WP.14

ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Madrid, Spain, 20-22 October 2003)

Topic (i): Development and use of data editing quality indicators

TITLE

Performance of Bootstrap Techniques with Imputed Survey Data

Submitted by INE, Spain[1]

I. INTRODUCTION

1. In this study we analyse the performance of the bootstrap, when estimating the variance and a confidence interval of a simple statistic in a survey with nonresponse and mean imputation. The relative bias and root mean square error of the variance estimator and also the coverage rate of the bootstrap confidence interval are calculated and compared with the ones obtained from the jackknife variance estimator in a previous work (Aparicio-Perez, F. and Lorca, D. 2002).

2. The data are taken from the Industrial Business Survey and are exactly the same as in our previous work.

3. In the following section, we explain the bootstrap methods that are used, next we present the simulation study and results and finally some conclusions are provided.

II. BACKGROUND

4. Shao and Sitter (1996) proposed a bootstrap method for variance estimation with imputed survey data. This is the method that we are using. There are another modified bootstrap methods, like the ones proposed in Saigo, Shao and Sitter (2001). See also Lee, Rancourt and Särndal (2001) for a review of variance estimation methods under imputation.

5. Assuming a rather general sampling design, the method consists of the following steps:

a) Let Y_I= $Y_R \cup Y_O$ where Y_R=$\{y_k : k \in A_R(respondents)\}$and Y_O=$\{z_k : k \in A_O(non-respondents)\}$. Y_O is obtained from Y_R using some imputation technique.

b) Let Y^*={y_i^* :i=1,….n} a simple random sample (bootstrap sample) drawn with replacement from Y_I, and Y^*_I= $Y^*_R \cup Y^*_O$ where Y^*_R=$\{y_k^* : k \in A_R^*(respondents)\}$/and

[1] Prepared by Félix Aparicio-Pérez and Dolores Lorca, (fapape@ine.es, mdlorca@ine.es).

Y*$_O$=$\left\{y_k^* : k \in A_O^* (non - respondents)\right\}$ and Y*$_O$ is obtained from Y*$_R$ using the same imputation technique that was used in step a)

c) Obtain the bootstrap estimator $\hat{q}_I^* = \hat{q}(Y_I^*)$ of $\hat{q}_I = \hat{q}(Y_I)$, based on the imputed bootstrap data set Y_I^*.

d) Repeat steps b) and c) B times.The following bootstrap variance estimator for $\hat{q}_I$ is calculated:

$$v_B(\hat{q}_I) = \frac{1}{B}\sum_{b=1}^{B}(\hat{q}_I^{*b} - \bar{q}_I^*)^2$$

where

$$\bar{q}_I^* = \frac{1}{B}\sum_{b=1}^{B}\hat{q}_I^{*b}$$

6. We are also calculating the percentile and the bootstrap-t confidence intervals of the statistic. The former is computed as the empirical confidence interval obtained after sorting the B bootstrap statistics. To build the bootstrap-t confidence intervals, we first compute the statistics $t_b^* = {(\hat{q}_I^{*b} - \bar{q}_I^*)}\big/{s_b^*}$ where $s_b^{*2} = v_B(\hat{q}_I^{*b})$, and then calculate $t_L^* = C\hat{D}F_t^{-1}(a)$, $t_U^* = C\hat{D}F_t^{-1}(1-a)$, where $C\hat{D}F_t^{-1}(x) = \#\left(t_b^*;\ t_b^* \le x,\ b = 1....B\right)/B$. The bootstrap-t confidence interval is then given by $\left(\hat{q}_I - t_U^* v_B(\hat{q}_I), \hat{q}_I - t_L^* v_B(\hat{q}_I)\right)$

III. MONTECARLO STUDY AND RESULTS

7. From our population of N=16438 industrial businesses we draw simple random samples without replacement of sizes n=100, 500, 1000 and 5000. We simulate a loss of about 30 per cent of our data with a uniform mechanism. The number of replications is 50000 for each sample size. The number of bootstrap samples is 999 for each replication.

8. An additional simulation with 23000 replications was conducted only for small sample sizes (n=100) calculating the bootstrap-t confidence interval. We used 999 bootstrap samples in the first level and 50 bootstrap samples in the second level (variance estimation of the first level bootstrap statistics).

9. The analysis variable is the turnover of the businesses. We use mean imputation.

10. Within each replication, we compute the percentage relative bias, the relative root mean square error and the coverage of a nominal 95% bootstrap confidence interval.

TABLE 1. Imputed variable: turnorver.

Imputation method: mean imputation, 95% confidence intervals

Sample size	BOOTSTRAP			BOOTSTRAP			JACKKNIFE		
	T			Percentile					
	RB(%)	MSE	COVR(%)	RB(%)	MSE	COVR(%)	RB(%)	MSE	COVR(%)
100	0.07	4.51	56.9	-0.42	4.47	60.6	0.29	4.52	56.9
500				1.00	2.00	68.3	-0.90	1.96	65.8
1000				4.24	1.45	75.6	-2.12	1.35	72.6
5000				27.15	0.76	91.9	-12.0	0.51	87.3

IV. CONCLUSIONS

11. The results show that the percentile bootstrap performs better than the jackknife for coverage rate of the confidence intervals and the reverse is true for mean square errors and bias of the variance estimators.

12. There seems not to be any advantage in using the bootstrap-t confidence interval, in spite of its higher computational time, though only small sample simulations and small number of second level bootstrap samples are considered for this method.

V. REFERENCES

Aparicio-Pérez, F. and Lorca, D. (2002). Performance of jackknife variance estimation using several imputation methods. UNECE work session on Statistical data Editing. Helsinki.

Lee, H., Rancourt, E. and Särndal, C.E. (2001). Variance estimation from survey data under single imputation. In Survey Nonresponse. Eds. R.M. Groves, D.A. Dillman , J.L. Eltinge and R.J.A. Little, 315-328. Wiley. New York.

Saigo, Shao and Sitter (2001). A repeated half-sample bootstrap and balanced repeated replications for randomly imputed Data. Survey Methodology, vol 27,No. 2,pp.189-196

Shao and Sitter (1996) Bootstrap for imputed survey data. Journal of American Statistical Association,91,1278-1288.