Estimating Usual Dietary Intake Distributions: Adjusting for Measurement Error and Nonnormality in 24-Hour Food Intake Data

Dietary Assessment Research Series Report 6

S.M. Nusser, W.A. Fuller, and Patricia M. Guenther

Staff Report 95-SR 80
December 1995

Estimating Usual Dietary Intake Distributions: Adjusting for Measurement Error and Nonnormality in 24-Hour Food Intake Data

Dietary Assessment Research Series Report 6

S.M. Nusser, W.A. Fuller, and Patricia M. Guenther

Staff Report 95-SR 80
December 1995

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

Sarah M. Nusser is assistant professor, Department of Statistics; Wayne A. Fuller is Distinguished Professor of Statistics, Iowa State University; and Patricia M. Guenther is nutritionist, Agricultural Research Service, U.S. Department of Agriculture.

This research was supported by Cooperative Agreement No. 58-3198-2-006 between the Agricultural Research Service, U.S. Department of Agriculture and the Center for Agricultural and Rural Development, Iowa State University. We thank S.B. Schaller, K.W. Dodd, and Z. Zheng, who developed the software and analyzed food consumption data for this paper.

CONTENTS

FIGURES

TABLES

ESTIMATING USUAL DIETARY INTAKE DISTRIBUTIONS: ADJUSTING FOR MEASUREMENT ERROR AND NONNORMALITY IN 24-HOUR FOOD INTAKE DATA

Introduction

Food consumption data are regularly collected by the U.S. Department of Agriculture (USDA) for the purposes of evaluating dietary status and formulating policy related to dietary consumption. When chronic phenomena such as inadequate dietary intake or long-term exposure to food contaminants are considered, researchers focus on the concept of an individual's *usual* intake. The usual intake of the individual is defined as the long-run average daily intake of a dietary component for the individual. The emphasis on usual intake focuses on long-term patterns of consumption rather than consumption levels on any given day.

Usual intake distributions for dietary components can be used to produce estimates of risk for populations. For example, the proportion of a population whose usual intakes are less (or greater) than a specified value indicative of dietary inadequacy (or excess) can be determined from a usual intake distribution. Alternatively, given a known concentration of a contaminant in a food and a toxic intake level, the proportion of individuals whose long-run consumption is above the toxic intake threshold can be estimated.

The usual intake of a food cannot be directly observed. Direct observation would require respondents to complete dietary intake questionnaires over a long period of time (e.g. a year) without altering their consumption patterns, and then averaging daily intakes from the respondent's sequence to obtain the individual's usual intake. A more realistic method of obtaining information on usual intakes involves asking each respondent to report their daily food intakes on a few randomly selected days. This reduces respondent burden and

increases the chances of obtaining accurate food intake records. However, because daily food intakes measure the usual intake with error and there is large day-to-day variability in consumption, a measurement error framework is needed to estimate usual intake distributions for foods and other dietary components.

Variation in daily intakes is due to two components. One component is measurement or response error, the failure of the respondent to correctly report the amounts of food actually consumed. The second component is associated with the individual's day-to-day variability in food consumption. The sum of these two components produces a large within-person variance component that tends to be heterogeneous across subjects (Hegsted 1972; Hegsted 1982; Beaton et al. 1979; National Research Council [NRC] 1986; Nusser et al.). In addition, the data often display systematic variation associated with day-of-week and day-of-interview. For studies in which 24-hour food intakes are recorded, intakes reported on the first interview day are believed to be more accurate than data collected on subsequent days. It is also well known that intake levels tend to be higher on the weekend than during the week.

When the observed data are approximately normally distributed, fixed and random effects can be easily estimated using simple measurement error models. However, researchers have shown that intake distributions for most dietary components are right skewed (Sempos 1985; NRC 1986; Emrich 1989; Aickin and Ritenbaugh 1991; Carriquiry et al. 1993). For foods that are not consumed daily, the distribution typically has a spike at zero corresponding to nonconsumers of the food, and a unimodal or J-shaped distribution of usual intakes for the consumers in the population (Nusser 1995). Thus, methodology based on a measurement error model must account for the different distributional shapes inherent in daily and usual intakes.

A National Research Council report (1986) represents one of the first attempts to develop a method of estimating usual intake distributions that recognized the presence of within-person variance and nonnormality in daily intake data. They proposed log-transforming the

data, shrinking the log mean intakes so that the shrunken means have variance equal to the estimated among-individuals component, and back transforming the shrunken means. The estimated distribution of usual intakes is the estimated distribution of the back transformed shrunken means.

In cooperation with USDA, researchers at Iowa State University (ISU) have extended these ideas for estimating usual intake distributions for dietary components that are consumed on a nearly daily basis (Nusser et al.). Daily intake data are adjusted for nuisance effects, and the intake data on each sample day are adjusted to have a mean and variance equal to that of the first sample day because it is believed to be the most accurate. The ISU approach then uses a semiparametric procedure to transform the adjusted daily intake data to normality. The transformed observed intake data are assumed to follow a measurement error model, and normal distribution methods are used to estimate the parameters of the model. Finally, a transformation that carries the normal usual intake distribution back to the original scale is estimated. The back transformation of the fitted normal distribution adjusts for the bias associated with applying the inverse of the nonlinear forward transformation to a mean distribution. The back transformation is used to define the distribution of usual intakes in the original scale. The approach was developed with the objective of producing an algorithm suitable for computer implementation and applicable to a large number of dietary components.

Neither the NRC or the ISU method is well-suited for estimating distributions of usual food intakes because many food items are consumed on only a fraction of the sample days for a portion of the population. Daily intake data for infrequently consumed foods contain a substantial number of zero intakes that arise from individuals who never consume the food, or from persons who are consumers, but do not consume the food on sample days. Thus, the measurement error approach must be augmented to account for the mixture of the consumer and nonconsumer distributions that arise with food intake data. Estimation

is complicated by the fact that subpopulation membership (consumer or nonconsumer) may not be identifiable for each subject.

In this paper, we describe a procedure for estimating usual intake distributions for foods and other dietary components that are not consumed on a daily basis. The procedure is an extension of the measurement error approach of Nusser, et al. to settings in which the data arise from a mixture of a single-valued nonconsumer distribution and a continuous, but not necessarily normal, consumer distribution. For the purposes of this paper, we concentrate on the case where an individual's usual intake is unrelated to their probability of consumption. The usual intake for individual i is modeled as the individual's usual intake on days that the food is consumed multiplied by their probability of consuming the food on any given day. The ISU method for estimating usual nutrient intake distributions is applied to the positive intakes to estimate a consumption-day usual intake distribution for the population. The distribution of the probability of consumption is estimated, and used to construct the joint distribution of consumption day usual intakes and consumption probabilities. This joint distribution is used to derive the usual intake distribution for all days.

