REPLICATE WEIGHTING METHODS FOR QUANTILE VARIANCE ESTIMATION *

Kevin W. Dodd, Iowa State University Alicia L. Carriquiry, Iowa State University
Wayne A. Fuller, Iowa State University
Kevin W. Dodd, Statistical Laboratory, Iowa State University, Ames, IA 50011

KEY WORDS: Measurement error, Survey sampling

1 Introduction

Data from complex surveys are often used to efficiently obtain estimates of population parameters. Variances of the estimators depend on the survey design and are often difficult to obtain analytically, especially for nonlinear statistics such as ratios or sample quantiles. Traditional methods for variance approximation, as discussed by Wolter (1985), involve one of two strategies: Taylor linearization or replication methods such as the bootstrap, the jackknife, and balanced repeated replication (BRR). In cases where the estimators have a complex form (as in the case of quantiles), replication methods are preferred by virtue of being easier to implement. A sample quantile is the classic example of a nonsmooth estimator for which the standard delete-1 jackknife is known to give an inconsistent variance estimator (Shao and Wu, 1989). However, asymptotic consistency of BRR variance estimators for sample quantiles has been established by Shao and Wu (1992) for the case where the data are observed without error.

The problem we address in this paper arises in the analysis of food consumption data. The United States Department of Agriculture conducts survey to assess the dietary adequacy of the population. An important concept in analyzing data from these surveys is that of *usual intake*, defined as the long-run average daily intake of a dietary component by an individual. To estimate distributions of usual intake, surveys collect daily intake measurements on individuals for a small number of days. Due to the small number of observations per individual, the distribution of individual mean intakes performs poorly as an estimate of the distribution of usual intakes. This is because the variance of the mean of a few daily intakes contains a sizable amount of within-individual variation. If we assume that daily intakes of a dietary component for an individual measure the individual's usual intake with error, the problem of estimating the distribution of usual intakes can be thought of as the problem of estimating the distribution of a random variable that is observed subject to measurement error. Once an estimator of the the usual intake distribution is obtained, we can estimate the proportion of the population with usual intake of some nutrient (say calcium) below a given level by evaluating the estimated distribution function at that value.

Several characteristics of dietary intake data make statistical analysis difficult. Intake data are nonnegative, and the distributions of both daily intakes and individual mean intakes are often highly skewed. Nuisance effects are often present in the data; daily consumption patterns differ according to day-of-week and month-of-year. Within-individual variances may vary across individuals, suggesting that the measurement error variance is not constant. Nusser et al. (1996) propose a method, based on semiparametric transformations, to estimate usual intake distributions of dietary components consumed on a daily basis. The proposed method has several steps and addresses attributes of the data mentioned above. A software package, C-SIDE, was developed to implement the procedure. The current version of C-SIDE estimates standard errors of usual intake percentiles using a form of Taylor linearization. These standard errors were derived under the assumption of simple random sampling and are not expected to perform well in the case of complex survey design. In this paper, we explore the use of balanced repeated replication and the jackknife as robust, computationally inexpensive alternatives to the Taylor linearization method in the special case of two-cluster-per-stratum designs. Our validation tool will be Monte Carlo simulation.

*This research was partly supported by Cooperative Agreement No. 58-3198-2-0006 with the Agricultural Research Service, U.S. Department of Agriculture.

2 The Model

Let Y_{ij}, $(i = 1, \ldots, n)$, $(j = 1, \ldots, r)$ denote the j^{th} observed daily intake for individual i, and let $y_i \equiv E(Y_{ij} \mid i)$ denote the usual intake for individual i. We wish to estimate the marginal distribution function of the random variables y_i. The transformation that maps observed daily intakes Y_{ij} into normal-scale daily intakes X_{ij} is denoted by $g(\cdot)$. Estimation of $g(\cdot)$ is discussed in Nusser et al. (1996). We assume a measurement error model in the normal scale:

$$X_{ij} = x_i + u_{ij}, \tag{1}$$

where $x_i \sim NI(\mu_x, \sigma_x^2)$; $u_{ij} \sim N(0, \sigma_u^2)$; x_i is the normal-scale usual intake for individual i; u_{ij} is the normal-scale measurement error for individual i on day j; the u_{ij} are independent given i; and x_i and u_{jk} are independent for all i, j, and k. Finally, a transformation $h(\cdot)$ that carries normal-scale usual intakes x_i to original-scale usual intakes y_i is obtained. We then estimate the p^{th} percentile, θ_p, of the usual intake distribution by

$$\hat{\theta}_p = h(\hat{\mu}_x + \hat{\sigma}_x \Phi^{-1}(p)), \tag{2}$$

where $\hat{\mu}_x$ and $\hat{\sigma}_x$ are estimates of the measurement error model parameters, and Φ^{-1} denotes the inverse of the standard normal distribution function.

3 Taylor Linearization for SRS

The C-SIDE software mentioned in Section 1 estimates the standard error of $\hat{\theta}_p$ in (2) using Taylor linearization to approximate the variance of $\hat{\mu}_x$ and $\hat{\sigma}_x$, and to approximate the variance of $h(\hat{Q}_p)$, where

$$\hat{Q}_p = \hat{\mu}_x + \hat{\sigma}_x \Phi^{-1}(p).$$

The effect of survey design on the variance component estimates and the function $h(\cdot)$ cannot easily be formulated mathematically, due to the nature of $g(\cdot)$ and $h(\cdot)$, so this particular linearization assumes that the data Y_{ij} (and hence X_{ij}) are a simple random sample of daily intakes.

Associated with the model in (1) is the ANOVA presented in Table 1, from which we obtain estimators of μ_x, σ_x^2, and σ_u^2 :

Source	df	SS	E(MS)
Model	$n-1$	$r\sum_{i=1}^{n}\left(\bar{X}_{i.}-\bar{X}_{..}\right)^2$	$\sigma_u^2+r\sigma_x^2$
Error	$n(r-1)$	$\sum_{i=1}^{n}\sum_{j=1}^{r}\left(X_{ij}-\bar{X}_{i.}\right)^2$	σ_u^2
Total	$nr-1$	$\sum_{i=1}^{n}\sum_{j=1}^{r}\left(X_{ij}-\bar{X}_{i.}\right)^2$	

Table 1: Analysis of variance under SRS.

$$\hat{\mu_x} = \bar{X}_{..} = \frac{1}{nr}\sum_{i=1}^{n}\sum_{j=1}^{r} X_{ij}, \tag{3}$$

$$\hat{\sigma}_u^2 = \frac{1}{n(r-1)}\sum_{i=1}^{n}\sum_{j=1}^{r}\left(X_{ij}-\bar{X}_{i.}\right)^2, \tag{4}$$

$$\hat{\sigma}_x^2 = \frac{1}{r}\left[\frac{r}{n-1}\sum_{i=1}^{n}\left(\bar{X}_{i.}-\bar{X}_{..}\right)^2 - \hat{\sigma}_u^2\right], \tag{5}$$

where $\bar{X}_{i.} = \frac{1}{r}\sum_{j=1}^{r} X_{ij}$. Under the model (1)

$$\hat{V}(\hat{\mu}_x) = \frac{1}{n}\hat{\sigma}_x^2 + \frac{1}{nr}\hat{\sigma}_u^2 \tag{6}$$

is an unbiased estimator of $V(\hat{\mu}_x)$. Because

$$\frac{r}{\sigma_u^2 + r\sigma_x^2}\sum_{i=1}^{n}\left(\bar{X}_{i.}-\bar{X}_{..}\right)^2 \sim \chi^2_{n-1}$$

independently of

$$\frac{1}{\sigma_u^2}\sum_{i=1}^{n}\sum_{j=1}^{r}\left(X_{ij}-\bar{X}_{i.}\right)^2 \sim \chi^2_{n(r-1)},$$

we have

$$V(\hat{\sigma}_x^2) = \frac{1}{r^2}\left[\frac{2\left(\sigma_u^2+r\sigma_x^2\right)^2}{n-1} + \frac{2\sigma_u^4}{n(r-1)}\right],$$

which suggests the estimator

$$\hat{V}(\hat{\sigma}_x^2) = \frac{1}{r^2}\left[\frac{2\left(\hat{\sigma}_u^2+r\hat{\sigma}_x^2\right)^2}{n-1} + \frac{2\hat{\sigma}_u^4}{n(r-1)}\right].$$

