The Stata Journal

H. Joseph Newton
Department of Statistics
Texas A & M University
College Station, Texas 77843
979-845-3142
979-845-3144 FAX
jnewton@stata-journal.com

Nicholas J. Cox
Department of Geography
University of Durham
South Road
Durham City DH1 3LE
United Kingdom
n.j.cox@stata-journal.com

The *Stata Technical Journal*, electronic version (ISSN 1536-8734) is a publication of Stata Press, and Stata is a registered trademark of Stata Corporation.

The Stata Journal (2002)
2, Number 2, pp. 190–201

From the help desk: It's all about the sampling

Allen McDowell
Stata Corporation
amcdowell@stata.com

Jeff Pitblado
Stata Corporation
jsp@stata.com

Abstract. Effective estimation and inference, when the data are collected using complex survey designs, requires estimators that fully account for the sampling design. This article explores, by means of Monte Carlo simulations of the power of simple hypothesis tests, the consequences of parameter estimation and inference when naive estimators are employed with survey data.

Keywords: st0016, cluster, design, power, strata, svy, svymean, svyset

1 Introduction

One of the most important concepts in statistics is the Central Limit Theorem. Informally, this theorem states that if we know the value of the variance S^2 of a random variable and take a sample of n independent measurements $(y_1, y_2, \ldots, y_n)$ on that variable, the distribution of the sample mean $\overline{y}$ may be approximated by the normal distribution with mean μ (the population mean) and variance $S_d^2 = S^2/n$. Although S_d^2 is typically unknown, we can usually estimate the variance of $\overline{y}$ and form a t or F statistic to test $H_0 : \mu = \mu_0$. Methods that directly employ the Central Limit Theorem rest on the assumption that the data were collected in the form of a simple random sample. Survey data, however, are typically collected according to complex designs involving stratification, clustering at one or more stages, and weighted sampling. In this article, we use Monte Carlo methods to investigate the effect that survey designs have on the significance level and power of a hypothesis test involving the population mean.

While there are many sampling designs to choose from, each having its own estimator for the population mean, we will limit our discussion to four simple designs: simple random sampling (SRS), stratified simple random sampling (STR-SRS), cluster sampling (PSU), and stratified-cluster sampling (STR-PSU). These designs, while not permitting us to demonstrate all of the possible pitfalls one might encounter when estimating a population parameter from survey data, will enable us to adequately demonstrate the importance of accounting for the sampling design. Failure to fully account for the sampling design when estimating population parameters can result in biased estimators of population parameters or biased variance estimators. A statistic used to estimate a population parameter is said to be unbiased if the mean of its repeated-sampling distribution is equal to the parameter of interest. A variance estimator of a statistic is said to be unbiased if the mean of its repeated-sampling distribution equals the repeated-sampling variance of the statistic of interest. In Stata, one specifies the sampling design prior to estimation using the `svyset` command.

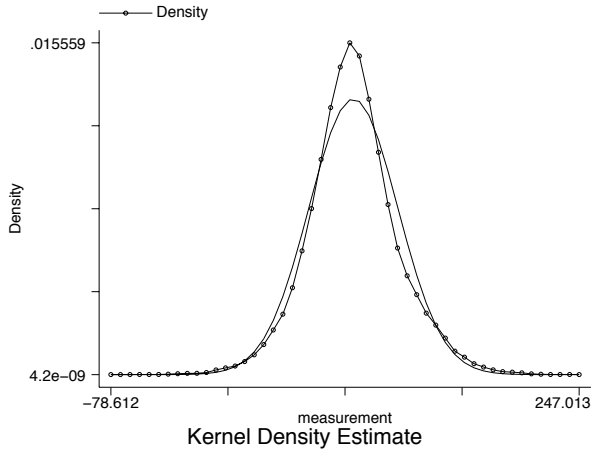

Figure 1: Population density estimate compared to a normal density.