We begin by describing characteristics of food intake data. The proposed methodology is presented and illustrated with data from the USDA's 1985 Continuing Survey of Food Intakes by Individuals (CSFII).

Characteristics of Food Intake Data

Nusser et al. provide a description of characteristics of daily intake data for dietary components that are consumed daily (or very nearly daily) from the USDA's 1985 CSFII. They found that distributions of daily intakes were generally skewed to the right, that the data contained sizable within-person variability in relation to among-individual variances,

that the within-person variances were related to the mean, and that day-of-the-week and sequence (day-of-interview) effects were significant.

To investigate the patterns of food intake data, daily intakes were examined for several foods from the 1985 CSFII 4-day data set (USDA, 1987). These data were used because they contain four daily intake observations on each individual, and thus contain considerable information regarding the underlying patterns of food consumption for individuals. Food groups were selected that provide a wide range of consumption patterns: dark green vegetables, apples, alcoholic beverages, diet soda, eggs, beef, fruit, and milk products. We used the same set of respondents used by Nusser, et al., consisting of 743 women aged 25-50 who were meal planners/preparers and were not pregnant or lactating. For each food, two data sets were considered: (1) a data set containing intakes for all interview days, and (2) the set of positive intakes from days on which the food was consumed by the respondent. These data sets were examined to consider patterns relevant to the usual intake distribution for all days and the usual intake distribution for consumption days only, respectively.

Table 1 provides information on the frequency of consumption for these food groups. The columns of the table contain the percentage of women who consumed the food group on k out of the 4 days of recorded intake, where $k = 0, 1, 2, 3,$ or 4 . Note that for the more specific and less commonly consumed food groups (dark green vegetables, apples, alcoholic beverages, and diet soda), a substantial proportion of respondents did not consume the food group on any of the sample days. When broader or more commonly consumed classes of foods are considered (e.g., beef, eggs, fruit, and milk products), the percentage of respondents consuming the food group on at least one day is much larger. Except for fruit and milk products, the percentage of women consuming the food group for a consumption frequency class is inversely related to the frequency of consumption. For fruit, however, the percentages across consumption frequency classes are quite similar, and for milk products, the percentage of women in the consumption frequency class increases with the frequency of

Table 1. Percentage of women who consumed from each food group on 0, 1, 2, 3, or 4 of the sample days

	Number of Sample Days on Which Food Group was Consumed				
Food Group	0	1	2	3	4
Dark Green Vegetables	68[a]	25	6	1	0
Apples	69	21	7	2	1
Alcoholic Beverages	70	15	7	3	4
Diet Soda	60	15	10	8	7
Eggs	40	32	19	7	2
Beef	36	38	20	6	0
Fruit	19	22	22	18	19
Milk Products	4	9	16	27	44

[a] Percentage of women consuming the specified food group on the indicated number of sample days.

Source: USDA 1985 Continuing Survey of Food Intakes by Individuals (unweighted).

consumption. In the procedure we propose, at least some women must have more than one day of intake data in order to estimate variance components. For these data, the number of women with two or more positive intakes varies from 48 for dark green vegetables to 646 for milk products.

The distribution of individual mean daily intakes is expected to reflect some of the characteristics that might appear in the usual intake distribution for all days. Relative frequency histograms for individual daily intake means are generally J-shaped, although the histogram for apples appears to be bimodal (Figure 1). The distributions for more frequently consumed foods, such as eggs, beef, fruit, and milk products, are less skewed than for other foods.

When all days of zero intake are removed, the distribution of individual mean daily intakes for consumption days provides information on the shape of the consumption-day usual intake distribution. These distributions are generally unimodal, and exhibit a high degree of skewness (Figure 2). The distributions for a few of the food groups (dark green vegetables, alcoholic beverages, and milk products) are more J-shaped.

As expected, tests based on the positive intake data indicated that the null hypothesis of normality is rejected for all of the food groups (Schaller 1993). In addition, we were unable to find power transformations that produced normal distributions for any of the food groups, indicating that the semi-parametric transformation proposed by Nusser, et al. is needed to transform consumption data to normality.

Within-individual standard deviations calculated from positive intake data plotted against individual means on consumption days indicate that for many food groups, there is a positive relationship between within-person variances and individual means (Schaller 1993). Some of the relationships are not as strong as those observed in Nusser, et al. for nutrient intakes, but a pattern of heterogeneous variances is still evident.

