VARIANCE ESTIMATION OF DRUG ABUSE EPISODES USING THE BOOTSTRAP

Arthur L. Hughes and Marc D. Brodsky, National Institute on Drug Abuse
Arthur L. Hughes, 5600 Fishers Lane, Rockville, MD 20857

Key words: Variance estimation, bootstrap, drug-related emergency room episodes, establishment surveys

1. Introduction

This paper examines the feasibility of using the bootstrap methodology to generate variance estimates of drug abuse episodes from a sample of hospital emergency rooms in the United States. The survey data comes from the Drug Abuse Warning Network (DAWN) which is a reporting system designed to monitor new and existing drugs of abuse based on estimates of total drug abuse episodes and the number of mentions of particular drugs. The sample was selected using a single-stage stratified random sample without-replacement scheme in the coterminous U.S. The motivation for conducting a study of this nature is to 1) examine less cost prohibitive alternatives to direct variance estimation methods currently being used on thousands of DAWN characteristics, 2) examine more simplistic variance estimators that could be used in standard software packages, and 3) explore the possibility of implementing nonparametric bootstrap confidence intervals and hypothesis tests (not examined in this paper). The emphasis of this paper will be on the application and evaluation of two methods: 1) a stratified without-replacement bootstrap methodology developed by Sitter (1992) and 2) a more traditional bootstrap method which assumes with-replacement selection and is easier to implement.

2. Background

Efron (1982), Hall (1992), Efron and Tibshirani (1986), and others have provided a great deal of theory and results of the bootstrap technique under the assumption of iid random sampling in a variety of estimation problems (e.g., means, correlation, regression, time series, etc.). However, the use of the bootstrap assuming iid observations may not be realistic when estimation and analysis procedures are based on data from complex sample surveys. In practice, many establishment sample surveys employ single-stage stratified sampling (this is especially true when reasonably complete list frames are available for use). Rao and Wu (1988); Sitter (1992); and Rao, Wu, and Yue (1992) present procedures for using the bootstrap when the sample design is complex.

A popular means for estimating totals and ratios for various characteristics of interests is through the use of the ratio estimator and its associated variance using methods such as the first order Taylor series approximation, generalized variance functions (GVFs), and replication methods (Wolter, 1985). The ratio estimator is popular since its associated variance will be less than the variance of the Horwitz-Thompson estimator when the correlation between the characteristic of interest (Y) and an ancillary variable (X) is relatively high (Cochran, 1977). However, these estimation procedures provide a paucity of information on the sampling distribution of the estimates. Also, there has always been some concern about the potential for the Taylor series variance estimator to exhibit a downward bias (Wolter, 1985).

3. Description of the Drug Abuse Warning Network

Hospital emergency room episodes resulting from the abuse of licit and illicit drugs are collected by DAWN which is a voluntary reporting system sponsored by the Substance Abuse and Mental Health Services Administration within the Department of Health and Human Services. Some of the major objectives of DAWN are: 1) to identify drugs or substances that are currently being abused and 2) to provide data for national and local area drug abuse policy and planning (including scheduling by the Food and Drug Administration). The sample design consists of a stratified random sample of hospitals in 21 metropolitan statistical areas (MSAs) and the residual area of the coterminous U.S. The key characteristic is the total number of drug-related episodes in a year (Y). The ancillary variable used in the ratio estimator is the number of emergency room visits for any reason during the year (X) which is known for all hospitals on the sampling frame. Currently about 500 hospital emergency rooms are participating in the survey (Hughes, et. al., 1991; NIDA, 1991). Each year, in preparation for a publication showing thousands of estimates, variances using Taylor series linearization are calculated in order to ascertain publishability. GVFs could be developed and used as an alternative;

however, given the wide range of values reported and the diversity of the estimates by MSA and demographic characteristics, the use of GVFs may not be adequate. Thus, as an alternative, the bootstrap is considered.

