A Semiparametric Transformation Approach to Estimating Usual Daily Intake Distributions

Dietary Assessment Research Series Report 2

S.M. Nusser, A.L. Carriquiry, K.W. Dodd, and W.A. Fuller

Staff Report 95-SR 74
December 1995 (Revised)

A Semiparametric Transformation Approach to Estimating Usual Daily Intake Distributions

Dietary Assessment Research Series Report 2

S.M. Nusser, A.L. Carriquiry, K.W. Dodd, and W.A. Fuller

Staff Report 95-SR 74 (Revised)
December 1995

Center for Agricultural and Rural Development
Iowa State University
Ames, Iowa 50011

S.M. Nusser is assistant professor of statistics; A.L. Carriquiry is assistant professor of statistics; K.W. Dodd is a graduate research assistant, Department of Statistics; and W.A. Fuller is Distinguished Professor of Statistics, Iowa State University.

This research was partly supported by Research Support Agreement No. 58-3198-9-032 with the Human Nutrition Information Service and Cooperative Agreement No. 58-3198-2-006 with the Agricultural Research Service, U.S. Department of Agriculture.

CONTENTS

Introduction 1
Application to CSFII Data 3
The CSFII Data 3
Initial Adjustments 4
Transformation Construction 7
Parameter Estimation in Transformed Space 11
The Distribution of Usual Intakes 13
Monte Carlo Study 17
Summary 23
References 24

FIGURES

Figure 1. Plot of grafted polynomial for iron 10

Figure 2. Estimated densities of usual intakes, four-day means, and one-day intakes for vitamin C 15

Figure 3. Plot of transformation normal scale to observed intakes used in simulation 20

Figure 4. (a) Average estimated root mean squared error in estimated percentiles and (b) absolute average estimated bias from 100 simulation runs, for the proposed (ISU) and best power (BP) estimation methods 22

TABLES

Table 1. Sample moments for adjusted intakes 6

Table 2. Statistics for the transformation to normality 8

Table 3. Estimated moments for transformed data 12

Table 4. Estimated moments for usual intakes in the original scale 16

Table 5. Sample moments for four-day means in the original scale 17

Table 6. Estimated percentiles for usual intake distributions 18

Table 7. Estimates of selected percentiles of the usual intake distribution using three estimation methods, averaged over 1000 simulations 21

A SEMIPARAMETRIC TRANSFORMATION APPROACH TO ESTIMATING USUAL DAILY INTAKE DISTRIBUTIONS

Introduction

The United States Department of Agriculture has been responsible for conducting periodic surveys to estimate food consumption patterns of households and individuals in the United States since 1936. Because dietary intake data from these surveys are used to formulate food assistance programs, consumer education efforts and food regulatory activities, it is crucial that appropriate methodologies be used in the analysis of these data. An important concept in analyzing food consumption data is that of usual intake, defined as the long-run average of daily intakes of a dietary component by an individual. This article outlines a methodology to estimate usual intake distributions of dietary components that are consumed on a nearly daily basis (e.g. nutrients, cholesterol, energy) from 24-hour dietary intake data.

To assess usual intake, daily dietary intakes are often collected from individuals for a number of days. If the individual's mean daily intake for a particular dietary component is used as an indication of the individual's usual intake, the variance of the mean intakes contains some within-individual variability and, hence, is greater than the variance of the usual intakes. Other parameters of the distribution of mean intakes may also differ from the parameters of the distribution of usual intakes. Because of these problems, using the distribution of the mean of a few days as an estimate of the usual intake distribution can lead to erroneous inferences regarding dietary status.

The problem of estimating the distribution of usual intakes can be formulated as the problem of estimating the distribution of a random variable that is observed subject to measurement error. There is considerable statistical literature on the general problem

of estimating the density of a random variable that has been contaminated by additive measurement errors. See Mendelsohn and Rice (1982), Stefanski (1990), Stefanski and Carroll (1990, 1991), and references cited by those authors.

Several approaches have been considered for the estimation of the distribution of usual intakes. Nusser et al. (1990) suggested a measurement error model for the observed intakes. The model decomposes the observed daily intake of an individual into the usual daily intake for that individual plus a measurement error associated with the individual on the day the intake was observed. To account for the heterogeneity of within-individual moments often observed in dietary intake data, the second and third moment of an individual's measurement errors are modeled as a function of the individual's usual intake. The first three moments of usual intake are estimated under the model. A parametric form for the usual intake distribution is assumed, and moment methods are used to estimate the parameters of the assumed distribution. While this approach has the advantage of working with the data in the original scale, it requires several parametric assumptions and is difficult to implement for dietary components that exhibit extreme behaviors (e.g. vitamin A).

A second approach to estimating the usual intake distribution involves transforming the daily intakes so that the transformed values approximately follow a normal distribution. The National Research Council (1986) recommended this approach and suggested a log transformation. As we explain, log transformations or simple power transformations do not consistently produce transformed data that are normally distributed.

We develop a semiparametric transformation that transforms dietary intake data into approximately normally distributed data. The transformation is a grafted cubic equation fit to a power of the original data. This fitting can be considered a semiparametric version of the Lin and Vonesh (1989) procedure. It is also related to the spline approach for estimating the distribution function. See Wahba (1975) and Wegman (1982). The transformed observed intake data are assumed to follow a measurement error model, and normal theory is used to estimate the parameters of the model. An estimated inverse

transformation carries the normal usual intake distribution back to the original scale and defines the distribution of usual intakes. Inferences concerning usual intakes can be made in the transformed space or in the original space. The approach was developed with the objective of producing an algorithm suitable for computer implementation and applicable to a wide variety of dietary components that are consumed on a nearly daily basis. Software is available from the authors to estimate usual intake distributions using this method.

Application to CSFII Data

The CSFII Data

The data for this study are a subset of the data from the 1985 Continuing Survey of Food Intakes by Individuals (CSFII) conducted by the U.S. Department of Agriculture (1985). Daily dietary intakes were collected from women between 19 and 50 years of age and from the preschool children of the women at approximate two-month intervals over the period, April 1985 to March 1986. Twenty-four-hour recall data were collected by personal interview for the first day and by telephone whenever possible for subsequent days. The sample was a multi-stage stratified area probability sample from the 48 coterminous states, and was designed to be self-weighting. Because of the relatively high attrition rate for the six-day sample, the USDA constructed a four-day data set for analyses which consisted of the first day of dietary intakes for all individuals who provided at least four days of data, plus a random selection of three daily intakes from the remaining three, four or five days of available data. Weights were developed to adjust for nonresponse and the analyses of this paper are constructed on the weighted data.

We analyze a subset of the four-day data set containing dietary intakes for 737 women between 25 and 50 years of age who were responsible for meal planning within the household and who were not pregnant or lactating during the survey period. Because of the time separation of the observations, we assume the four observations on each individual to be independent observations on that individual. The dietary components, calcium,

energy, iron, protein, vitamin A and vitamin C, were selected for analysis because of their nutritional importance and because of their varying distributional behaviors.