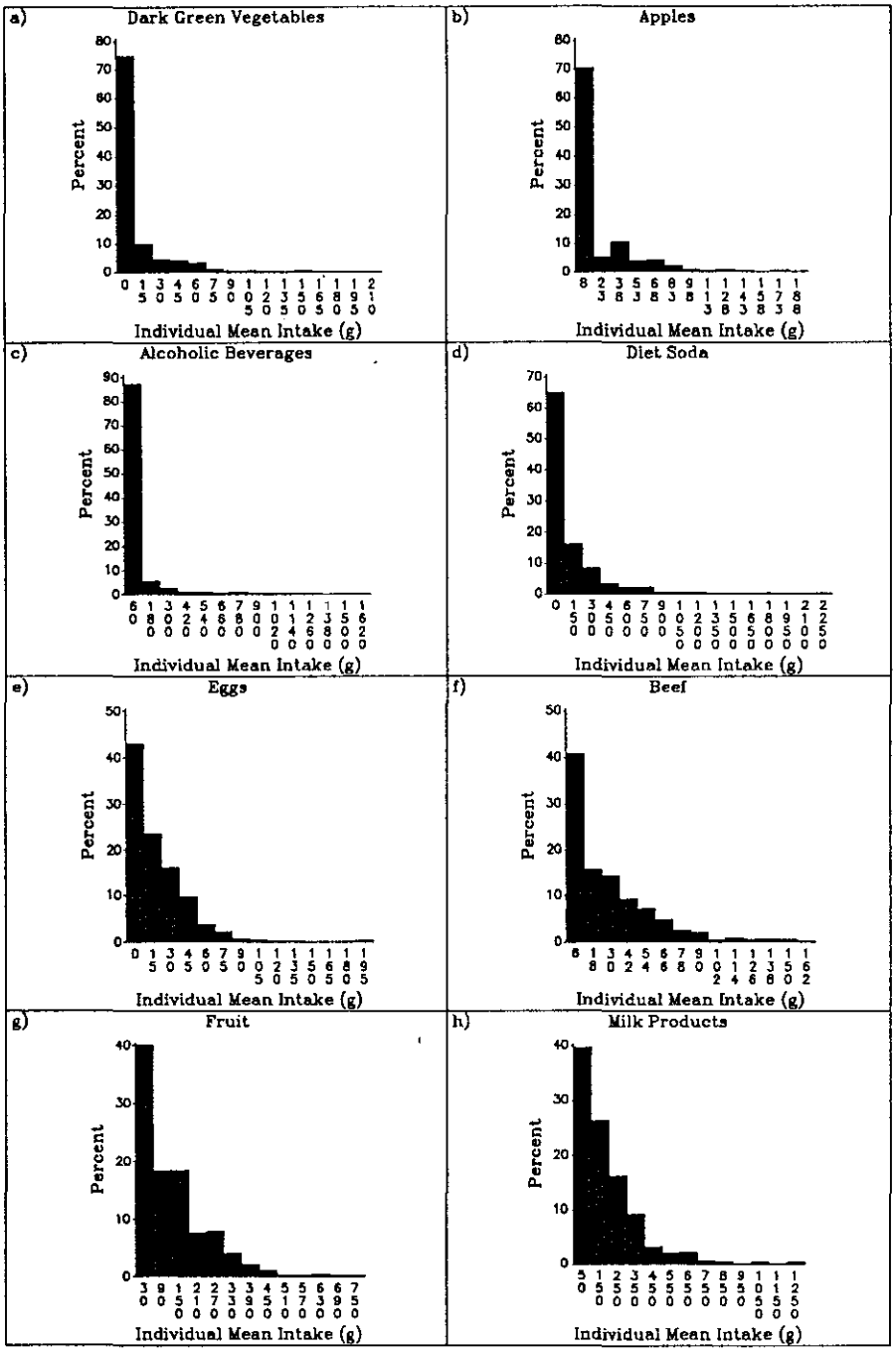

Figure 1. Relative frequency histograms for individual daily intake means calculated using all four days of intake per respondent for (a) dark green vegetables, (b) apples, (c) alcoholic beverages, (d) diet soda, (e) eggs, (f) beef, (g) fruit, and (h) milk products

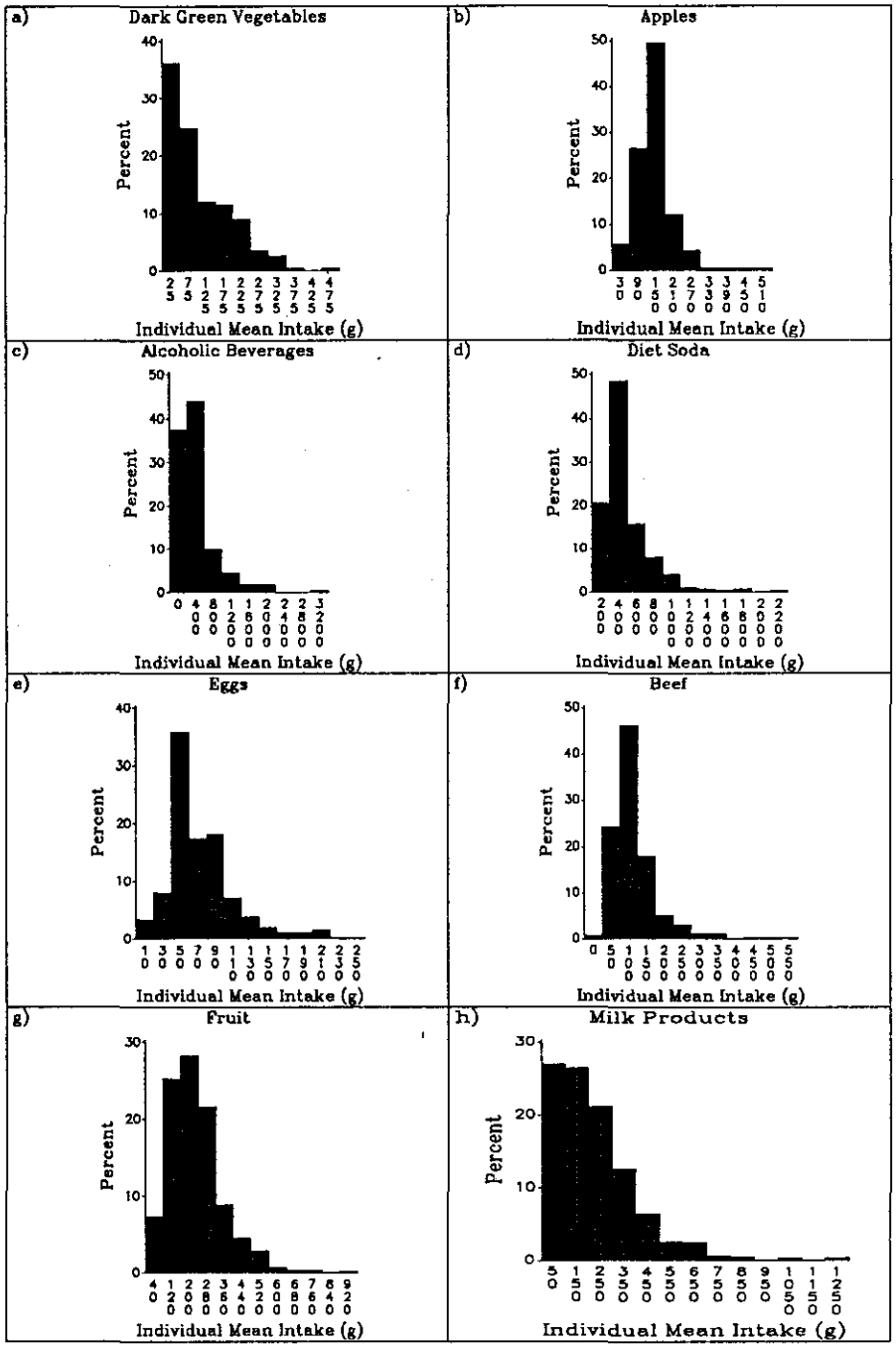

Figure 2. Relative frequency histograms for individual daily intake means calculated using only the positive intake values for each respondent for (a) dark green vegetables, (b) apples, (c) alcoholic beverages, (d) diet soda, (e) eggs, (f) beef, (g) fruit, and (h) milk products

Considerable within-person variability is present in the positive food intake data. Ratios of within-person to among-person variation range from 1.4 to 8.2, with several foods exhibiting ratios in the 3-6 range.