1.1 Population under study

The Monte Carlo studies for each of these designs are based on repeated sampling from the same simulated population. This population was designed to have 4 strata, each containing 200 clusters of individuals. Each cluster contains 30 individuals, and thus our population is balanced and contains $N = 4 \times 200 \times 30 = 24{,}000$ observations. The population was purposely generated to possess cluster and stratification effects in the measurement y in order to fully appreciate the need to understand design effects. We first generated data according to the model

$$x_{hij} = \mu_h + u_{hi} + e_{hij}$$

where μ_h is constant within stratum according to

$$\mu_h = \begin{cases} -10, & h = 2 \\ 15, & h = 3 \\ 0, & \text{otherwise} \end{cases}$$

and $h = 1, \ldots, 4$, $i = 1, \ldots, 200$, and $j = 1, \ldots, 30$. The random cluster means u_{hi} were generated from a normal distribution with zero mean and a variance of 16, and the e_{hij} are from a normal distribution with zero mean and a variance of $100 + 100h^2$. The population values used in the Monte Carlo studies were the values $y_{hij} = \mu + x_{hij} - \overline{x}$, where $\overline{x}$ is the average of the x_{hij}'s, and $\mu = 90$. Thus, by construction, the distribution of y is a mixture of normal random variables; Figure 1 depicts a kernel density estimate of the distribution of y compared to a normal density. From this point forward, the population is fixed for the purposes of sampling. The standard deviation of our population is $S = 30.86$.

1.2 Significance level and power

We are interested in how misspecifying the sampling design affects the power of testing $H_0 : \mu = \mu_0$, where μ_0 is the hypothesized mean. For theoretical results on these issues, see Cochran (1977), Kish (1965), Korn and Graubard (1999), Levy and Lemeshow (1999), and Thompson (1992).

We chose an evenly spaced grid of hypothesized means over the interval $\mu \pm 10S_d$, where S_d^2 is the variance of the mean estimator for a given sampling design. The power can then be estimated by the proportion of times H_0 was rejected. Each proportion is based on the two-sided t test using the results from `svymean` applied to 1000 samples chosen randomly according to the sampling design under study. All tests in the study will be performed with a significance level $\alpha = .05$. Since we have our population in a Stata dataset, we know with certainty the value of the population mean and can compute the variance of the mean estimator for each sampling design, implying that we can test H_0 using the normal distribution. The power of a two sided Z test is

$$\beta(\mu_0) = \Phi\left(-z_{\alpha/2} - \frac{\mu_0 - \mu}{S_d}\right) + \Phi\left(-z_{\alpha/2} + \frac{\mu_0 - \mu}{S_d}\right) \tag{1}$$

where $z_{\alpha/2}$ is the $100(1 - \alpha/2)$ percentile of the standard normal distribution, and Φ is the standard normal CDF.

We use the power curve based upon the normal distribution as the standard for comparison. A power curve generated from employing a biased estimator will exhibit a horizontal location shift relative to the Z test power curve. Power curves resulting from biased variance estimators will exhibit nonhomothetic vertical shifts relative to the Z test power curve.

2 Simple random sampling

There are $\binom{N}{n}$ possible samples of size n that can be drawn from a population of size N using simple random sampling without replacement. For this design, each individual in the population has an equal chance of being observed. Data from this design is the least difficult to analyze, since most of the results we learned in our introductory statistics courses apply. The sample mean estimates the population mean without bias, and its variance is

$$S_d^2 = \frac{S^2}{n}(1 - f)$$

where $f = n/N$ is called the sampling fraction or finite population correction (FPC). Ignoring the fact that the sample is being drawn from a finite population results in an estimate of variance that is biased upwards. However, if the sample size is a small percentage of the population size, the bias becomes negligible, and the FPC can be effectively ignored. With the variance estimated by

$$s_d^2 = \frac{s^2}{n}(1 - f)$$

the usual t statistic will tend to follow Student's t distribution with $df = n - 1$ degrees of freedom.

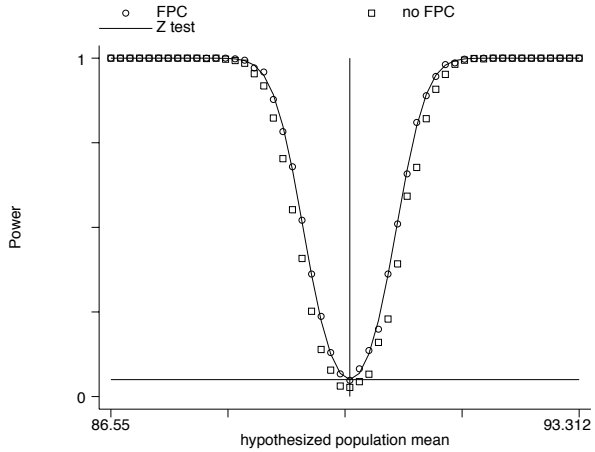

Figure 2: Monte Carlo power compared with the power from a Z test for the SRS design.