The report of the National Research Council (1986) provides a review of factors that influence observed daily intakes. Some effects, such as errors in reported food intake and translation of food intake to nutrient intake, are not estimable from the data of our study. The effect of other factors, such as day of the week, season (month), interview method, and interview sequence can be investigated.

Initial Adjustments

We begin by adjusting the data for several nuisance effects. The adjustment variables will vary with each study. In the case of the 1985 CSFII data, the daily intakes were examined using least squares methods to determine whether day of the week, month, interview mode (telephone or in-person) and interview sequence (first, second, third, or fourth interview) effects were important. Month and interview sequence are confounded to a large degree because the first interview was conducted at nearly the same point in time for all individuals.

Let W_{0ij} denote the observed intake for the ith individual on the jth day in the interview sequence plus a constant equal to 0.0001 times the sample mean for the nutrient. This small amount is added to avoid problems in subsequent procedures that depend on the derivative of a power of the data, which can be infinite when evaluated at zero. Consider the sample of n individuals, and let the ith individual have a weight w_i , where $\sum_{i=1}^{n} w_i = 1$.

Because dietary intake data are often skewed, a power transformation is applied to the data to make the distributions of the observed data more symmetric. The original observations W_{0ij} are used to estimate the power γ by minimizing the error sum of squares

$$\sum_{i=1}^{n} w_i \sum_{j=1}^{k} (U_{ij} - \beta_0 - \beta_1 W_{0ij}^{\gamma})^2 \ , \tag{1}$$

over a grid of values of γ , where U_{ij} is the normal score for the ij^{th} observation, and β_0 and

β_1 are estimated for each value of γ. The normal scores are computed as

$$U_{ij} = \Phi^{-1}[(s_{ij} - 3/8)/(nk + 1/4)] \ , \tag{2}$$

where Φ is the standard normal distribution function, and s_{ij} is the rank of the ij^{th} observation. The grid of values for γ is [1, $(1.5)^{-1}$, $(2.0)^{-1}$, ..., $(10)^{-1}$, *log*], where *log* denotes the natural logarithm and corresponds to $\gamma = 0$.

Once the power has been selected, a model containing day of the week, interview mode, and interview sequence as additive classification variables is fit by weighted least squares to the power transformed observations, W^{γ}_{0ij}. (A different set of variables may be more appropriate when analyzing other data sets.) The weights in the regression are the sampling weights. Interview mode is not significant for any dietary component. Day-of-week effects are significant for energy ($p < 0.001$) and protein ($p < 0.05$) intakes, primarily because of higher consumption on weekends for both dietary components. Sequence effects (confounded with month effects) are significant at the $\alpha = 0.001$ level for calcium, energy, iron, and protein intakes, and are principally attributable to higher intake levels on the first interview day versus the other three days.

Because of these results, data were adjusted for weekday and interview sequence effects. If $Z_{0ij} = W^{\gamma}_{0ij}$ represents the power-transformed observed intake for the ith individual on the jth day, then the ij^{th} observation adjusted for weekday and interview sequence effects is $Z_{ij} = \hat{Z}^{-1}_{0ij}\bar{Z}_{0.1}Z_{0ij}$, where $\bar{Z}_{0.1}$ is the mean of the power-transformed observed intakes for the first interview day and $\hat{Z}_{0ij}$ is the predicted intake from the regression for the ith individual on the jth day. The ratio adjustment, which is the multiplicative analog to the standard additive adjustment, is used to increase the chance that adjusted intake values are nonnegative. The data are adjusted to the mean of the first interview day (rather than the grand mean) because the data are believed to be more accurate on the first interview day than on subsequent days. (Other adjustments, such as to the grand mean, can be alternatively implemented when deemed to be more appropriate for data sets other than

Table 1. Sample moments for adjusted intakes

Dietary Component	Mean	Among-Individual Std. Dev.	Within-Individual Std. Dev.	Ratio of Within-to-Among-Individual Variances	Skewness
Calcium (mg)	622.3	242.4	318.2	1.72	1.35
Energy (kcal)	1683.4	441.0	579.4	1.73	1.10
Iron (100 mg)	1105.3	312.2	488.7	2.45	1.76
Protein (10 g)	668.8	156.0	262.4	2.83	1.38
Vitamin A (μg/RE)	801.0	524.5	1557.9	8.82	11.75
Vitamin C (10 mg)	792.5	413.9	631.2	2.33	1.87

the CSFII.)

It is well established that the characteristics of responses in a repeated survey are a function of the time-in-sample at which a respondent is observed. See, for example, Bailar (1975). Our initial regression adjustment modifies the data so that there is no sequence effect in the mean of the intake distributions for the different days. Because of the possibility of other higher-order time-in-sample effects, we standardize the sample variance of transformed observations for the second, third, and fourth times-in-sample to the sample variance observed on the first day. The adjusted observations in the original scale are defined by $Y^*_{ij} = [\hat{\mu}_{.1} + S_{.j}^{-1} S_{.1}(Z_{ij} - \hat{\mu}_{.j})]^{1/\gamma}$, where $i = 1,\ 2,\ \ldots,\ n$ individuals, $j = 1,\ 2,\ \ldots, k$ days, $S^2_{.j} = (n-1)^{-1} \sum_{i=1}^{n} (Z_{ij} - \hat{\mu}_{.j})^2$, and $\hat{\mu}_{.j} = n^{-1} \sum_{i=1}^{n} Z_{ij}$. For a very few observations (fewer than four for every component), the slope of the transformation is modified near zero to guarantee nonnegative adjusted data.

The among- and within-individual standard deviations for the adjusted intakes in original units are presented in Table 1. These statistics indicate that there is considerable within-individual variability relative to among-individual variability. The ratios of within- to among-individual variances are similar to those for comparable dietary components reported in The National Research Council report (1986). Vitamin A is unusual in that there is one

very large observation and a few other large observations that are responsible for the very large within-individual variance. Table 1 also contains the estimator of skewness, where skewness is the third central moment divided by the cube of the standard deviation. The skewness coefficient indicates that for most dietary components, an assumption of normality is unreasonable. In addition, analyses (not shown here) indicate that within-individual standard deviations are positively correlated with individual means.