Table 2 presents information on the observed intake distributions for each of the four sample days. For some food groups, the mean positive intake changes substantially across interview days. Mean positive intakes for dark green vegetables and alcoholic beverages are approximately 20% higher on the first interview day than on subsequent days. Intake of beef on the first day is also higher than the mean for the remaining three days. For the other food groups, mean positive intakes are roughly constant across interview days. The patterns in means across days are also observed for the standard deviations across interview days (data not shown). The number of respondents reporting consumption is higher on the first sample day for alcoholic beverages and diet sodas, but is relatively constant across interview days for most of the other food groups.

Exploratory analyses indicate that for some food groups, intakes on consumption days are correlated with the probability of consuming the food. Table 3 presents the mean and standard deviation for intakes on consumption days for women who consume the food on 1, 2, 3, or 4 out of the 4 sample days. No significant correlation was detected between an individual's mean consumption day intake and the number of sample days on which the food was consumed by the individual for dark green vegetables, apples, beef and eggs. However, statistically significant ($p < .01$) positive correlations exist for alcoholic beverages ($r = .19$), diet soda ($r = .43$), fruit ($r = .25$), and milk ($r = .31$). The standard deviation also appears to increase with consumption frequency for diet soda. Results from a test of homogeneity of means across consumption classes confirmed these observations.

Table 2 Mean positive intake (and the number of respondents with positive intakes) for each sample day and for each of eight foods

Food	Sample Day 1	2	3	4
Dark Green Vegetables	122[a] (64)[b]	84 (63)	103 (76)	107 (84)
Apples	130 (83)	135 (60)	137 (98)	153 (92)
Alcoholic Beverages	504 (121)	425 (108)	392 (93)	406 (90)
Diet Soda	534 (183)	534 (172)	537 (143)	524 (158)
Eggs	73 (178)	74 (189)	73 (176)	68 (196)
Beef	118 (177)	112 (191)	109 (162)	103 (180)
Fruit	227 (350)	249 (365)	240 (376)	238 (359)
Milk Products	245 (563)	234 (537)	231 (562)	246 (550)

[a]Mean positive intake (g) on the sample day.
[b]Number of respondents (out of 743) with positive intakes on the sample day.

Source: USDA 1985 Continuing Survey of Food Intakes by Individuals (unweighted).

Table 3. Mean (and standard deviation) of individual mean intakes calculated from positive intakes only, for each consumption frequency, and for each of eight foods

Food	Number of Sample Days on Which Food was Consumed by Respondent			
	1	2	3	4
Dark Green Vegetables	102[a] (89)[b]	106 (79)	122 (120)	-
Apples	141 (71)	133 (46)	148 (45)	137 (51)
Alcoholic Beverages	303 (359)	441 (525)	587 (537)	465 (323)
Diet Soda	358 (177)	442 (200)	512 (282)	705 (383)
Eggs	70 (43)	72 (35)	70 (28)	81 (48)
Beef	111 (72)	112 (54)	106 (39)	-
Fruit	181 (130)	220 (125)	230 (108)	272 (118)
Milk Products	126 (144)	165 (125)	193 (129)	279 (199)

[a] Mean and standard deviation (in parenthesis) of individual means calculated using positive intake values only.

[b] Standard deviation of individual mean intakes (g) calculated using positive intakes.

Source: USDA 1985 Continuing Survey of Food Intakes by Individuals (unweighted).

Estimating Usual Intake Distributions for Infrequently Consumed Dietary Components

Model

For most foods or food groups, at least a portion of the population never consumes the food. Thus, the usual intake distribution for infrequently consumed dietary components generally consists of a spike at zero corresponding to the nonconsumers of the component, and a continuous distribution of positive usual intakes for consumers. The usual food intake on all days for an individual can be modeled as the individual's usual intake on consumption days (usual intake conditional on positive intake) multiplied by the individual's probability of consuming the food on any day. The proposed approach is to set aside the zero intakes and transform the positive intakes to normality using a modification of the measurement error approach described in Nusser, et al. The conditional distribution of usual intake for consumption days is estimated in the normal scale using a measurement error model, and transformed back to the original scale. Next, a distribution of individual consumption probabilities (i.e., the probability that an individual consumes the dietary component on any given day) is estimated. The unconditional usual intake distribution for all days is then estimated from the joint distribution of conditional usual intakes and individual consumption probabilities.

In what follows, let Y_{ij} be the observed intake for individual i on day j and let y_i represent the usual intake of individual i , let p_i be the probability that individual i consumes the dietary component on any given day, let Y_{ij}^* be the observed intake when intake is positive, and let y_i^* be usual intake when intake is positive. Note that $y_i^* = E\{Y_{ij}|i \text{ and } Y_{ij} > 0\}$ is the expected value of the positive intakes for individual i. Our model is

$$y_i = y_i^* p_i, \tag{1}$$

$$p_i \sim g(p\,;\,\theta), \tag{2}$$

$$X_{ij}^* = T\left(Y_{ij}^*;\ \mathbf{a}_{ij},\ \boldsymbol{\alpha}\right), \tag{3}$$

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

$$x_i^* \sim NI\left(\mu_{x^*},\ \sigma_{x^*}^2\right), \tag{5}$$

$$u_{ij} \sim NI\left(0,\ \sigma_u^2\right), \tag{6}$$

$$y_i^* = \eta\left(x_i^*,\ \boldsymbol{\beta}\right), \tag{7}$$

where T is a transformation of Y_{ij}^* that is a function of characteristics of the observations, such as day-of-week and day-of-interview, denoted by $\mathbf{a}_{ij}$, and of a parameter vector $\boldsymbol{\alpha}$; p_i has a distribution g that depends on the parameter vector $\boldsymbol{\theta}$, and η is the back transformation depending on parameter vector $\boldsymbol{\beta}$ that carries normal conditional usual intakes to the original scale. The distribution of usual intakes is determined by the joint distribution of p_i and y_i^*. We assume that the u_{ij} are independent given i, and that p_i is independent of y_i^*.