Figure 2 exhibits three power curves generated for this design using $f = 0.25$. The Z test power curve, $\beta(\mu_0)$ from (1) with $S_d = 0.3450$, is plotted using a solid line. The circles plot the estimated power for the case where the FPC was specified. The squares plot the estimated power where the FPC was not specified. Note that the test that excludes the FPC (represented by the squares) does not have the correct significance level. Due to the large sample size, it is not surprising to see that the power curve for the estimator that includes the FPC, which has the correct significance level, closely approximates the power curve based on the Z test.

3 Stratified simple random sampling

In the stratified SRS design, the population is partitioned into L mutually exclusive and exhaustive strata, thus $N = N_1 + \cdots + N_L$, with N_h denoting the number of individuals in stratum h. The strata samples are then independently drawn according to the SRS design within each stratum. The survey designer identifies the strata and the strata sample sizes n_h to be drawn from each stratum. If stratification is being employed in order to improve estimation precision, some prior information regarding the population is implied. There are $\binom{N_1}{n_1}\binom{N_2}{n_2}\ldots\binom{N_L}{n_L}$ possible samples of size $n = n_1 + \cdots + n_L$ that can be drawn from the population with the STR-SRS design. Since the sample space is reduced compared to that of the simple random sampling design, individuals may no longer have the same probability of being selected. In fact, each stratum has its own sampling fraction, $f_h = n_h/N_h$. Thus, except for the special case when the sampling fractions are equal across strata, the inclusion probabilities are not the same across all strata. Failing to account for this will bias the estimator for the population mean.

Weighting each observation by the inverse of its probability of inclusion will eliminate the bias; these weights are called probability sampling weights or `pweights` in Stata. Failing to account for potential differences in location or scale across strata will result in biased variance estimators that will tend to be larger than that of the variance of the mean for this design.

The mean estimator for this design is

$$\overline{y}_{st} = \sum_{h=1}^{L} W_h \overline{y}_h$$

where $\overline{y}_h$ is the mean for stratum h and $W_h = N_h/N$. The W_h are a result of correctly specifying `pweights`. The variance of this estimator is

$$S_d^2 = \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{n_h}(1 - f_h) \tag{2}$$

where S_h denotes the standard deviation of the population values in stratum h. The variance estimator results from estimating S_h from the sample values in stratum h. The degrees of freedom for this variance estimator is typically taken to be $df = n - L$.

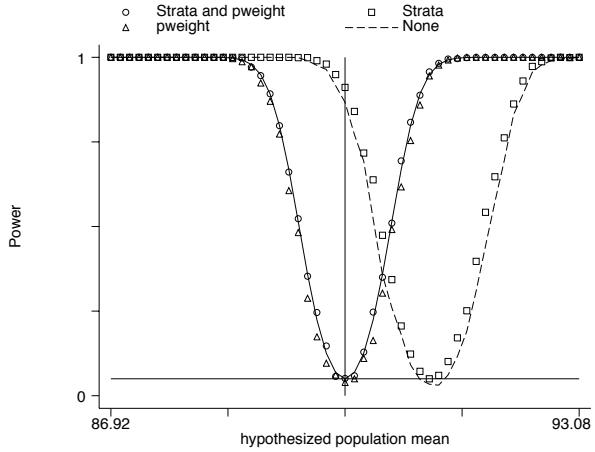

Figure 3: Monte Carlo power compared with the power from a Z test for the STR-SRS design.

The results of the power study for this design are shown in Figure 3. Here we drew samples with stratum sampling fractions according to $f_1 = .10$, $f_2 = .20$, $f_3 = .30$, and $f_4 = .40$. The Z test power curve, with $S_d = 0.3084$, is depicted as a solid line. The power values plotted using a dashed line depict the results from specifying FPC only. The squares depict the results from specifying both strata and FPC, but not `pweights`. The triangles depict the results from specifying `pweights` and FPC, but not strata. The circles depict the results from fully specifying the design.