Transformation Construction

Our estimation scheme is designed to handle samples with unequal weights. In order to apply classical equal-weight methods for the estimation of the components of variance model, we use the empirical cumulative distribution function to create an equal-weight sample from the adjusted, unequal-weight sample. The first step to creating the equal-weight sample is to construct an empirical cumulative distribution function from the nk observations, defined by

$$\hat{F}_Y(a) = \sum_{i=1}^{n} w_i \sum_{j=1}^{k} I_{Y_{ij}}(a) \ ,$$

where $I_{Y_{ij}}(a)$ is the indicator function with

$$\begin{aligned} I_{Y_{ij}}(a) &= 1 \text{ if } \quad Y_{ij}^* \leq a \\ &= 0 \text{ otherwise} \ . \end{aligned}$$

A continuous function, denoted $\tilde{F}_Y(a)$, is created by connecting the midpoints of the rises in the steps of $\hat{F}_Y(a)$. This function is used to define nk observations of an equal-weight sample that gives nearly the same distribution function as that of the adjusted data. The equal-weight observations are defined by $Y_{ij} = \tilde{F}_Y^{-1}[(nk)^{-1}(s_{ij} - 0.5)]$ for $i = 1, 2, ..., n$ and $j = 1, ..., k$, where s_{ij} is the rank of the Y_{ij}^* . These adjusted, equal-weight intakes Y_{ij} are hereafter called *observed intakes.*

The first step to transforming the observed intakes to normality is to calculate normal scores U_{ij} as defined in (2) for the Y_{ij} . The pairs (U_{ij}, Y_{ij}) are used to estimate a

Table 2. Statistics for the transformation to normality

Dietary Component	Inverse of Power	Anderson-Darling for Fitted Values[a]	Number of Join Points	t for Heterogeneous Variances[b]	t for Linear Effect[b]
Calcium	3.5	0.16	4	3.08	-2.14
Energy	2.0	0.23	5	3.04	0.35
Iron	2.5	0.22	8	2.72	-0.71
Protein	2.0	0.22	8	2.20	0.71
Vitamin A	5.5	0.25	12	3.12	-0.71
Vitamin C	3.5	0.24	6	1.95	-1.22

[a]Reject at 10% level if Anderson-Darling statistic greater than 0.656.
[b]Reject null hypothesis of zero slope at 5% level if $|t| > 1.96$.

semiparametric function that transforms the observed intakes into approximately normal variables. The transformation function is fit to the data in two phases. First, a power is determined that produces observations that are close to normally distributed by minimizing (1), where $w_i \equiv 1$ and Y_{ij} replaces W_{0ij}. Let the selected power be denoted by α . The inverses of the powers of the first transformation step for the CSFII data are given in the first column of Table 2.

In the next phase of the normality transformation, a grafted cubic polynomial is fit to the $\left(U_{ij},\ Y_{ij}^{\alpha}\right)$ pairs, minimizing deviations in the Y-direction. Let the join points for the polynomial be B_1, B_2, ..., B_p . The value B_1 is -2.33 and $B_p = 2.33$. The values B_2, ..., B_{p-1} are defined such that the intervals $(B_i,\ B_{i+1})$, $i = 1, 2, ..., p-1$, are of equal length. The function is constructed to be linear for $U_{ij} < B_1$, linear for $U_{ij} > B_p$, and cubic in the intervals $(B_i,\ B_{i+1})$, $i = 1, 2, ..., p-1$, with continuous first and second derivatives at the join points. See, for example, Fuller (1976, p. 393) for a description of the function. The fitted grafted polynomial function is also constrained to be monotone increasing. At least three join points are included in the model for each component. The number of parameters, p, in the grafted polynomial model is equal to the number of join points.

The number of join points for the grafted cubic is chosen to be the minimum number of join points (up to 12) required to make the value of the Anderson-Darling test statistic for

normality less than 0.25 when applied to the transformed data. The number of join points is set to 12 if the Anderson-Darling test statistic does not fall below 0.25 with twelve or fewer join points. The Anderson-Darling statistic computed for the observed intake data transformed with the grafted polynomial is given in the second column of Table 2. The number of join points is given in column three of Table 2. Vitamin A is the only dietary component that required 12 join points. Figure 1 contains a plot of the normal scores against the $(2.5)^{-1}$ power of the iron observations. The S-shaped plot indicates that simple power transformations are not adequate to transform the plot into a straight line.

As an additional check on the transformation, the Anderson-Darling statistic was computed for the individual means of the transformed intakes. In no case was the statistic significant at the ten percent level.

To check the hypothesis that the within-individual variances calculated from the transformed data are constant over individuals, let

$$A_i = (k-1)^{-1} \sum_{j=1}^{k} \left(X_{ij} - \bar{X}_{i.}\right)^2$$

and

$$\hat{M}_4 = 3\bar{A}^{-2} n^{-1} \sum_{i=1}^{n} \left[1 + 2\,(k-1)^{-1}\right]^{-1} A_i^2, \tag{3}$$

where X_{ij} is the transformed value for individual i on day j, $\bar{X}_{i.} = k^{-1} \sum_{j=1}^{k} X_{ij}$, $\bar{A} = n^{-1} \sum_{i=1}^{n} A_i$, and $k = 4$ is the number of observations per individual. If the transformed observations are normally distributed with homogeneous variances and four observations per person, $\hat{M}_4$ estimates 3, the fourth moment of the standard normal distribution. The approximate variance is

$$\begin{aligned} \hat{V}\left(\hat{M}_4\right) &= 9n^{-1}(k-1)^{-4}\left[1 + 2\,(k-1)^{-1}\right]^{-2} \\ &\quad \times \left\{ \left[2^4 \Gamma\left(\frac{k-1}{2}\right)^{-1} \Gamma\left(\frac{k-1}{2} + 4\right)\right] - \left[2\,(k-1) + (k-1)^2\right]^2 \right\}, \end{aligned}$$