Because of the complexity of the model, components of the model are estimated separately. First, the transformation T that carries the original Y_{ij}^* into X_{ij}^* is estimated. Using the X_{ij}^*, the parameters $(\mu_{x^*},\ \sigma_{x^*}^2,\ \sigma_u^2)$ and the transformation η are estimated. The parameters of the distribution of p_i are estimated using information on the number of days individuals consume the food. The distribution of y^* is combined with the distribution of p to obtain the distribution of usual intakes, y. Measurement error is important at two points in the estimation. A part of the transformation T adjusts for systematic measurement error by transforming reported consumption on the second, third, and fourth days to the level observed on the first day. The variable u_{ij} represents the day-to-day variability due to variation in consumption and to errors in reporting.

In the work of Nusser et al., the error variance σ_u^2 is permitted to vary from individual to individual. In our analysis of food intakes, we adopt the simpler model of common variance in the normal scale. Extensions to the more complicated model in Nusser et

al. are straightforward. Empirical studies indicate adjustments that account for the heteroscedasticity in the normal scale have little influence on the results.

Estimating the Consumption Day Usual Intake Distribution

Data Adjustments. The positive food intake data are adjusted for nuisance effects, which will vary with the study. For the 1985 CSFII, adjustments are made for day-of-week and day-of-interview effects. The distribution of daily intakes on the first sample day is taken to be the reference standard because it is considered to be the most accurate information available in the sample.

The adjustment for nuisance effects is similar to a standard regression approach. The data are regressed on variables representing nuisance effects using a linear model. For example, dummy variables for day-of-week are included in the regression. The estimated model is used to adjust the data to the first-day mean (rather than the grand mean) using a procedure that is the multiplicative analog of the standard linear adjustment to increase the chances that adjusted intakes are nonnegative. Let $\left\{Y_{ij}^{*} : i = 1,\ 2,\ ...,\ n^{*}\right.$ individuals, $\left. j = 1,\ 2,\ ...,\ r_i \text{ days}\right\}$ be the set of unadjusted positive observed intakes for a dietary component, where n^{*} is the number of individuals with at least one positive intake and r_i is the number of positive intakes for individual i . Let $\bar{Y}_{.1}^{*}$ be the (weighted) mean of the day one positive intakes, and $\hat{Y}_{ij}$ be the predicted values from the (weighted) multiple regression of Y_{ij}^{*} on the nuisance effect variables. The data adjusted for nuisance effects are

$$Y_{aij}^{*} = \hat{Y}_{ij}^{*-1}\bar{Y}_{.1}^{*}Y_{ij}^{*} . \tag{8}$$

A second transformation is applied to these data to produce approximately homogeneous distributions across days in the normal scale. The positive data are transformed to approximate normality using a power transformation. A grid search is used to determine the power transformation that brings the data closest to normality. A segmented linear transformation is used to center and scale the data on day j $(j = 2,\ 3,\ ...,\ r)$ such that

the data on day j have the mean and variance of day one. Let γ be the power that best transforms the positive data Y^*_{aij} to normality, and let

$$V^*_{ij} = Y^{*\gamma}_{aij}\,.$$

The sample mean and variance of the transformed positive intakes on day j are denoted by $\hat{\mu}_j$ and $\hat{\sigma}^2_j$, respectively. The data for day j are adjusted to the day one mean and variance as follows:

$$\tilde{V}^*_{ij} = \begin{cases} \hat{\mu}_1 + \hat{\sigma}_j^{-1}\hat{\sigma}_1\left(V^*_{ij} - \hat{\mu}_j\right), & \text{if } V^*_{ij} > 2\,|a_j| \\ \\ \hat{\mu}_1 + \hat{\sigma}_j^{-1}\hat{\sigma}_1\left(V^*_{ij} - \hat{\mu}_j\right) - b_j\left[1 - (2\,|a_j|)^{-1}\,V^*_{ij}\right], & \text{otherwise} \end{cases} \tag{9}$$

where

$$a_j = \hat{\mu}_j - \hat{\sigma}_1^{-1}\hat{\sigma}_j\hat{\mu}_1$$

and

$$b_j = \hat{\mu}_1 - \hat{\sigma}_j^{-1}\hat{\sigma}_1\hat{\mu}_j\,.$$

The constants a_j and b_j are the points of intersection between the line defined in the first component of equation (9) and the V^* and $\tilde{V}^*$ axes, respectively. The second line of equation (9) is a modification to the linear transformation in the first line to insure that adjusted transformed intakes are positive and that zero intakes are transformed into zero intakes (empirical results indicate that very few, if any, observations fall into the $[0,\ 2\,|a_j|]$ interval). Adjusted original-scale intakes are defined by

$$\tilde{Y}^*_{ij} = \tilde{V}^{*1/\gamma}_{ij}.$$

Obtaining an Equal-Weight Sample. Our procedure is designed for complex samples in which individuals have different sample weights. For ease of analysis and to increase computational efficiency, a new set of sample values is constructed in which each individual has the same weight. The sample distribution function of the new values is a close approximation to that of the adjusted data.

The new values are generated using a smooth estimate of the cumulative distribution function for the adjusted intakes. To estimate the cumulative distribution function, the weight for a positive intake on day j for individual i is defined by

$$w_{ij} = r_i^{-1} w_i \ ,$$

where w_i is the original sample weight for individual i , and r_i is the number of positive intakes recorded for individual i , $i = 1, 2, ..., n^*$. A piecewise linear estimator, $\hat{F}$, of the distribution function for observed intakes, F, is developed by connecting the midpoints of the rises in the empirical cumulative distribution function using procedures outlined in Nusser, et al.

An equal-weight sample is constructed from $\hat{F}$ by calculating n^* intake values at equal probability intervals. These n^* equal-weight values replace the ranked set of the first observed positive intakes recorded for each of the n^* individuals. Using $\ddot{Y}^*_{s(1)}$ to denote the equal-weight sample value for the first observed positive intake for an individual whose adjusted positive intake has rank s , we have

$$\ddot{Y}^*_{s(1)} = \hat{F}^{-1}\left(\frac{s - 0.5}{n^*}\right),$$

where $s = 1,\ 2,\ ...,\ n^*$. The first positive reported intake value of individual i , which is of rank s among the $\tilde{Y}^*_{i1}$, is replaced by $\ddot{Y}^*_{s(1)}$ and denoted by $\ddot{Y}^*_{i(1)}$. Positive intakes from subsequent interviews are adjusted to maintain the individual's day-to-day structure. For the t-th subsequent day of observed positive intake for individual i , where $t = 2, ..., r_i$, the intake for individual i is defined by

$$\ddot{Y}^*_{i(t)} = \tilde{Y}^{*-1}_{i(1)} \ddot{Y}^*_{i(1)} \tilde{Y}^*_{i(t)}.$$

We let $\ddot{Y}^*_{ij}$ denote the adjusted equal weight values, where j denotes the original sample day.