4. Application of the Bootstrap to Stratified Designs

4.1 Description of Procedure

Like other resampling procedures (jackknife, random groups, etc), bootstrap samples should be selected to mirror the original sampling plan. Rao and Wu (1988) proposed a method for selecting bootstrap samples in the stratified without-replacement setting that first involve the transformation of the y's and x's. The key feature about the transformed variables is that they incorporate a finite-correction factor found in without-replacement-design variance estimators. Other scale factors are included to insure that the bootstrap variance is consistent and that the bootstrap-estimated third moment matches the estimated third moment from the original sample.

Sitter suggested a somewhat different procedure for stratified without-replacement sample designs. This procedure eliminates the need to rescale the observations. Instead, within each bootstrap iteration, he selects n_h' samples without replacement k_h times where $n_h'=(n_h/N_h)n_h$ and $k_h=N_h/n_h$ (the sampling weight in stratum h). This results in n_h units selected at each iteration. The procedure is repeated for each stratum B times in an independent manner using the Monte Carlo variance estimator to estimate the variance. Sitter states that this procedure parallels the original sampling plan while matching third moments with the proper choice of the resample size within each stratum.

While these procedures appear to be appropriate for the type of design used in the DAWN study, they may be somewhat difficult to implement in practice. Therefore, in addition to the implementation of Sitter's method, the following procedure was used realizing that the resulting variance may not be asymptotically equivalent to the sampling variance from the original sample without some rescaling of the observations as suggested by Bickel and Freeman (1984) and Rao and Wu (1988); rescaling of the sample weights (Rao, Wu and Yue (1992); or subsampling in the appropriate manner (Sitter, 1992):

1. In each stratum, select a simple random sample of size n_h, with-replacement, from the n_h units in the base sample. A base sample is defined to be the original sample that may be replicated s times using different random starts.

2. Once the resample is selected in each stratum, use the resampled variables (x,y) to calculate the ratio estimate

$$\hat{Y}_R^*=X\frac{\hat{Y}^*}{\hat{X}^*}$$

where

$$\hat{Y}^*=\sum_{h=1}^{L}\sum_{i=1}^{n_h}\frac{N_h}{n_h}y_{hi}$$

$$\hat{X}^*=\sum_{h=1}^{L}\sum_{i=1}^{n_h}\frac{N_h}{n_h}x_{hi}$$

and X is the known population total of the x's. Also calculate the Taylor series variance for each replicated base sample.

3. Repeat steps 1 and 2 independently B times to produce the ratio estimates

$$\hat{Y}^*_{Rs,1},\ \hat{Y}^*_{Rs,2},\ \ldots,\ \hat{Y}^*_{Rs,B}$$

4. Estimate the variance ɥsing the Monte Carlo estimator

$$v_s=\frac{1}{B-1}\sum_{j=1}^{B}(\hat{Y}^*_{Rs,j}-Y_{Rs})^2$$

where Y_{Rs} is the average bootstrapped estimate from the s^th^ base sample.

As for the number of bootstrap samples, Efron and Tibshirani (1986) recommend a value of B between 50 and 200 as adequate for estimating standard errors. They further state that when the CV of the standard error from the full sample is greater than 10 percent, there is little improvement in the CV of the bootstrap estimate of standard error when B is 100 or more. However, if confidence intervals are desired, they claim that at least 1000 bootstrap samples may be needed due to the need to accurately estimate the tails of the sampling distribution generated by the bootstrap process. In literature where examples of simulation studies were given, the base sample was replicated along with the bootstrap samples. It appears that it is useful to do this so that one may be able to compare the distributional behavior of the bootstrap with that from the original sample.