To estimate $V(\hat{\sigma}_x) = V(\sqrt{\hat{\sigma}_x^2})$ we use the Taylor expansion of the square root function at the point σ_x^2 to obtain

$$\sqrt{\hat{\sigma}_x^2} \doteq \sqrt{\sigma_x^2} + \frac{1}{2\sqrt{\hat{\sigma}_x^2}}\left(\hat{\sigma}_x^2 - \sigma_x^2\right).$$

It follows that

$$V(\hat{\sigma}_x) \doteq \frac{1}{4\sigma_x^2} V(\hat{\sigma}_x^2),$$

so we take

$$\hat{V}(\hat{\sigma}_x) = \frac{1}{4\hat{\sigma}_x^2} \hat{V}(\hat{\sigma}_x^2). \quad (7)$$

Define

$$\begin{aligned} Q_p &= \mu_x + \sigma_x \Phi^{-1}(p), \\ \hat{Q}_p &= \hat{\mu}_x + \hat{\sigma}_x \Phi^{-1}(p). \end{aligned}$$

Then $\hat{\theta}_p$ in (2) is simply $h(\hat{Q}_p)$. To estimate the variance of $\hat{\theta}_p$, we use the Taylor expansion of $h(\hat{Q}_p)$ at Q_p to obtain

$$h(\hat{Q}_p) \doteq h(Q_p) + \frac{\partial h(x)}{\partial x}(\hat{Q}_p - Q_p).$$

It follows that

$$V(\hat{\theta}_p) \doteq \left[\frac{\partial h(x)}{\partial x}\right]^2 V(\hat{Q}_p). \quad (8)$$

An estimator of $V(\hat{Q}_p)$ is given by

$$\hat{V}(\hat{Q}_p) = \hat{V}(\hat{\mu}_x) + [\Phi^{-1}(p)]^2 \hat{V}(\hat{\sigma}_x), \quad (9)$$

where $\hat{V}(\hat{\mu}_x)$ and $\hat{V}(\hat{\sigma}_x)$ are as in (6) and (7), respectively. Combining (8) and (9), we obtain

$$\hat{V}(\hat{\theta}_p) = \left[\frac{\partial h(x)}{\partial x}\right]^2 \left\{\hat{V}(\hat{\mu}_x) + [\Phi^{-1}(p)]^2 \hat{V}(\hat{\sigma}_x)\right\}. \quad (10)$$

4 Taylor and BRR in SRS

The estimator $\hat{\theta}_p$ of (2) is an estimator of a population quantile, but by construction is expected to be a smoother estimator than the usual sample quantile. To investigate the smoothness properties of $\hat{\theta}_p$ and to assess the general performance of BRR, we first consider the case where the observed daily intakes come from a simple random sample. In this case, the Taylor linearization method derived in Section 3 is expected to yield acceptable standard error estimates. In the simulation, the usual intake distribution is constructed to be moderately skewed and to require a fair amount of effort to estimate with the C-SIDE software. For each of 1,000 samples, an observation Y_{ij} for the j^{th} day ($j = 1, 2$) on the i^{th} individual ($i = 1, \ldots, 700$) is generated as follows:

- Draw x_i, the individual's usual intake in normal scale from a $N(0, 0.36)$ distribution.

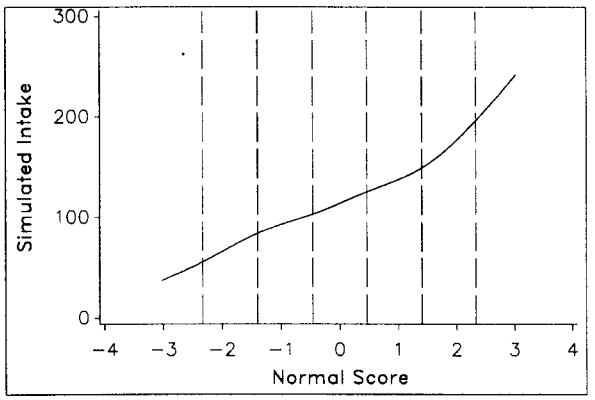

Figure 1: Transformation of X_{ij} to Y_{ij}.

