootstrap sampling performance

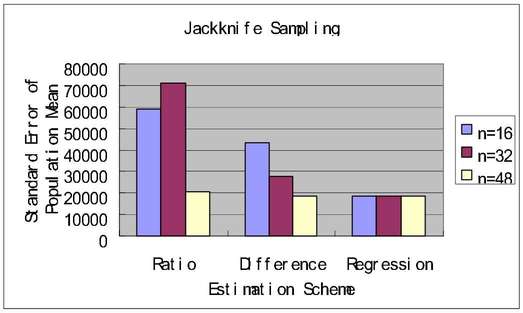

Figure 6: Jackknife sampling performance

3. Ratio estimation is worse than the other two estimation schemes when given the size of the second sample in double sampling and the resampling method. Difference estimation performs nearly the same as regression estimation, but it does better than that of regression estimation when bootstrap resampling method is applied.

3. When the resampling technique is given, the performance of regression estimation is not sensitive to the sample size in the second phase. While the performance of difference estimation is relative to the sample size; the larger the size is, the better the difference estimation performs. No obvious rule can be concluded between ratio estimation and the sampling size in the second phase.

In this experiment, we did not change the parameter how many runs the bootstrap sampling is applied, which is set to a fixed number "20"; hence, different runs of bootstrap method can be tried in future work. Also, only one guess about the coefficient k in difference estimation was tried; so, different guesses can be tried in the future. The more important point is that only one dataset (population) was tested. More datasets should be put into experiment to see to what extent these conclusions hold. We can not tell if these conclusions are useful in practice until we have tried enough different populations.

REFERENCES

Hogg, R.V, McKean, J.W., and Craig, A.T. (2005), *Introduction to Mathematical Statistics,* 6 th ed., New Jersey, Prentice Hall.

Lohr, S. L. (1998), *Sampling: Design and Analysis,* Pacific Grove, CA: Duxbury Press.

Neyman, J. (1938). Contribution to the theory of sampling human populations, *Journal of the American Statistical Association* 33, pp 101-116.

Quenouille, M. (1956), Notes on Bias in Estimation, *Biometrika*, 43, 357-356.