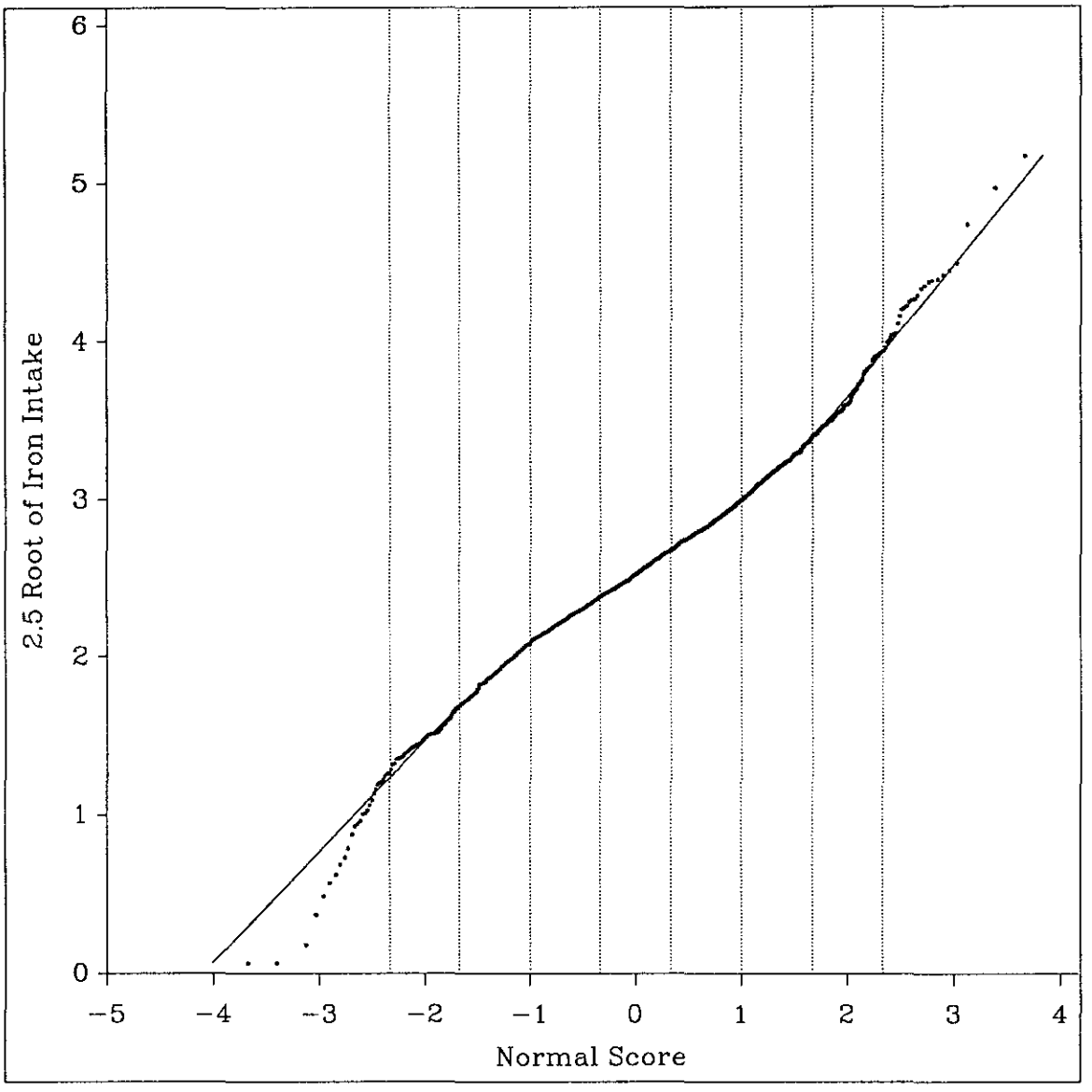

Figure 1. Plot of grafted polynomial for iron.
Notes: The normal scores are represented by points. The smooth line is the fitted grafted polynomial. Vertical dashed lines designate join points of the grafted polynomial.

which is equal to .039077 for $n = 737$ and $k = 4$. The values of the test statistic

$$\left[\hat{V}\left(\hat{M}_4\right)\right]^{-1/2}\left(\hat{M}_4 - 3\right)$$

calculated using the transformed observed data, are given in Table 2 under the heading "t for Heterogeneous Variances." This ratio is greater than 1.96 for all nutrients analyzed except vitamin C, indicating that the within-individual variances vary across individuals.

To investigate the hypothesis that the heterogeneity of within-individual variances in the transformed space is due to a relationship between within-individual standard deviations and individual means, the model $A_i^{1/2} = \beta_0 + \beta_1 \bar{X}_{i.}$ was fit using least squares. The t-statistics for testing the hypothesis that $\beta_1 = 0$ are presented in Table 2 in column 4. The statistic for calcium is -2.14 while the remaining statistics are less than 1.5 in absolute value. When within-individual standard deviations are plotted against individual means, no obvious patterns are revealed. Therefore, it was decided to complete the analysis for all nutrients under the assumption that the variances are not related to the means.

Parameter Estimation in Transformed Space

A measurement error model is used for estimating the distribution of usual intakes in normal space. Let

$$X_{ij} = x_i + u_{ij} \ , \tag{4}$$

where $x_i \sim NI(\mu_x, \sigma_x^2)$, $u_{ij} \sim N(0, \sigma_{ui}^2)$, $\sigma_{ui}^2 \sim (\mu_A, \sigma_A^2)$, x_i is the unobservable normal usual intake value for individual i ; u_{ij} is the unobservable measurement error for individual i on day j ; the u_{ij} are independent given i ; and x_i and $u_{\ell j}$ are independent for all i, ℓ and j .

On the basis of the empirical analyses presented in Table 2, we permit heterogeneous within-individual variances. The errors u_{ij} represent variation of two kinds. There is the day-to-day variability in the true amounts eaten by individual i , and there is the difference

Table 3. Estimated moments for transformed data

Dietary Component	Among-Individual Variance $(\hat{\sigma}_x^2)$	Average Within-Individual Variance $(\hat{\sigma}_{u.}^2)$	Within- to Among-Ratio $\frac{\hat{\sigma}_{u.}^2}{\hat{\sigma}_x^2}$	Variance of Individual Variances $(\hat{\sigma}_A^2)$
Calcium	0.367	0.637	1.74	0.082
Energy	0.382	0.637	1.67	0.081
Iron	0.320	0.692	2.16	0.086
Protein	0.276	0.734	2.66	0.078
Vitamin A	0.261	0.739	2.83	0.112
Vitamin C	0.319	0.680	2.13	0.059

between the true amount eaten and the amount reported for an individual. It is believed that the day-to-day variance for an individual is much larger than the variance of the reporting error. The transformed observed daily intakes X_{ij} have $\mu_X \doteq 0$ and $\sigma_X^2 \doteq 1$. The conditional distribution of X_{ij}, given $(x_i,\ \sigma_{ui}^2)$ is $N\left(x_i,\ \sigma_{ui}^2\right)$. However, the unconditional distribution is not normal if $\sigma_A^2 > 0$. We conduct our analysis under the operational assumption that the initial transformation produces x_i and u_{ij} satisfying (4). Under (4), the individual means, $\bar{X}_{i.} = k^{-1}\sum_{j=1}^{k} X_{ij}$, are independent $\left(0, \sigma_{\bar{X}}^2\right)$ random variables, where $\sigma_{\bar{X}}^2 = \sigma_x^2 + k^{-1}\mu_A$. For our purposes, it is not necessary to specify a form of the distribution of the individual error variances because we will only use the variance of the distribution of variances.

Estimators for the moments are $\hat{\mu}_x = n^{-1}\sum_{i=1}^{n}\bar{X}_{i.}$, $\hat{\sigma}_{\bar{X}}^2 = (n-1)^{-1}\sum_{i=1}^{n}(\bar{X}_{i.} - \hat{\mu}_x)^2$, $\hat{\sigma}_{u.}^2 = [n(k-1)]^{-1}\sum_{i=1}^{n}\sum_{j=1}^{k}(X_{ij} - \bar{X}_{i.})^2$, $\hat{\sigma}_x^2 = \hat{\sigma}_{\bar{X}}^2 - k^{-1}\hat{\sigma}_{u.}^2$, $\hat{\sigma}_A^2 =$ $n^{-1}\left(1 + 2\left[k-1\right]^{-1}\right)\sum\limits_{i=1}^{n} A_i^2 - \bar{A}^2$, where A_i and $\bar{A}$ are defined in (3). The within- and among-individual variances for the transformed data are given in Table 3. In all cases, the sum of the within-individual and among-individual variances is close to one because the transformed data have mean zero and variance one. The average of the within-individual variances exceeds the among-individual variance for all dietary components. The ratio of within- to among-individual variance is smallest for energy with a value of 1.67 and

is largest for vitamin A with a ratio of 2.83. The ratios of within- to among-individual variance of Table 3 are similar but larger than the corresponding ratios computed from the standard deviations in original scale of Table 1 with the exception of vitamin A. In the original scale, the vitamin A data are skewed, the individual standard deviations are positively correlated with the individual means, and a few very large observations made a large contribution to the within-individual variance in the original scale.