The bias in the estimated mean caused by not using sampling weights is clearly illustrated by the rightward location shifts of the dashed line and squares relative to the Z test power curve. The triangles exhibit a significance level that is too low, indicating that the variance estimators are biased upward. Finally, the fact that the circles closely approximate the Z test power curve demonstrates that specifying stratification along with proper weighting results in an unbiased estimator of the population mean with a variance estimator that provides the appropriate significance level.

4 One-stage clustered designs

In one-stage cluster sampling designs, the population is partitioned into groups of individuals called clusters or primary sampling units (PSU). A simple random sample of n clusters is then taken from the population, and all individuals within a sampled cluster are observed. Thus, it is the cluster as a whole that contributes information instead of the individuals. Suppose we have a population where we identify N clusters, each with M individuals, and denote y_{ij} as the observed value for the jth individual within the ith cluster. An unbiased estimator for the population mean is

$$\overline{\overline{y}} = \frac{1}{nM} \sum_{i=1}^{n} \sum_{j=1}^{M} y_{ij}$$

The variance of this estimator is

$$S_d^2 = \frac{1-f}{nM^2} \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \overline{y})^2 \tag{3}$$

where y_i is the total for cluster i and $\overline{y}$ is the average of the cluster totals. From our population, we keep the clusters identified even as we ignore the strata; thus, there are $N = 4 \times 200 = 800$ clusters in the population, each containing $M = 30$ individuals.

There are $\binom{N}{n}$ possible samples. Except for the special case where all of the clusters are of equal size, the number of observations in the sample will not be equal; the sample size becomes a random variable. As a consequence of this sampling design, individuals within a cluster are dependent upon each other with respect to the probability of being sampled. Note that this alone does not appreciably affect the sampling distribution of the mean; however, the association of individuals within each cluster does.

So the question is: How does this within-cluster association affect the variance of our mean estimator? We will try to explain this using intracluster correlation. Cochran (1977) defines the intracluster correlation as "the correlation coefficient between pairs of units that are in the same systematic sample". By this definition, intracluster correlation is

$$\rho = \frac{E(y_{ij} - \overline{\overline{Y}})(y_{ik} - \overline{\overline{Y}})}{E(y_{ij} - \overline{\overline{Y}})^2} = \frac{2 \sum_i \sum_{j<k} (y_{ij} - \overline{\overline{Y}})(y_{ik} - \overline{\overline{Y}})}{(M-1)(NM-1)S^2} \tag{4}$$

where $\overline{\overline{Y}}$ denotes the population average per individual ($\overline{\overline{Y}} = \mu$), and S^2 is the variance among the individuals. Correlation coefficients take on the values from -1 to 1, where

values near -1 indicate very strong negative associations, values near 1 indicate very strong positive associations, and values near 0 indicate no association at all. Let's look at some simulated populations to see what values ρ can take on, and how to interpret these values.

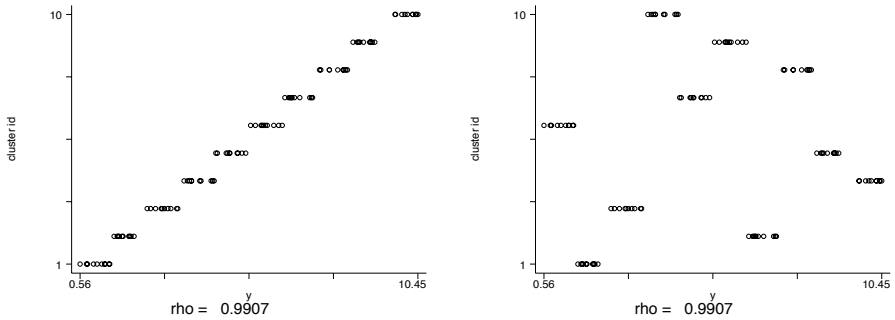

Figure 4: Scatter plots of cluster id i by measure y_{ij}. Left: Population with clusters from the uniform distribution centered at the cluster id and with unit range. Right: Same population with cluster id permuted.

Figure 4 contains two scatter plots of the cluster id versus y. The y_{ij} in the left plot are simple random samples from the continuous uniform distribution centered at i (the cluster id). The right plot contains the same data but with the cluster id randomly assigned. The relationship between the values in y_{ij} and the artificial cluster id i do not play a role in the value of $\rho = 0.99$. The fact that the clusters do not overlap gives us a visual clue that there is a strong positive intracluster association.