- Draw σ_{ui}^2, the measurement error variance, from a discrete uniform distribution on the values 0.32, 0.50, 0.64, 1.1. The measurement error variance distribution has mean 0.64 and variance 0.0834.

- Draw the measurement error u_{ij} from a normal distribution with mean zero and variance σ_{ui}^2, and form $X_{ij} = x_i + u_{ij}$, where X_{ij} is the observed intake in normal scale. If X_{ij} is less than -6.97, X_{ij} is set to -6.97.

By construction, the marginal distribution of X_{ij} has mean zero and unit variance. Let $Y_{ij} = L_{ij}^{2.5}$, where L_{ij} is a grafted cubic function of X_{ij}. The function relating Y_{ij} to X_{ij} is shown in Figure 1. The definition of L_{ij} ensures that no power transformation can be applied to the Y_{ij} to achieve normally distributed daily intakes.

For each sample, each individual's observed intakes are randomly assigned to one of 32 approximately equal-sized clusters, which are then grouped into 16 strata. Let $Y_{ij}^{(kl)}$ denote the observation for the i^{th} individual on the j^{th} day, where the i^{th} individual has been assigned to cluster k of stratum l. The columns h_n, $(n = 1, \ldots, 16)$, of a Hadamard matrix of order 16 are used to construct sixteen balanced half-samples as follows: $Y_{ij}^{(kl)}$ is assigned to the n^{th} half-sample if $k = 1$ and the l^{th} element of h_n is 1, or $k = 2$ and the l^{th} element of h_n is -1. This procedure is equivalent to assigning weights 0 and 2 to observations according to the standard BRR procedure presented in McCarthy (1969).

The usual intake distributions are estimated using C-SIDE for each of the sixteen half-samples and for the full sample. Let θ denote the p^{th} percentile of the usual intake distribution. Let $\hat{\theta}_i$, $(i = 1, \ldots, 16)$ and $\hat{\theta}_0$ denote the estimates of θ obtained from the i^{th} half-sample and the full sample, respectively. We

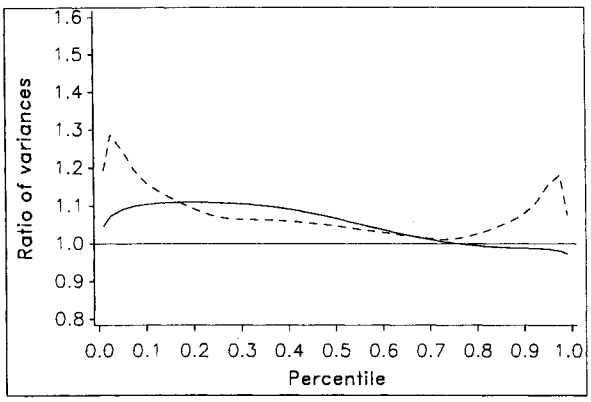

Figure 2: Ratio of Mean Estimated Variance to True Variance in the SRS Simulation.

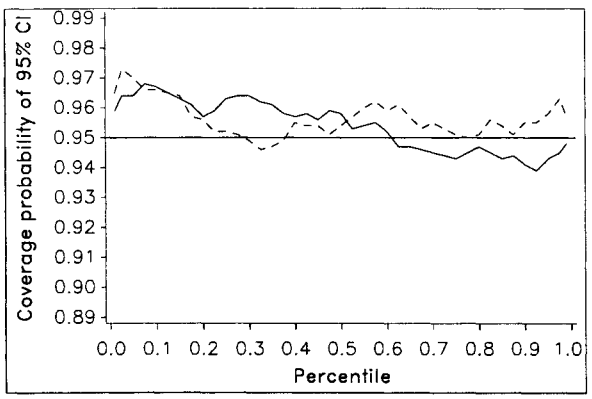

Figure 3: Coverage Probability of Nominal 95% CI in the SRS Simulation.