The last column of Table 3 contains an estimate of the variance of the individual variances, denoted by $\hat{\sigma}_A^2$. The coefficients of variation of the individual variances are about 35 to 45%.

The Distribution of Usual Intakes

Under model (4), the conditional distribution of observed daily intakes in normal scale for all individuals with usual intake $\ddot{x}_i$ is the average of all normal distributions with common mean $\ddot{x}_i$ and variance A , where $A \sim (\mu_A, \sigma_A^2)$. Thus, observed intake in normal scale is the sum of $\ddot{x}_i$ and u , where $E\{(u, u^2, u^4)|x = \ddot{x}_i\} = (0, \mu_A, 3\mu_A^2 + 3\sigma_A^2)$, and the distribution of u is symmetric about zero.

Let $\ddot{y}_i$ denote the usual intake in original scale for all individuals with usual normal intake $\ddot{x}_i$, and let g denote the transformation taking the adjusted observed intakes Y to normality. Then, $\ddot{y}_i = E\{Y|x = \ddot{x}_i\} = E\{g^{-1}(x + u)|x = \ddot{x}\} = h(\ddot{x})$. The transformation h is estimated by approximating the conditional expectation of Y at a set of values of $\ddot{x}$, and then fitting a grafted polynomial to the $(\ddot{y}, \ddot{x})$ pairs. We specified 400 values of $\ddot{x}$ such that the first five moments of the points match the first five moments of a $N(0, \hat{\sigma}_x^2)$ distribution. At each value of $\ddot{x}$, we use a nine-point approximation to the distribution of u . The distribution of u has mean zero and a variance with estimated mean $\hat{\sigma}_{u.}^2$ and estimated variance of variance equal to $\hat{\sigma}_A^2$. Nine points, c_ℓ , and nine weights, w_ℓ , where $\sum w_\ell = 1$, are constructed such that the first five moments of the discrete nine-point distribution match the first five estimated moments of the conditional distribution of $\ddot{x} + u$

conditional on $\ddot{x}$. For each of the 400 values of $\ddot{x}$, the usual intake in the original scale is approximated by $\ddot{y}_i = \sum_{\ell=-4}^{4} w_\ell g^{-1}(\ddot{x}_i + c_\ell)$, where $\ddot{x}_i$ is the ith value in normal scale, and c_ℓ and w_ℓ $(\ell = -4, -3, ..., 4)$ are the values and weights, respectively, for the nine-point approximation to the distribution of u . The 400 $\ddot{y}$-values provide a 400-point estimator of the usual intake distribution. A grafted cubic created from the pairs $(\ddot{y}_i, \ddot{x}_i)$, denoted by $\hat{h}$, is an estimator of the transformation of the normal $\ddot{x}$ into the usual intakes.

Densities for the dietary components were constructed by multiplying the derivative of $\hat{h}^{-1}(y)$ by the normal ordinate for the usual intake density of the component in the normal scale. The estimated density of usual intakes for vitamin C is the solid line in Figure 2. Also in the figure is the estimated density for observed intake identified by the long dashed line and the estimated density of the four-day mean identified by the short dashed line. The estimated density for four-day means was approximated by using the same smoothing algorithm that was used to estimate the distribution function. The degree of skewness in the density of the mean declines as the number of daily intakes in the mean increases.

Table 4 contains the mean, variance and skewness coefficient for the estimated usual intake distributions calculated from the 400-point approximation. The estimated means of the usual intakes are very close to the means of Table 1. Also, the estimated standard deviations of usual intakes are close to the among-individual standard deviations of Table 1 for calcium, energy, iron and protein. This is to be expected because the estimates of Table 1 are the sample moment estimators of the same quantities. The estimated standard deviations of usual intakes for the two vitamins differ considerably from the direct moment estimators of Table 1. As previously mentioned, the original distributions for the vitamins are very skewed and, hence, the original sample moments are heavily dependent upon a few large observations. The effect on the large observations is reduced for estimators constructed using our procedure.

Comparison of the sample moments for usual intakes in Table 4 with the estimated moments for individual means (Table 5) indicates that the distribution of four-day means