(*Continued on next page*)

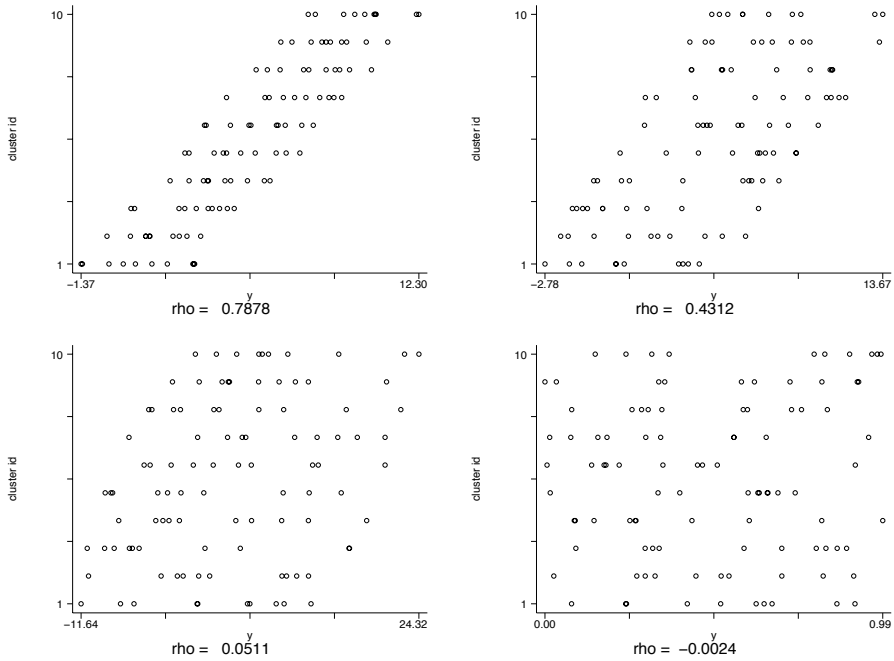

Figure 5: Scatter plots of cluster id i by measure y_{ij}. Top left: Clusters centered at i with range 5. Top right: Clusters centered at i with range 10. Bottom left: Clusters centered at i with range 30. Bottom right: Clusters each iid samples from the uniform.

Figure 5 contains similar plots but with increasing amounts of overlap. This was accomplished by changing the range of the uniform distribution we used to generate the y_{ij} from 5 (top left) to 10 (top right) to 30 (bottom left). The remaining plot contains clusters taken from the uniform distribution without shifting or rescaling. Thus, as we look at these plots from left to right and top to bottom, we can see the intracluster association getting weaker and weaker.

(*Continued on next page*)

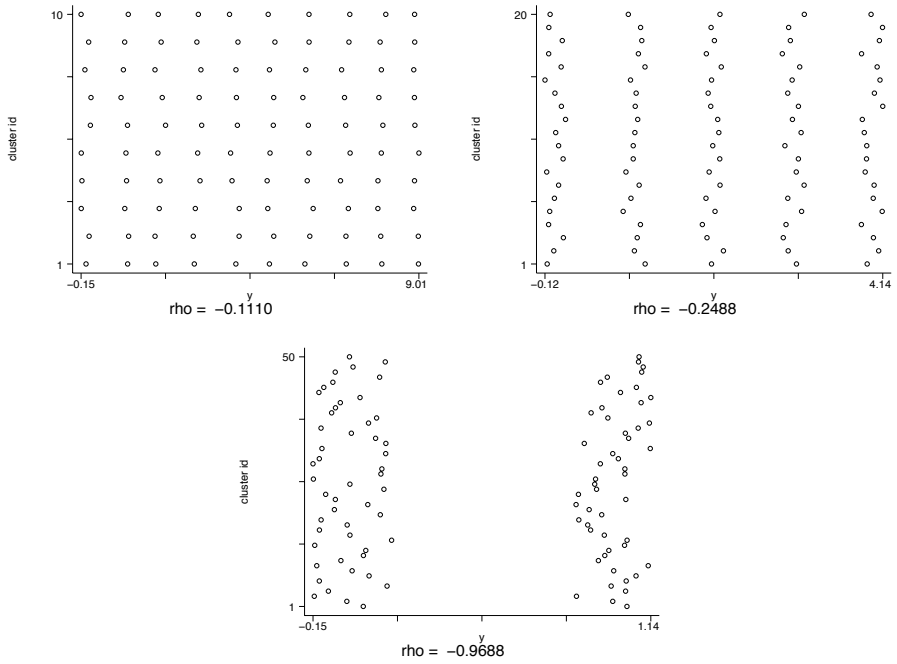

Figure 6: Examples of negative intracluster association.