Transformation to Normality. The adjusted equal-weight sample values, $\ddot{Y}^*_{ij}$, are transformed to normality using the 2-step semiparametric procedure defined in Nusser, et

al. A grid search is used to determine the power transformation that minimizes the squared deviation between the normal score and the power transformed data. Then a smooth cubic spline is used to estimate the function that takes the power transformed data to normality. This procedure can be viewed as a semi-parametric version of the Lin and Vonesh (1989) procedure. The function created by this two-step process is defined to be the function carrying the adjusted data to normality. The transformed positive intakes are defined by

$$X_{ij}^{*} = \omega\left(\ddot{Y}_{ij}^{*}\right),$$

where ω is used to denote the transformation composed of the spline transformation applied to the power of the adjusted equal-weight observations.

Estimating the Conditional Usual Intake Distribution in Normal Scale. The normal data are used to estimate a distribution of usual intakes based on the measurement error model proposed by Nusser, et al. The transformed positive intakes, X_{ij}^{*} are assumed to satisfy (4), (5), and (6).

Under this model, the mean μ_x of the normal usual intake distribution is estimated by the simple estimator

$$\hat{\mu}_x = n^{*-1} \sum_{i=1}^{n^*} \bar{X}_{i\cdot}^{*},$$

where

$$\bar{X}_{i\cdot}^{*} = r_i^{-1} \sum_{j=1}^{r_i} X_{ij}^{*}.$$

The variance $\sigma_{x\cdot}^2$ is estimated using Henderson's method III (Graybill, 1976), which is a method-of-moments variance-components estimation procedure for unbalanced data.

Distribution of Positive Usual Intakes. Given estimates of $\sigma_{x\cdot}^2$ and σ_u^2, the transformation η of (7) is estimated. This requires two steps. A set of estimated x^*-values, denoted by $\ddot{x}^*$, are created with the property that the mean and variance of the set is equal to the estimated mean and variance of x^*. The estimated x^* values are defined by

$$\ddot{x}_i^{*} = \left(\hat{\sigma}_{x^*}^2 + r_i^{-1}\hat{\sigma}_u^2\right)^{-1/2} \hat{\sigma}_{x^*} \bar{X}_{i\cdot}^{*}.$$

Then the estimated usual intake that would be generated by an individual with usual intake $\ddot{x}_i^*$ is calculated as

$$\ddot{y}_i^* = \sum_{i=-4}^{4} b_i \omega^{-1} \left(\ddot{x}_i^* + c_i \right),$$

where ω^{-1} is the inverse of the function ω and $(b_i,\ c_i)$, $i = -4, -3, ..., 4$ is such that $\sum_{i=-4}^{4} b_i c_i = 0$, $c_i = -c_{-i}$, $\sum_{i=-4}^{4} b_i c_i^2 = \sigma_u^2$, and $\sum_{i=-4}^{4} b_i c_i^4 = 3\sigma_u^4$.

The function η is then estimated by the spline regression of $\ddot{x}_i^*$ on the power of the $\ddot{y}_i^*$. The power used in ω is used in η, and the spline procedure is that defined in the semiparametric transformation ω. The distribution of positive usual intakes is defined as the distribution of $y^* = \eta(x^*)$, where $x^* \sim N(\mu_{x^*},\ \sigma_{x^*}^2)$.

Estimating the Distribution of Individual Consumption Probabilities

To estimate the unconditional distribution of usual intakes, an estimate of the distribution of the individual probabilities of consumption, $g(p)$, is required. The information available to support estimation is the proportion of sample days on which the food is consumed by individual i . Attempts to model the distribution of consumption probabilities with logistic regression and with a Beta distribution using these data produced unsatisfactory results.

Therefore it was decided to model the consumption probability distribution as a discrete distribution with K probability values, p_k, each with probability mass θ_k. In the examples below, we use $K = 51$ equally spaced mass points, $\{p_k\} = \{0.0, 0.02, 0.04, ..., 1.0\}$. Let $\hat{\Psi}_l$ denote the observed (weighted) relative frequency of individuals who consume the food on l out of r days, where $l = 0, 1, ..., r$. The $\hat{\Psi}_l$ are assumed to arise from a mixture of the K binomial probabilities of consumption on l out of r days, with binomial parameters of r and p_k, and mixture parameters $\boldsymbol{\theta} = (\theta_1,\ \theta_2,\ ...,\ \theta_K)$, where $\theta_k = [0,\ 1]$ and $\sum_{k=1}^{K} \theta_k = 1$. Hence, the expected value for $\hat{\Psi}_l$ is equal to

$$\Psi_l(\boldsymbol{\theta}) = \sum_{k \in A_l} \theta_k \binom{r}{l} p_k^l (1 - p_k)^{r-l},$$

where

$$A_l = \begin{cases} \{1, 2, .., K-1\} & \text{if } l < r \\ \{2, 3, ..., K\} & \text{if } l = r. \end{cases}$$

Using the notation above, the minimum chi-squared estimator for this problem is defined as the value of $\boldsymbol{\theta}$ that minimizes

$$n \sum_{l=0}^{r} \left[\hat{\Psi}_l - \Psi_l(\boldsymbol{\theta})\right]^2 \left[\Psi_l(\boldsymbol{\theta})\right]^{-1}$$

(Agresti, 1990, p. 471). However, in our problem, the number of parameters, K, exceeds the number of terms in the chi-squared objective function, $r+1$. Thus, we include an entropy term in the objective function to smooth the observed distribution over the K mass points of the distribution.