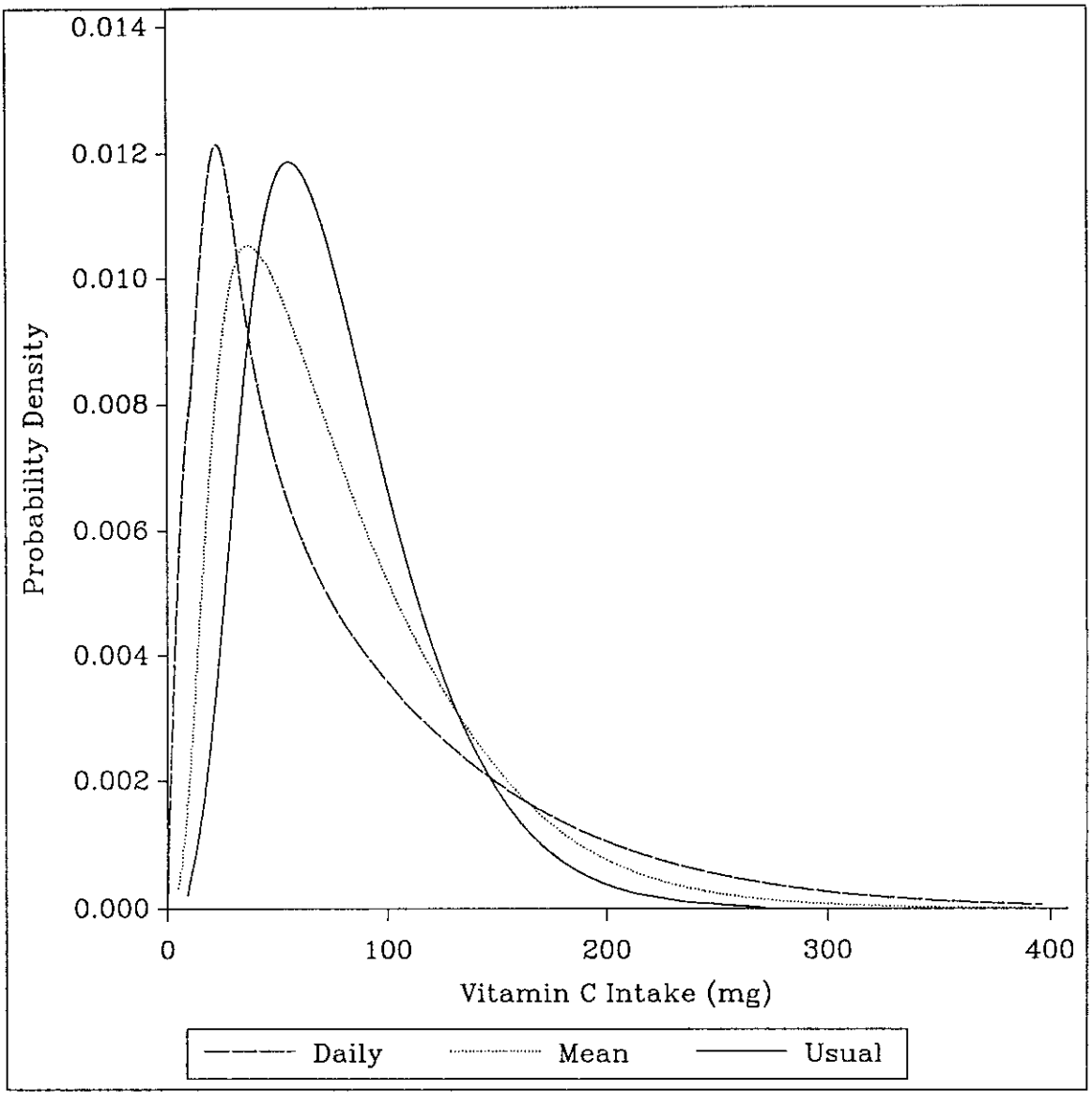

Figure 2. Estimated densities of usual intakes, four-day means, and one-day intakes for vitamin C

Table 4. Estimated moments for usual intakes in the original scale

Dietary Component	Mean	Standard Deviation	Skewness
Calcium (mg)	622.8	240.2	0.82
Energy (kcal)	1684.2	441.1	0.49
Iron (100 mg)	1107.7	305.2	0.72
Protein (10 g)	669.7	159.2	0.47
Vitamin A (μg/RE)	806.6	499.9	2.47
Vitamin C (10 mg)	789.7	395.8	1.09

is a poor estimate of the usual intake distribution. For all dietary components, the standard deviation and skewness coefficient are larger for the mean distribution than for the estimated usual intake distribution.

Table 6 contains estimated percentiles for the dietary components. The percentiles were computed with the estimated transformation function using the percentiles of the estimated distribution of usual intakes in normal scale. For example, the estimated mean and variance of vitamin C usual intakes in normal space are zero and 0.319, respectively. Therefore, the estimated 95 percent point in normal space is $\hat{\sigma}_x \Phi^{-1}(0.95) = 0.565 \times 1.645 = 0.929$. Using the estimated h-transformation, the 95 percent point of the usual intake distribution in original space is 154.2 mg.

A balanced repeated replication method was used to estimate the standard deviations of the estimated percentiles. The procedure is based on Fay (1989). The sample is a stratified sample with two primary sampling units per stratum. Some strata were combined to create a sample of 48 strata, each with two primary sampling units. Sixteen replicates were created based on orthogonal contrasts using weights of 1.5 and 0.5 for the two primary sampling units. All operations including the power and grafted polynomial estimation, were carried out for each of the replicate samples. The estimated standard errors of the estimated percentiles, given below the estimates in Table 6, are the square roots of

Table 5. Sample moments for four-day means in the original scale

Dietary Component	Mean	Standard Deviation	Skewness
Calcium (mg)	622.3	295.3	1.18
Energy (kcal)	1683.4	532.9	0.70
Iron (100 mg)	1105.3	387.7	1.17
Protein (10 g)	668.8	208.4	0.89
Vitamin A (μg/RE)	800.8	864.9	7.17
Vitamin C (10 mg)	792.4	515.6	1.24

$\hat{V}\{\hat{\theta}_0\} = 0.25 \sum_{i=1}^{16}(\hat{\theta}_i - \hat{\theta}_0)^2$, where $\hat{\theta}_i$ is the estimated percentiles for the ith replicate and $\hat{\theta}_0$ is the estimate for the original sample.

Monte Carlo Study

A Monte Carlo study was conducted to evaluate the performance of the estimation procedure described in the last section, and to compare our method with two other procedures. The first alternative procedure for estimating the distribution of usual intakes is composed of the following steps.

1. Power-transform the original observations, where the selected power is chosen to minimize the Anderson-Darling statistic.

2. Compute the mean daily intake for each individual using the transformed data.

3. Shrink the individual means of the transformed data for individual i as follows: $\widetilde{x}_i = \hat{\mu}_x + \hat{\sigma}_{\bar{X}}^{-1}\hat{\sigma}_x\left(\bar{X}_{i.} - \hat{\mu}_x\right)$, where $\hat{\mu}_x$ is the mean of the transformed observations, $\hat{\sigma}^2_{\bar{X}}$ is the estimated variance of the transformed means, and $\hat{\sigma}^2_x$ is the estimator of the among-individual variance. The shrunken means have the mean and variance of the usual distribution in the transformed scale.

4. Back-transform the shrunken means to the original scale using a Taylor series approximation to adjust for bias when applying the inverse nonlinear transformation to usual intakes.

Table 6. Estimated percentiles for usual intake distributions

	Percentile						
Component	0.01	0.05	0.10	0.50	0.90	0.95	0.99
Calcium (mg)	209	291	344	590	944	1066	1323
	(10)	(11)	(11)	(14)	(25)	(31)	(44)
Energy (kcal)	796	1020	1148	1652	2258	2459	2883
	(25)	(23)	(24)	(35)	(57)	(65)	(85)
Iron (100 mg)	528	672	751	1074	1506	1656	1986
	(22)	(24)	(24)	(31)	(49)	(59)	(88)
Protein (10 g)	348	430	476	658	877	948	1098
	(12)	(13)	(13)	(15)	(21)	(24)	(34)
Vitamin A (μg/RE)	218	310	371	669	1397	1743	2687
	(10)	(12)	(14)	(31)	(130)	(207)	(474)
Vitamin C (10 mg)	184	286	355	717	1323	1542	2013
	(11)	(15)	(18)	(33)	(59)	(68)	(94)

Note: Values in parentheses are estimated standard errors

5. Estimate the cumulative distribution function of usual intakes from the back-transformed shrunken means by the empirical distribution function.

This procedure is an extension of the suggestions in the report of the National Research Council (NRC 1986). Because the primary difference between the procedure described above and our procedure of the last section is in the transformation of step (a), we call the outlined procedure the *best power* procedure.

The second alternative procedure for estimating usual intake distributions is based on the smoothed empirical distribution of individual mean intakes. This method has been used in the past by practitioners, and is expected to do poorly in the tails because of the presence of within-individual variation in the distribution.

In the simulation, a true usual intake distribution was generated that displays distributional characteristics similar to those of protein. Protein is in the center of the

components studied with respect to skewness and with respect to number of join points. For each of 1,000 samples, an observation Y_{ij} for the jth day ($j = 1, 2$) on the ith individual ($i = 1, ..., 700$) was generated as follows.

