Inferring Random Samples from a Cluster Sample

Hee-Choon Shin
NORC, 55 E. Monroe St., Suite 4800, Chicago, IL 60603

Key Words: Variance, Complex sample, Bootstrap, Resampling

I. Introduction

Most of the applied statistical techniques assumed the existence of an independent and identically distributed (*iid*) sample. Correspondingly, the major commercial statistical analysis packages (e.g., SAS and SPSS) were developed to apply these techniques to real data. In sample surveys, however, most samples are complex rather than simple (Cochran, 1977; Kish, 1965; Wolter, 1985). In other words, most of survey data are dependent. What is more, the depth of statistical theories on the interrelated or dependent data is very shallow. And, in general, substantive analysts are not willing to use some specific techniques which have been forwarded within survey research community. Also, based on these techniques, some specific analysis packages (e.g., SUDAAN, PCCARP, and WESVAR) were developed to deal with the interdependency.

Our goal in this research is to find an alternative way to analyze complex sample surveys with the assumption of independent samples. If we can derive independent random samples from a given complex sample, the standard techniques can be directly applied to each sample by using a standard statistical software. Our approach can be considered as a variation of various re-sampling methods (Efron, 1982; Kovar, Rao, and Wu, 1988; Shao, 1996; Sitter, 1992).

II. Independent Sub-samples

To suggest out approach, let us consider a stratified two-stage sample design. The population under consideration is stratified into L strata with N_h clusters in the hth stratum. For each stratum, n_h clusters are independently selected. Within ith cluster in hth stratum, n_{hi} individuals are independently sampled. The total number of individuals in the sample is $n = \sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} n_{hij}$.

Correct formulae for the estimates and their variances under various specific sample designs can be found in Cochran (1977), Kendall and Stuart (1976), Kish (1965), and Wolter (1985) among many others.

Our approach to data analysis is to derive a set of independent sub-samples from a given cluster sample. Then, each of sub-sample can be analyzed as an independent sample. Consider a sub-sample of individuals with one and only one individual from the ith cluster in hth stratum. The αth sub-sample is

$$s_\alpha \ni y_{hi.\alpha_{hi}}\ (h=1,2,\cdots,L;\ i=1,2,\cdots,n_h),$$

where the randomly (uniform) generated number α_{hi} is $1 \le \alpha_{hi} \le n_{hi}$.

For a moment, assume $n_{hi} = n^*$. If we sub-sample without replacement, there will be n^* possible sub-samples. The estimated mean for the αth sub-sample is

$$\bar{y}_\alpha = \frac{\sum_{h=1}^{L}\sum_{i=1}^{n_h} y_{hi\cdot\alpha_{hi}} w_{hi\cdot\alpha_{hi}}}{\sum_{h=1}^{L}\sum_{i=1}^{n_h} w_{hi\cdot\alpha_{hi}}},$$

where $w_{hi\cdot\alpha_{hi}}$ is the weight. The estimated mean for individual level is

$$\bar{\bar{y}} = \frac{\sum_{\alpha=1}^{n^*}\bar{y}_\alpha}{n^*}.$$

The estimated variance is

$$var(\bar{\bar{y}}) = \frac{(\bar{y}_\alpha - \bar{\bar{y}})^2}{n^*-1}.$$

Now let $n_{hi} \ne n^*$ and consider the concept of mass in physics. The mass is

$$m = \delta V,$$

where δ is the density and V is the volume. In our case, δ is the weight and V can be regarded as the size of each cluster, i.e., n_{hi}. The estimated mean for the αth sub-sample with unequal cluster size is

$$\bar{y}_{\alpha} = \frac{\sum_{h=1}^{L}\sum_{i=1}^{n_h} y_{hi\cdot\alpha_{hi}} w_{hi\cdot\alpha_{hi}} n_{hi}}{\sum_{h=1}^{L}\sum_{i=1}^{n_h} w_{hi\cdot\alpha_{hi}} n_{hi}}.$$

The above steps can be repeated k times to obtain $\bar{\bar{y}}_{(k)}$ and $var(\bar{\bar{y}}_{(k)})$.

Beyond the simple discussion of means, let us consider regression analysis. The major difficulty of applying the standard regression analysis to a survey data with a cluster design is the presence of *dependent* observations. To be able to specify the likelihood function, all the variance-covariance structure (under a certain model) should be known. All the known techniques including balanced replication, Taylor series, and jackknife methods cannot completely deal with the interdependency. Meanwhile, the standard generalized regression techniques can be directly applied to our independent sub-samples.

III. Examples and Discussion

Data are from the National Survey of Family Growth, 1995. We are interested in the estimation of the average number of male sexual partners of U.S. women (15-44) and their age at the first sexual intercourse. Also, we are interested in the effect of age at first sex on the number of male sexual partners. Table 1 shows the means and their standard errors, and Table 2 shows the regression coefficients and their standard errors. The example shown in the Tables supports the effectiveness of our approach as compared to other alternatives. As we see in the Tables, direct application of standard statistical software (i.e., SAS or SPSS) results in the underestimation of standard errors in relation to means and regression coefficients. However, we do not see the advantage of using a special software (e.g., SUDAAN) to analyze the clustered data.

In addition to the problem of interdependency of data, there is a limit in specifying the design parameters. Because of the confidentiality restriction or the presence of disclosure risk, major surveys does not provide the full design structure. In many cases, the design is too complex to specify all the aspects of sample design. Indeed, a particular specific software does not have all the possible options to take into account the complex sample design. Therefore, for example, applying Taylor series method to these data is at best an approximation. *An approximation* of an approximation since Taylor series is an approximation of a function.

Reference

Cochran, William G. (1977). *Sampling Techniques, Third Edition.* John Wiley & Sons: New York.

Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans.* SIAM:
Philadelphia.

Kendall, Maurice and Alan Stuart. 1976. *The Advanced Theory of Statistics, Vol. 3. Third Ed.* New York: Hafner.

Kish, Leslie (1965). *Survey Sampling.* John Wiley & Sons: New York.

Kovar, J.G., J.N.K. Rao, and C.F.J. Wu (1988). "Bootstrap and other methods to measure errors in survey estimates." *The Canadian Journal of Statistics* 16: 25-54.

Shao, Jun (1996). "Resampling methods in sample surveys." *Statistics* 27: 203-254.

Sitter, *R.R.* (1992). *"Comparing three bootstrap methods for survey data." The* Canadian Journal of Statistics 20: 135-154.

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

Table 1. Means and standard errors of the number of male sexual partners and age at first sexual intercourse, U.S. women 15-44.

		Number of male sexual partners during the last 12 months		Age at the first sexual intercourse	
		mean	standard error	mean	standard error
SAS (unweighted)		1.1672	.0211	17.3423	.0326
SAS (weighted)		1.1641	.0235	17.3967	.0328
Taylor series meth.		1.1641	.0275	17.3967	.0445
Sets of independent sub-samples (k)	10	1.1542	.0231	17.3839	.0429
	50	1.1542	.0227	17.3971	.0441
	100	1.1687	.0270	17.3965	.0445
	200	1.1630	.0250	17.3965	.0436
	500	1.1642	.0251	17.3956	.0446

Table 2. Simple regression coefficients and standard errors: the effect of age at the first sex on the number of male sexual partners.

		Coefficient	Standard error
SAS (unweighted)		- .0569	.0066
SAS (weighted)		- .0606	.0073
Taylor series method		- .0606	.0139
Sets of independent sub-samples (k)	10	- .0577	.0105
	50	- .0579	.0105
	100	- .0640	.0133
	200	- .0616	.0120
	500	- .0617	.0122