take $\hat{\theta}_0$ as the point estimate of θ, and compute the BRR estimate

$$\hat{V}_B(\hat{\theta}) = (16)^{-1} \sum_{i=1}^{16} (\hat{\theta}_i - \hat{\theta}_0)^2 \qquad (11)$$

Figures 2 and 3 show the results of the first simulation. The true variance is taken to be the sample variance of the 1,000 estimated percentiles for each value of p. The ratio of the average estimated variance to the true variance is shown in Figure 2. The observed coverage probabilities for nominal 95% confidence intervals constructed using the estimated standard errors are shown in Figure 3. In both Figures, the dotted lines correspond to the Taylor estimates and the solid lines to the BRR estimates. Note that we use a critical point of 1.96 for constructing confidence intervals with the Taylor standard errors, but a critical point of $t_{.975,16} = 2.12$ for the confidence intervals constructed with the BRR standard errors. The BRR variances average somewhat higher than the Monte Carlo variances σ_M^2 over the the first three-fourths of the distribution, and somewhat lower over the last fourth. The general performance of the BRR estimates is quite good, and in fact, is better than the Taylor estimates for percentiles in the tails. The Taylor estimates are also satisfactory, indicating that estimated usual intake quantiles may be sufficiently smooth for the jackknife (a replication approximation to Taylor linearization) to give acceptable results as well.

5 Application to the CSFII

For the second set of simulations, we generate the observed daily intakes Y_{ij} by resampling from a data set that is a subset of the 1994 Continuing Survey of Food Intakes by Individuals (CSFII) conducted by the U.S. Department of Agriculture. The CSFII was a multi-stage stratified area probability sample from the 48 coterminous states. The sample was divided into 43 variance estimation strata, each with two clusters, and 43 sets of jacknife weights were constructed to facilitate variance estimation. The i^{th} set of jacknife weights deletes a randomly selected cluster from stratum i and doubles the weights for the observations in the remaining cluster. The subset of the data contains dietary intake data for 1082 men between 20 and 59 years of age. Two daily intake observations were recorded for three dietary components for each individual. The nutrients iron, protein, and vitamin C were analyzed in Nusser et al. (1996). These three nutrients display a wide range of distributional behaviors, so that if BRR and the jackknife perform satisfactorily for these nutrients, we will have some evidence that they will perform well for a variety of dietary components consumed on a near-daily basis. The data set also contains information about nuisance effects such as day-of-week and interview sequence.

The $1082 \times 2 = 2164$ observations in the base data set are used to create 1000 simulated data sets with 43 two-cluster strata as follows.

- For the l^{th} stratum ($l = 1, \ldots, 16$), select numbers $k_j^{(l)}$, ($j = 1, 2$) with replacement from the set $\{1, 2\}$.
- Select multipliers $m_j^{(l)}$ with replacement from the set $\{.90, .91, \ldots, 1.09, 1.10\}$. If $k_1^{(l)} = k_2^{(l)}$ and $m_1^{(l)} = m_2^{(l)}$, then reselect $m_2^{(l)}$ until the $m_j^{(l)}$ are distinct.
- Multiply every daily intake observation in the $k_j^{(l)\text{th}}$ cluster of the l^{th} stratum by $m_j^{(l)}$ to form

two simulated pseudo-clusters for stratum l. The weight for each simulated observation is the weight for the generating observation in the base data set. The values of the variables for day-of-week and interview sequence are similarly retained for the simulated data. Note that, under this procedure, the sizes of the simulated clusters are allowed to vary, in contrast to the procedure used in Section 4.

For each of the 1000 simulated data sets, 44 balanced half-samples are constructed as described in Section 4, using a Hadamard matrix of order 44. Both BRR and Jackknife variance estimates are computed for selected percentiles for each of the six dietary components. BRR variance estimates are obtained using the formula

$$\hat{V}_B(\hat{\theta}) = (44)^{-1} \sum_{i=1}^{44} (\hat{\theta}_i - \hat{\theta}_0)^2 \qquad (12)$$

where $\hat{\theta}_i$ is computed using the i^{th} half-sample. Jacknife variance estimates are obtained using the formula

$$\hat{V}_J(\hat{\theta}) = \sum_{i=1}^{43} (\hat{\theta}_i - \hat{\theta}_0)^2 \qquad (13)$$

where $\hat{\theta}_i$ is computed using the i^{th} set of jackknife weights.