- Draw x_i , the individual's usual intake in normal scale from a N (0, 0.36).
- Draw σ^2_{ui} , the measurement error variance, from a 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^2_{ui} , for $j = 1, 2$, and form $X_{ij} = x_i + u_{ij}$, where X_{ij} is the observed intake in normal scale. If X_{ij} falls below -6.97, X_{ij} is set equal to -6.97.

Let $Y_{ij} = L_{ij}^{2.5}$ be the observed intake in the original scale, where L_{ij} is a grafted cubic function of X_{ij}. The definition for L_{ij} is such that no power of the generated intakes is normally distributed. This function relating Y_{ij} and X_{ij} is presented in Figure 3.

Two hundred ninety-one percentiles of the estimated usual intake distributions were computed using the three procedures, and were averaged over the 1,000 samples. The set of percentiles is defined by the 41 percentiles 0.01, 0.025 to 0.975 by 0.025, 0.99, plus percentiles corresponding to 250 equally spaced probabilities. The estimated percentiles were compared to the percentiles of the true usual intake distribution, generated by numerical integration. Results for selected percentiles are shown in Table 7. Figure 4 shows the estimated root mean squared error and the absolute value of the average estimated bias for each of the 291 percentiles for the best power and ISU estimation methods. The root mean squared error and to some degree the bias are larger in the tails than in the center of the distribution for all procedures. While all three methods display some bias, the method proposed in the second section, called ISU in the table, produces estimates of percentiles that are nearly always less biased than the other two methods. The ISU method generally has smaller standard errors than the best power procedure, especially in the tails, and is

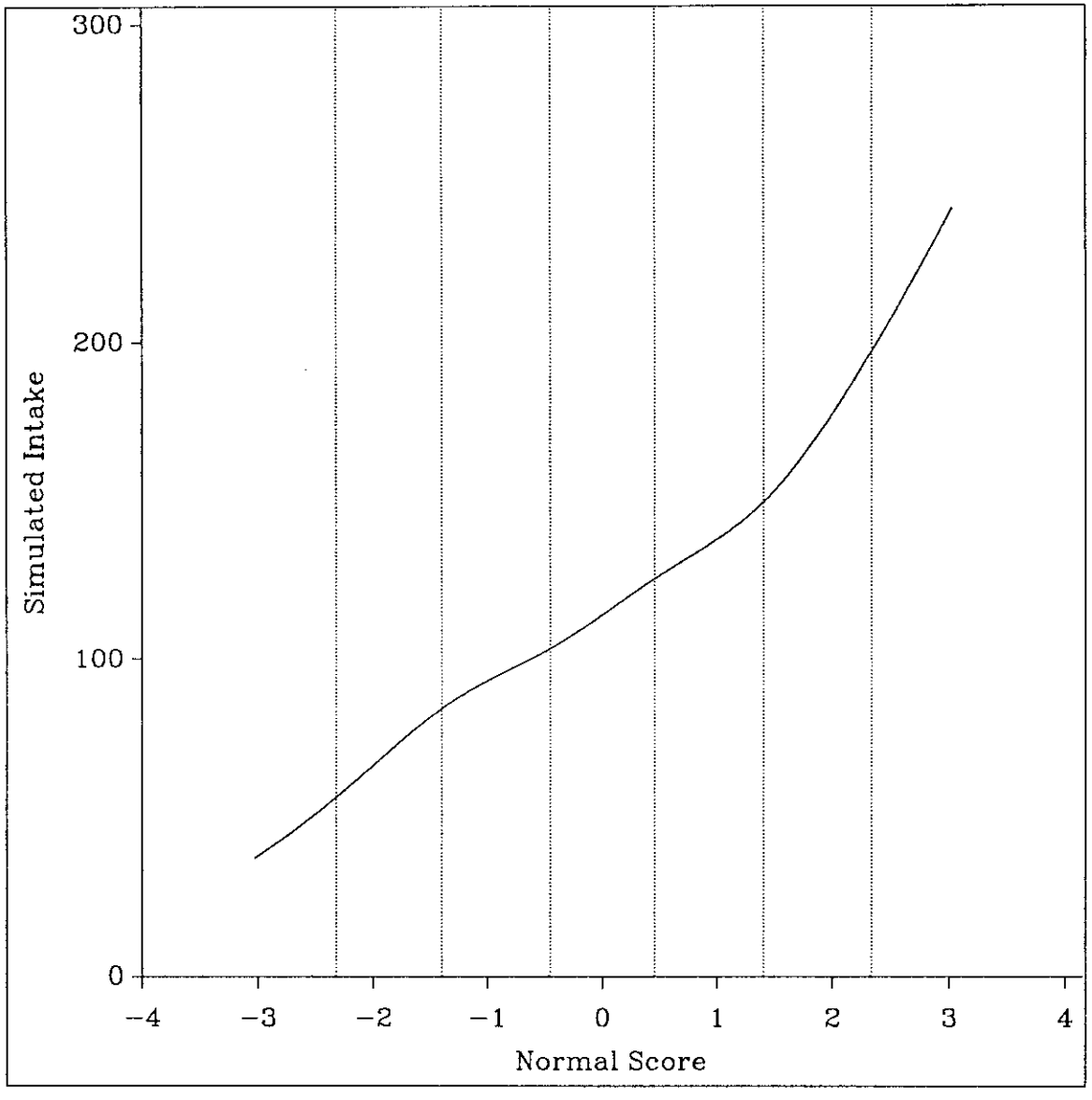

Figure 3. Plot of transformation of normal scale to observed intakes used in simulation

Table 7. Estimates of selected percentiles of the usual intake distribution using three estimation methods, averaged over 1000 simulations

		Estimation Method		
Percentile	True	ISU	Best Power	2-day Mean
0.01	81.85	81.90	79.31	67.56
		(0.078)	(0.103)	(0.108)
		[2.461]	[4.142]	[14.689]
0.05	91.86	91.91	92.03	83.40
		(0.052)	(0.062)	(0.059)
		[1.645]	[1.952]	[8.658]
0.10	96.99	96.87	97.97	91.02
		(0.043)	(0.048)	(0.045)
		[1.358]	[1.806]	[6.138]
0.25	105.44	105.17	106.18	101.99
		(0.033)	(0.036)	(0.035)
		[1.075]	[1.352]	[3.625]
0.50	115.03	115.04	115.17	114.49
		(0.030)	(0.032)	(0.033)
		[0.943]	[1.011]	[1.175]
0.75	125.23	125.62	124.61	127.77
		(0.039)	(0.039)	(0.040)
		[1.301]	[1.389]	[2.833]
0.90	135.42	135.74	134.13	141.62
		(0.058)	(0.060)	(0.065)
		[1.871]	[2.288]	[6.530]
0.95	142.23	142.38	141.50	152.69
		(0.076)	(0.083)	(0.095)
		[2.419]	[2.731]	[10.875]
0.99	157.00	157.16	159.27	180.05
		(0.132)	(0.157)	(0.208)
		[4.165]	[5.468]	[23.960]

Notes: Values in parentheses are estimated standard errors for the Monte Carlo mean percentiles.
Values in brackets are estimated root mean squared errors.