It may not seem so, but it is possible for clustering to result in sampling distributions that are less variable than those from simple random sampling. Consider a population where y_{ij} is from the uniform distribution centered at j instead of i and with range 0.3. Figure 6 contains plots of data from three such populations. The top left plot comes from data generated with 10 individuals within each cluster, the top right from a population with 5 individuals within each cluster, and the bottom with 2 individuals per cluster. In this case, the clusters are so similar to each other that the negative intracluster association becomes stronger as the number of individuals per cluster decreases. Perfect negative association is achieved only when all the clusters contain the same two distinct values.

Now, Theorem 9.2 of Cochran (1977) shows how the variance of the mean estimator for this design is related to ρ,

$$S_d^2 = \frac{1-f}{nM}\frac{NM-1}{M(N-1)}S^2\{1+(M-1)\rho\}$$

where nM is the sample size. Thus, positive intracluster associations result in larger variances for the PSU design than for the SRS design. As a consequence, not specifying the clusters will result in variance estimators that are biased downward. Conversely, negative intracluster associations result in smaller variances for the PSU design than for the SRS design, so not specifying the clusters will result in variance estimators that are biased upward. For our simulated population described in Section 1.1, $\rho = 0.1019$, so

we would expect our variance estimates to be too small if clusters are not specified.

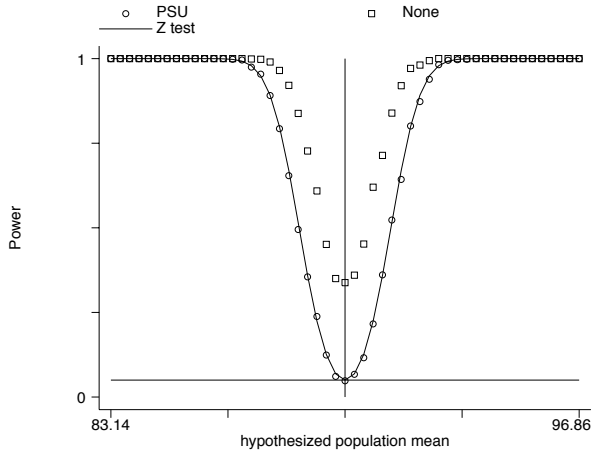

Figure 7: Monte Carlo power compared with the power from a Z test for the PSU design.

The results of the power study for this design are shown in Figure 7; we used a sampling fraction of $f = 0.25$. The Z test power curve, with $S_d = 0.6864$, is depicted as a solid line. The squares plot out estimated power where only the FPC was specified. The circles represent the case where both the PSU and FPC were specified. Note that sampling weights are unnecessary for estimating the mean since all of our clusters have the same chance of being observed. The fact that the estimated power plotted by the squares is so far above the Z test power curve clearly indicates that the variance estimator is biased downwards. As a result, we are rejecting the null hypothesis far too often when it is true.

5 Stratified-cluster sampling

In a stratified and clustered design, the population is, as in the case of the STR-SRS design, segmented into L mutually exclusive and exhaustive strata, and stratum h has N_h clusters. The sampling then involves taking an independent simple random sample of clusters from each stratum. The survey designer is responsible for choosing the number of clusters n_h to be sampled from stratum h. Once a cluster is selected, every individual within the cluster is observed. In the case of the stratified one-stage clustered design, there are $\binom{N_1}{n_1}\binom{N_2}{n_2}\ldots\binom{N_L}{n_L}$ possible samples that can be drawn from the population. Except for cases like ours where all of the clusters are of equal size M, the total sample size is dependent upon the particular sample drawn.

The mean estimator for this design is given by

$$\overline{\overline{y}}_{st} = \sum_{h=1}^{L} W_h \overline{\overline{y}}_h \tag{5}$$

where $\overline{\overline{y}}_h$ is the mean of the observations from stratum h and $W_h = N_h/N$. Its variance is

$$S_d^2 = \sum_{h=1}^{L} W_h^2 \frac{1-f_h}{n_h M^2} \sum_{i=1}^{N_h} \frac{(y_{hi} - \overline{y}_h)^2}{N_h - 1} \tag{6}$$

where $f_h = n_h/N_h$ is the sampling fraction for stratum h, y_{hi} is the total of the measurements in cluster i of stratum h, and $\overline{y}_h$ is the average of the cluster totals within stratum h.