4.2 Preparation of DAWN Data

The 1991 DAWN sampling frame contain the number of hospital emergency room visits (X) for every hospital unit (selected or not selected) and the number of drug related hospital emergency room visits (Y) for the responding hospitals. Logistic regression models were used to impute Y values for the nonrespondents and nonselected units in order to obtain nonmissing (X,Y) pairs for the entire frame (N=5300) while restricting imputed values to be greater then zero. Ideally these models should be DAWN MSA and stratum specific; however, due to the limited number of respondents in numerous cells (n=2 or 3 in many cases) DAWN MSAs were grouped based using SAS PROC CLUSTER. This procedure was used to group the MSAs according to the ratio of $r=\ln(Y/X)$ from the respondents. Following the cluster analysis a t-test was performed to test the null hypothesis that $Ho{:}r_1=r_2= \ldots =r_k$ (i.e., all of the ratios are equal within a cluster of k MSAs). After imputation, the overall correlation between Y and X was 0.54 for the respondents and 0.60 for the full frame.

The base sample was replicated s=100 times with B=100 bootstrap samples generated for each replicated sample. This will allow for observation of the behavior of the sampling distribution of the estimate and variance from the base samples as well as from the bootstrap samples. The original sampling fraction was used to select all of the samples from the original DAWN area and stratum. Because DAWN is a highly stratified sample containing what is believed to be sufficient data on the extremes of the distribution, the original stratum definitions were kept in order to capture the extreme values of Y and X.

5. Results

Tables 1 and 2 presents estimates of total drug-related emergency room abuse episodes and the associated standard errors for the three methods (SIT, BWR, and base sample). The SIT and BWR estimates of total episodes appear to be very similar to those from the base sample and all appear to exhibit a sampling distribution that is skewed to the right. But in spite of this, the standard deviations of the estimates and standard errors from SIT and BWR appear to be higher than the base sample estimates and the skewness estimates are somewhat lower than that for the base sample. The large difference in kurtosis between ORG and the bootstrap methods is a concern, even though these methods were not designed to match fourth moments. In table 3 the average relative bias of the estimated total is reasonably low with little variation; however, for the standard errors, both the SIT and BWR procedures exhibit quite a bit of variation. Also the coverage of the 90, 95 and 99 percent confidence intervals are quite inconsistent (table 4), even for the base sample estimates. This is especially apparent in the SIT procedure and may be in part due to the use of randomization to account for noninteger sample sizes in many of the strata. Increasing the number of samples may help to determine if some inherent problem in the estimation procedures exists. In summary, the BWR appears to be as effective as SIT and is less costly to compute; however, more work is needed before an acceptable method can be determined.

6. Suggested Future Work

The following are some areas that should be pursued:

1. One of the most critical areas requiring further work is to determine the minimum number of bootstrap iterations needed to accurately estimate the tails (2.5 and 97.5 percentiles) of the sampling distribution. Other methods for generating confidence intervals such as the "bias-corrected", "percentile-t", and "other percentile method" (Hall, 1992, chapters 1 and 3) should be investigated. Also, the use of importance sampling (Johns, 1988 and Hall, 1992) in complex sample data should be investigated as a possible means of reducing the number of bootstrap iterations.

2. Investigate the performance of Sitter's method and the BWR with other DAWN survey characteristics, including cocaine, heroin/morphine, and marijuana mentions. Also, the performance of these resampling methods by metropolitan area and demographic characteristics should examined as well.

3. Compare the bootstrap variance estimates to estimates based on other resampling procedures such as the jackknife and balanced repeated replication.

REFERENCES

Bickel, P.J., and Freedman, D.A. (1984), "Asymptotic Normality and the Bootstrap in Stratified Sampling," *The Annals of Statistics*, Vol. 12, No. 2, 470-482.

Cochran, W.G. (1977), *Sampling Techniques*, (3rd edition), John Wiley & Sons, New York.

Efron, B. (1982), "*The Jackknife, the Bootstrap and Other Resampling Plans*," Society for Industrial and Applied Mathematics: Philadelphia, CBMS-NSF, Monograph No. 38, SIAM.

Efron, B. and Tibshirani, R. (1986), "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy," *Statistical Science*, Vol. 1, No. 1, 54-77.

Hall, P. (1992), *The Bootstrap and Edgeworth Expansions*, Springer-Verlag, New York.