Entropy, as a measure of uncertainty, was introduced by Shannon (1948), and its use as a principle in statistical estimation was discussed by Jaynes (1957). Maximum entropy estimation is often used when the number of parameters to be estimated exceeds the amount of data available for estimation. The K probabilities, θ_k, of a discrete distribution with K mass points are obtained by maximizing

$$\Gamma = -\sum_{k=1}^{K} \theta_k \ln \theta_k$$

subject to $\sum_{k=1}^{K} \theta_k = 1$ and constraints that represent the known information regarding the θ_k, where $\theta_k \in [0,1]$ and $\theta \ln \theta$ is zero for $\theta = 0$. In the absence of any prior information, Γ is maximized when $\theta_k = K^{-1}$ for all k ; that is, when there is complete uncertainty about the probability of the K events. The function Γ reaches a global minimum when θ_k is one for some k and zero for all other values of k.

The modified minimum chi-squared estimator for our problem is defined as the value of $\boldsymbol{\theta}$ that minimizes

$$n \sum_{l=0}^{r} \left[\hat{\Psi}_l - \Psi_l(\boldsymbol{\theta})\right]^2 \tilde{\Psi}_l^{-1} + \sum_{k=2}^{K} \frac{\theta_k}{1-\theta_1} \ln\left(\frac{\theta_k}{1-\theta_1}\right),$$

where $\sum_{k=1}^{K} \theta_k = 1$, $\theta_k \in [0, 1]$,

$$\tilde{\Psi}_l = \begin{cases} \max\left\{\hat{\Psi}_0,\ \left(1 - \bar{\Psi}\right)^r\right\} & \text{if } l = 0 \\ \max\left\{\hat{\Psi}_r,\ \bar{\Psi}^r\right\} & \text{if } l = r \\ \hat{\Psi}_l \left(1 - \tilde{\Psi}_0 - \tilde{\Psi}_r\right) \left(1 - \hat{\Psi}_0 - \hat{\Psi}_r\right)^{-1} & l = 1, 2, ..., r-1 \end{cases}$$

and $\bar{\Psi} = n^{-1} \sum_{i=1}^{n} (r_i/r)$. The chi-squared term contains the sample information. The modified denominator in the chi-squared term prevents numerical difficulties that can arise when $\hat{\Psi}_l = 0$. The resulting estimator is closely related to the modified chi-squared estimator (Agresti, 1990, p. 472). The maximum entropy term smooths the mass across all possible values of p_k given the sample information in the chi-squared term. Note that θ_1, which is the proportion of the population that never consumes the food, is not inlcuded in the entropy term. The θ_2, θ_3, ..., θ_K are the parameters associated with consumers and θ_1 is the fraction of nonconsumers in the population.

A FORTRAN program using an IMSL subroutine was written to produce estimates of the consumption probability distribution parameters $\boldsymbol{\theta}$. Details of this procedure are presented in Zheng (1995). Let $g\left(p,\ \hat{\theta}\right)$ denote the estimated consumption probability distribution.

Estimating the Unconditional Usual Intake Distribution

The unconditional usual intake for individual i , denoted by y_i , is

$$y_i = y_i^* p_i ,$$

where y_i^* is the usual intake given that the food is consumed and p_i is the probability of positive consumption. The distribution of usual intakes for consumers is the distribution of $y_i^* p_i$, which can be derived from the joint distribution of y_i^* and p_i . If y_i^* is independent of p_i , then the joint distribution of y_i^* and p_i is

$$f\left(y_i^*\right) g\left(p_i\right).$$

The cumulative distribution function of the unconditional usual intakes is then

$$H\left(y\right) = \Pr\left(y^* p \leq y\right) = \theta_1 + \sum_{k=2}^{K} \theta_k \int_0^{y/p_k} f\left(y^*\right) dy^*.$$

Numerical integration is required to estimate the unconditional usual intake distribution H. An adaptive quadrature algorithm was developed to perform these calculations.

Application to 1985 CSFII Data

Our estimation methods were applied to dark green vegetable, apple, egg, and beef consumption data for the 743 nonpregnant, nonlactating, 25-50 year-old women from the 1985 Continuing Survey of Food Intakes by Individuals (CSFII) described in Section 2. On the basis of the analyses associated with Table 3, it is assumed that the conditional usual intake for these foods is independent of the probability of consumption. First, adjusted equal-weight positive intakes were used to estimate the consumption-day usual intake distribution using the semiparametric transformation approach described in Section 3.2. The semiparametric transformation provided a substantial improvement over the power transformation initially selected as the best power for transforming the data to normality for all foods. The plot of the estimated transformation of the observed apple intakes to normality presented in Figure 3 demonstrates that power transformations are inadequate. The ratio of the estimated measurement error variance to the usual intake variance in normal scale $\left(\hat{\sigma}_{x^*}^{-2}\hat{\sigma}_u^2\right)$ ranged between 1.1 and 5.6, indicating large within-person variability relative to among-person variance for these foods. A plot of the estimated consumption day usual intake distribution for each of the four foods is presented in Figure 4. The estimated mean, standard deviation, and skewness coefficient for the conditional usual intake distribution are presented in Table 4. The mean consumption day intakes are roughly one medium serving for dark green vegetables and apples, and about 1 - 1.5 servings for eggs and beef (Pao et al., 1982).