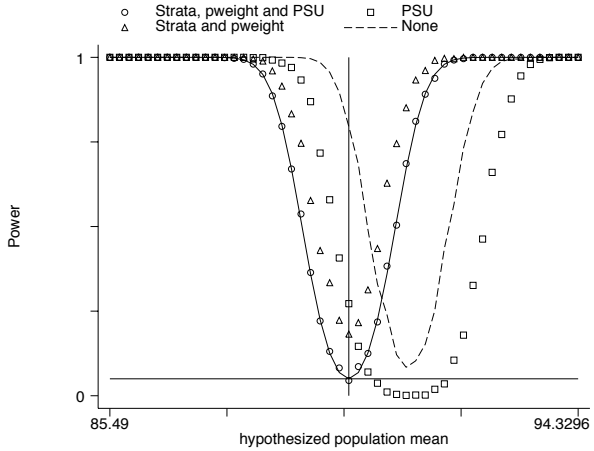

Figure 8: Monte Carlo power compared with the power from a Z test for the STR-PSU design.

The results of the power simulations for this design are shown in Figure 8. Here, as in STR-SRS, we drew samples according to $f_1 = .10$, $f_2 = .20$, $f_3 = .30$, and $f_4 = .40$. As in all the previous graphs, the Z test power curve, with $S_d = 0.4507$, is represented by a solid line. The power values plotted using a dashed line depict the results from specifying only the FPC. The rightward and upward shifts of this curve, relative to the Z test power curve, indicate that this mean estimator is biased and results in a variance estimator that is biased downwards. The squares depict the results from specifying the PSUs and the FPC. Here too we see that the mean estimator is similarly biased, but that the variance is now biased upwards. The triangles depict the results from specifying strata, `pweights`, and the FPC. The mean estimator in this case is unbiased, but the variance is biased downwards. Finally, the circles depict the results from fully specifying the design. As for the previous designs, the circles closely approximate the Z test power curve, indicating that the mean estimator and its variance estimator are unbiased.

6 Conclusion

Table 1 contains summary information from the Monte Carlo studies where the sampling designs were fully specified. It contains the values of S_d, the standard error of the mean

estimator, for each design along with the associated degrees of freedom. A power curve evaluated at the true population mean represents the achieved significance level of the test performed. Let $\widehat{\beta}(\mu)$ denote the Monte Carlo estimator of power evaluated at the true population mean. Since all of the tests performed were based on a 5% significance level, $\widehat{\beta}(\mu)$ is just an estimator of 0.05. The evidence supports the conclusion that, for the simple designs considered here, the estimators from the fully specified designs produce unbiased estimates and variances such that inference for a specified significance level can be achieved with correct probability coverage. It was also clearly demonstrated that estimators that do not fully take into account the sampling design are biased or have biased estimates of variance. Thus, statistical inference based upon such estimators will be flawed.

Table 1: Overview of results

				Binomial Exact	
Design	S_d	*df*	$\widehat{\beta}(\mu)$	95% Confidence Interval	
SRS	0.3450	5999	0.048	0.0356001	0.0631401
STR-SRS	0.3084	5996	0.050	0.0373352	0.0653910
PSU	0.6864	199	0.048	0.0356001	0.0631401
STR-PSU	0.4507	196	0.045	0.0330098	0.0597525

7 References

Cochran, W. G. 1977. *Sampling Techniques*. 3d ed. New York: John Wiley & Sons.

Kish, L. 1965. *Survey Sampling*. New York: John Wiley & Sons.

Korn, E. L. and B. I. Graubard. 1999. *Analysis of Health Surveys*. New York: John Wiley & Sons.

Levy, P. S. and S. Lemeshow. 1999. *Sampling of Populations*. New York: John Wiley & Sons.

Thompson, S. K. 1992. *Sampling*. New York: John Wiley & Sons.

About the Authors

Allen McDowell is Director of Technical Services at Stata Corporation.

Jeff Pitblado is a Senior Statistician at Stata Corporation.