Hughes, A.L., Elliott, S.D., Colliver, J.D., and Gruberg, R.E. (1991). "Estimating Drug Abuse Episodes from a Sample of Hospital Emergency Rooms," In Proceedings of the Section of Survey Research Methods, American Statistical Association, 326-331.

Johns, M.V. (1988), "Importance Sampling for Bootstrap Confidence Intervals," *Journal of the American Statistical Association*, Vol. 83, No. 403, 709-714.

National Institute on Drug Abuse (1991), Annual Emergency Room Data-1991, Series I, No. 11-A, DHHS Publication No. (ADM) 92-1955, U.S. Government Printing Office, Washington, D.C.

Rao, J.N.K., and Wu, C.F.J. (1988), "Resampling Inference With Complex Survey Data," *Journal of the American Statistical Association*, Vol. 83, No. 401, 231-241.

Rao, J.N.K., Wu, C.F.J., and Yue, K. (1992), "Some Recent Work on Resampling Methods for Complex Surveys," *Survey Methodology*, Vol. 18, No. 2, 209-217.

Sitter, R.R. (1992), "A Resampling Procedure for Complex Survey Data," *Journal of the American Statistical Association*, Vol 87, No. 419, 755-765.

Wolter, K.M. (1985), *Introduction to Variance Estimation*, Springer-Verlag, New York.

Table 1. Estimates of total drug abuse episodes by sampling method

Method	Mean	Minimum	Maximum	Std. Dev.	Skewness	Kurtosis
SIT	415,645	385,136	461,084	16,872	0.71	0.01
BWR	418,673	387,482	469,938	16,748	0.74	0.16
ORG	416,198	388,796	470,184	14,191	1.06	2.18

NOTE: BWR=with-replacement bootstrap, SIT=Sitter's method, ORG=base sample. Population total=416,066. Estimates for the SIT and BWR procedures are based on the set of 100 weighted sample estimates, where each of these estimates equal the average of 100 estimates within a given base sample. Thus, the SIT and BWR estimates were generated from 10,000 estimates.

Table 2. Estimates of standard error of total drug abuse episodes by sampling method

Method	Mean	Minimum	Maximum	Std. Dev.	Skewness	Kurtosis
SIT	13,527	7,128	28,466	5,362	1.18	0.34
BWR	14,223	7,118	30,559	5,420	1.27	0.70
ORG	12,354	6,804	29,357	4,487	1.76	3.34

NOTE: BWR=with-replacement bootstrap, SIT=Sitter's method, ORG=base sample. The true standard error is assumed to be equal to 12,354, which is the average standard error over the 100 base samples. Estimates for the SIT and BWR procedures are based on the set of 100 Monte Carlo standard errors, where each of the 100 estimates equal to the average of 100 variance estimates within a given base sample (followed by taking the square root). Thus, the SIT and BWR estimates were generated from 10,000 estimates.

Table 3. Percent relative bias of estimates and standard errors of total drug abuse episodes by sampling method

Method	Mean	Minimum	Maximum	Std. Dev.
Estimates				
SIT	-0.10	-7.43	10.82	4.06
BWR	0.62	-6.87	12.95	4.03
ORG	0.03	-6.55	13.01	3.41
Standard errors				
SIT	9.49	-42.31	130.42	43.41
BWR	15.13	-42.38	147.36	43.87
ORG	0.00	-44.93	137.63	36.32

NOTE: BWR=with-replacement bootstrap, SIT=Sitter's method, ORG=base sample. Population total=416,066. The true standard error is assumed to be equal to 12,354, which is the average standard error over the 100 base samples.

Table 4. Percent of 100 original samples for which the 100(1-α)% confidence interval does not contain the population value

	Confidence coefficient		
Method	100(1-α)=90%	100(1-α)=95%	100(1-α)=99%
SIT	24	20	6
BWR	20	10	1
ORG	14	10	3

NOTE: BWR=with-replacement bootstrap, SIT=Sitter's method, ORG=base sample.