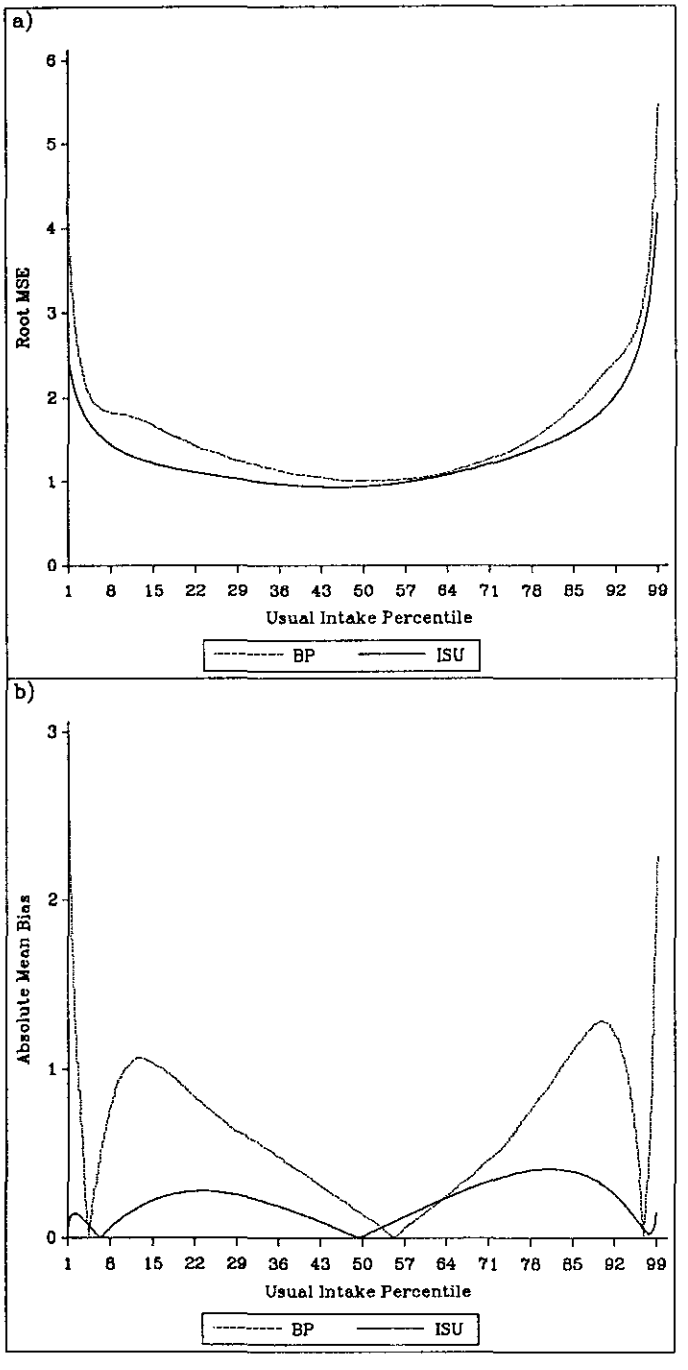

Figure 4. (a) Average estimated root mean squared error in estimated percentiles and (b) absolute average estimated bias from 1000 simulation runs, for the proposed (ISU) and best power (BP) estimation methods

uniformly superior to the best power procedure with respect to mean squared error for all 291 percentiles calculated in the simulation. As expected, the method proposed in Section 2 provides less biased and less variable estimates than estimates based on individual means of the two days. The distribution estimated using individual means is comparable to the other procedures only for percentiles near the mean of the usual intake distribution.

Summary

We have presented a method for estimating distributions of usual intakes based on daily intakes of dietary components that are consumed nearly daily. This method is applicable more broadly to settings where the distribution of nonnormal unobservable means is of interest, and the observed data measure the mean with considerable error that is possibly heteroscedastic. An additional example of such a problem is determining an individual's average blood pressure using multiple measurements.

Our goal was to develop a method that is applicable to many distributional shapes, accommodates a variety of behaviors in the observed data, and is easily implementable via a computer program. A software package called SIDE (Software for Intake Distribution Estimation) is available from the authors that executes the proposed method for data that arise from continuous, unimodal, positive-valued distributions. The user specifies one or more dietary components, adjustment variables, and sample design features. The program provides estimated moments and percentiles for the usual intake distribution along with diagnostics on the performance of various steps in the procedure. Additional modules for this software are being developed to estimate usual food intake distributions from data containing many zeros using a related method (Nusser et al. 1996).

REFERENCES

Bailar, B. A. (1975), "The effects of rotation group bias on estimates from panel surveys," *Journal of the American Statistical Association* 70, 23-29.

Fay, R. E. (1989), "Theory and application of replicate weighting for variance calculations," *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 212-217.

Fuller, W. A. (1976), *Introduction to Statistical Time Series.* Wiley: New York.

Lin, L. I.-K. and Vonesh, E. F. (1989), "An empirical nonlinear data-fitting approach for transforming data to normality," *The American Statistician* 43, 237-243.

Mendelsohn, J. and Rice J. (1982), "Deconvolution of microfluorometric histograms with B-splines," *Journal of the American Statistical Association* 77, 748-753.

National Research Council (1986), *Nutrient Adequacy.* National Academy Press: Washington, DC.

Nusser, S. M., Battese, G. E., and Fuller, W. A. (1990), "Method of moments estimation of usual nutrient intakes distributions," Working Paper 90-WP52, Center for Agricultural and Rural Development, Ames, IA.

Nusser, S. M., W. A. Fuller, and P. M. Guenther (1996), "Estimating usual dietary intake distributions: adjusting for measurement error and nonnormality in 24- hour food intake data," accepted for publication in Survey Measurement and Process Quality International Conference monograph.

Stefanski, L. A. (1990), "Rates of convergence of some estimators in a class of deconvolution problems," *Statistics and Probability Letters* 9, 229-235.

Stefanski, L. A. and Carroll, R. J. (1990), "Deconvoluting kernel density estimators," Statistics 21, 169-184.

Stefanski, L. A. and Carroll, R. J. (1991), "Deconvolution-based score tests in measurement error models," *Annals of Statistics* 19, 249-259.

U.S. Department of Agriculture, Human Nutrition Information Service. (1987), *Continuing Survey of Food Intakes by Individuals, Women 19-50 years and their children 1-5 years, 4 days, 1985.* CSFII Report No 85-4, p. 182.

Wahba, G. (1975), "Interpolating spline methods for density estimation. I. Equal spaced knots," *Annals of Statistics* 3, 30-48.

Wegman, E. J. (1982), "Density estimation," in *Encyclopedia of Statistical Sciences*, eds. S. Kotz and N. L. Johnson, Wiley: New York.