Adjustments for day-of-week and interview sequence effects are made separately for each parent sample and replicate sample. The results of the second simulation for the nutrients protein, iron, and vitamin A are given in Figures 4-9.

The ratios of the average estimated variance to the true variance are shown in Figures 4-6 for the three dietary components. The observed coverage probabilities for nominal 95% confidence intervals constructed using the estimated standard errors are shown in Figures 7-9. In all of these Figures, the dotted lines correspond to the jackknife estimates and the solid lines to the BRR estimates. Note that we use a critical point of $t_{.975,43} = 2.01$ for the confidence intervals constructed with both kinds of standard errors, even though 44 replicates were used to construct the BRR standard errors. In the case of a linear statistic, the BRR procedure reproduces the textbook estimator of variance, which would have 43 degrees of freedom. Hence, we assume the same holds true for our nonlinear statistic, the usual intake quantile.

In general, both the BRR and jackknife variance estimates are conservative, but the BRR estimates are less biased. However, confidence intervals constructed with the estimated variances are rather liberal, suggesting that the distributions of usual intake quantiles are heavier-tailed than the t-distribution. The larger positive bias of the jackknife variance estimator therefore yields more "accurate" confidence intervals.

6 Summary

Our simulation studies suggest that both the balanced repeated replication method and the delete-1 jackknife method yield generally acceptable variance estimates. Both methods yield conservative variance estimates, but somewhat liberal confidence intervals. The performance of the estimates varies quite a bit on a per-nutrient basis. The BRR method is less biased than the jackknife, and is expected to be more stable than the jackknife method when the measurement error variance is small, in which case the usual intake distribution is almost identical to the daily intake distribution.

References

McCarthy, P. J. (1969), "Pseudo-replication: half-samples," *Rev. Intl. Statist. Inst.*, **37**, 239-264.

Nusser, S. M., Carriquiry, A. L., Dodd, K. W., and Fuller, W. A. (1996), "A semiparametric transformation approach to estimating usual daily intake distributions," *J. Amer. Statist. Assoc.* to appear.

Shao, J. and Wu, C. F. J., (1989) "A general theory for jackknife variance estimation," *Ann. Statist.*, **17**, 1176-1197.

Shao, J. and Wu, C. F. J., (1992) "Asymptotic properties of the balanced repeated replication method for sample quantiles," *Ann. Statist.*, **20**, 1571-1593.

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

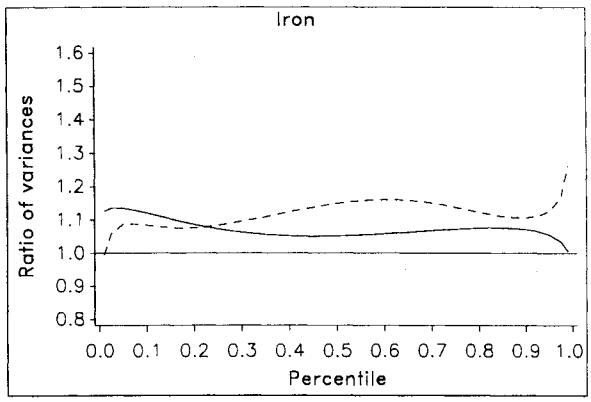

Figure 4: Ratio of Mean Estimated Variance to True Variance for Iron in the CSFII Simulation.

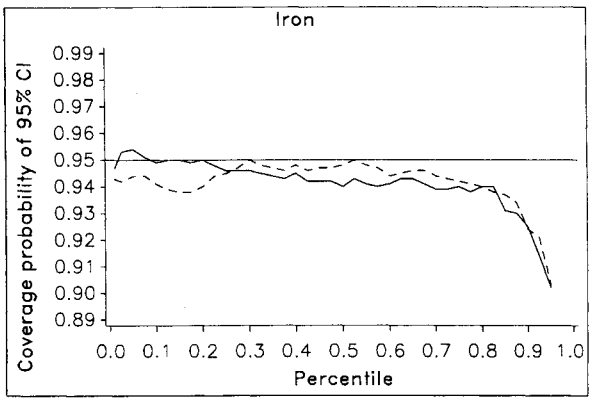

Figure 7: Coverage Probability of Nominal 95% CI for Iron in the CSFII Simulation.

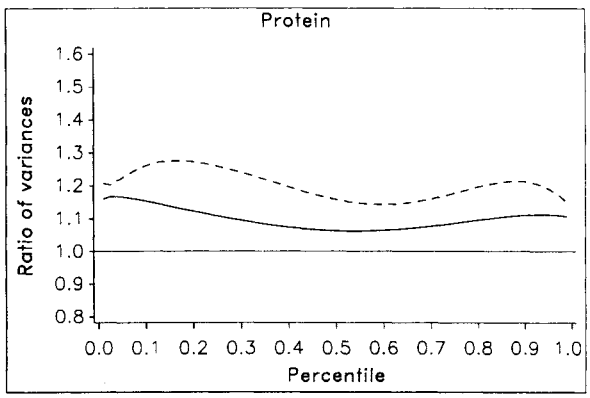

Figure 5: Ratio of Mean Estimated Variance to True Variance for Protein in the CSFII Simulation.

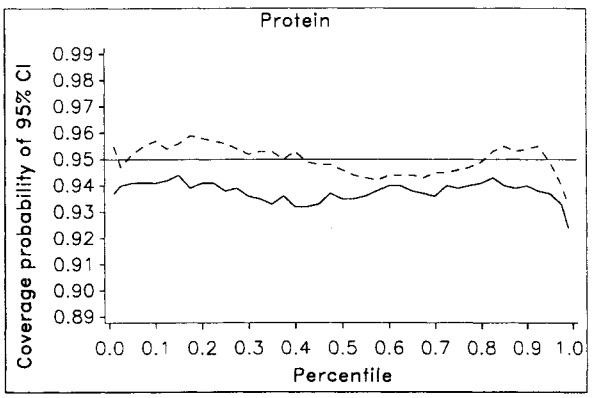

Figure 8: Coverage Probability of Nominal 95% CI for Protein in the CSFII Simulation.

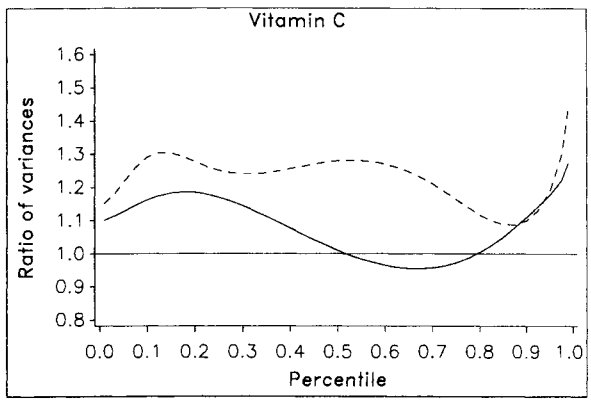

Figure 6: Ratio of Mean Estimated Variance to True Variance for Vitamin C in the CSFII Simulation.

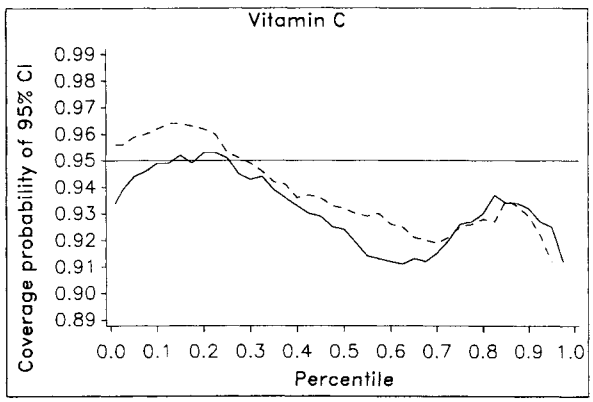

Figure 9: Coverage Probability of Nominal 95% CI for Vitamin C in the CSFII Simulation.