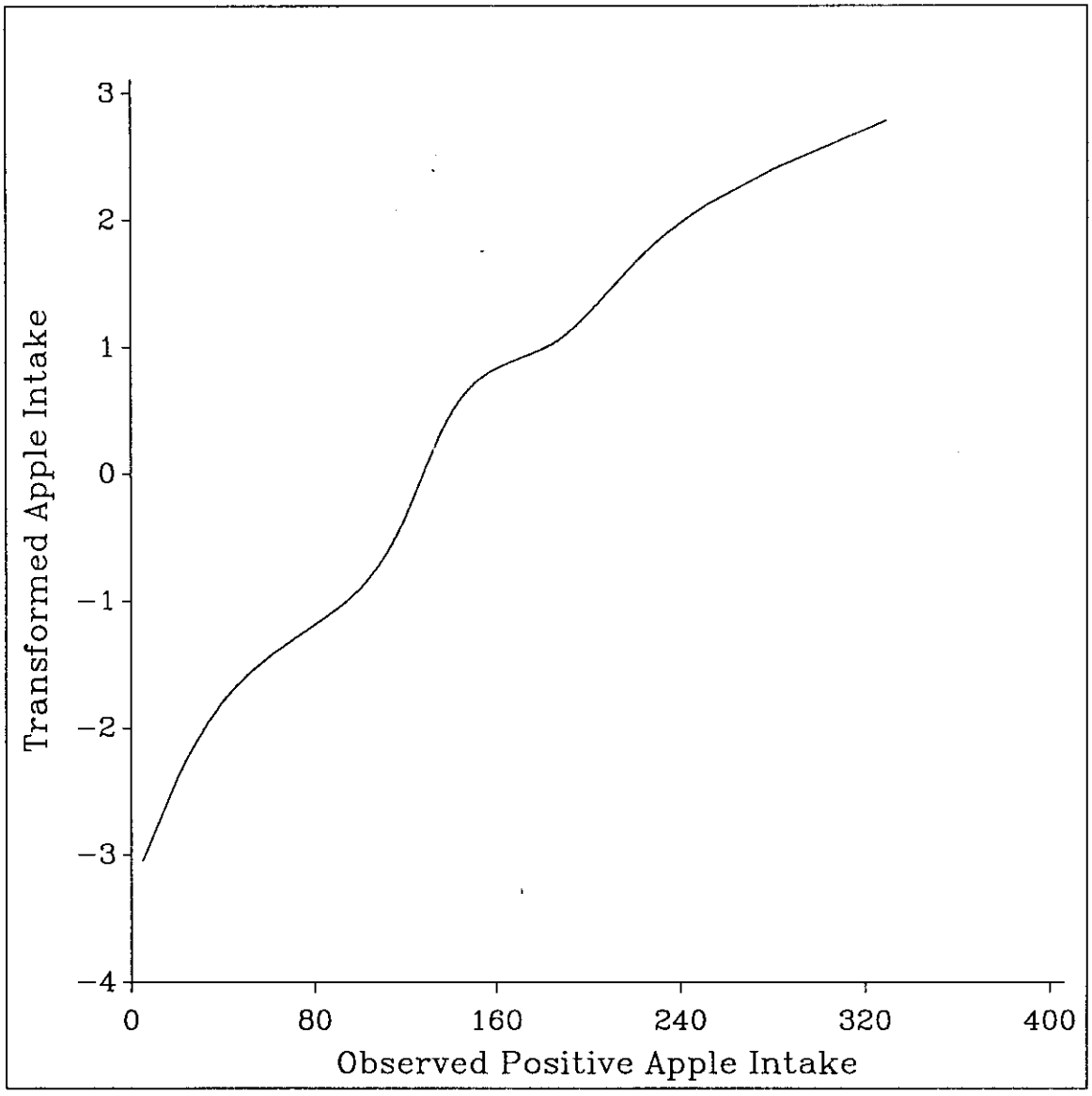

Figure 3. Estimated transformation from original scale to normality for observed apple intakes

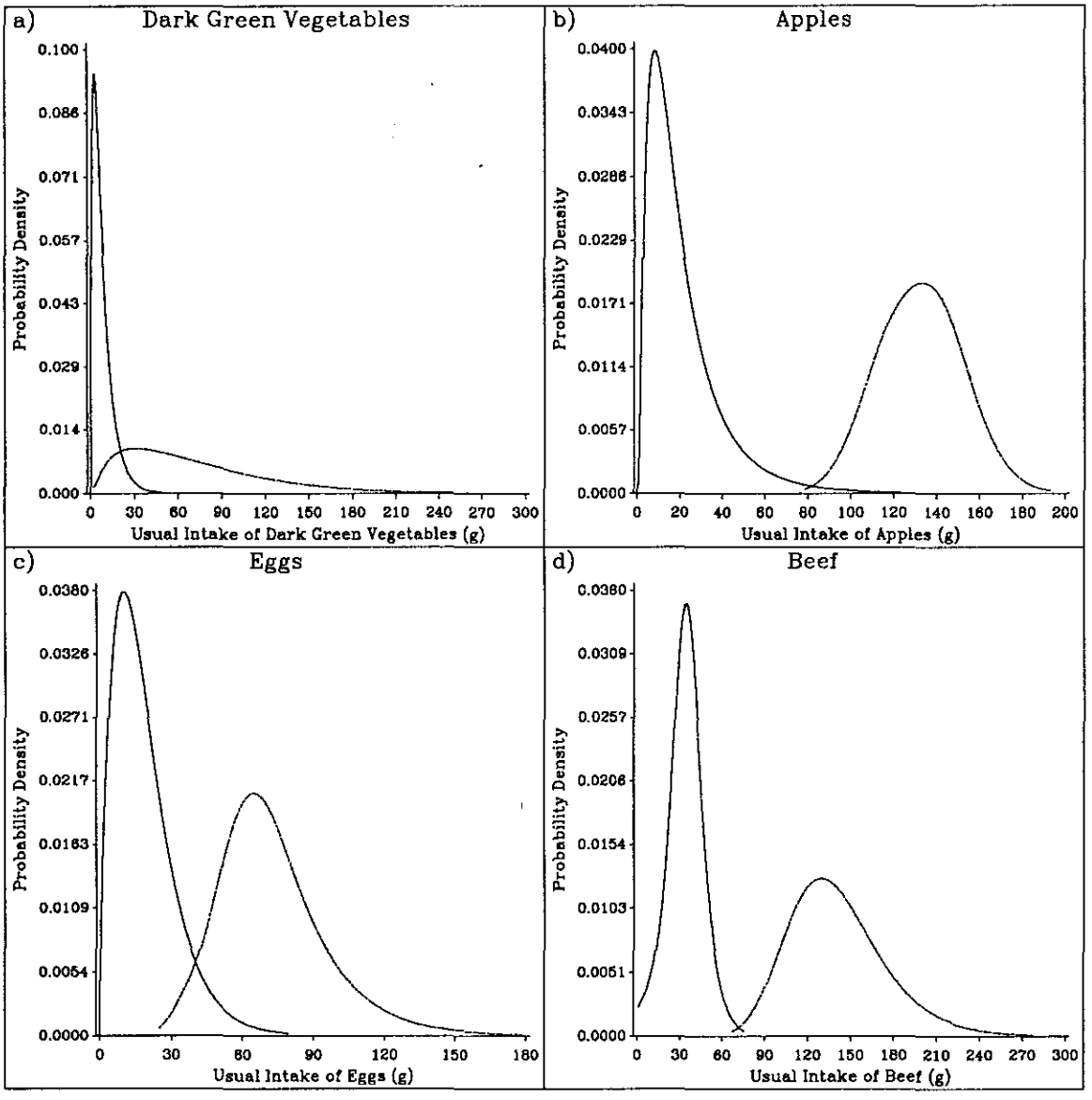

Figure 4. Estimated usual intake distributions on consumption days (dashed line) and on all days (solid line) for (a) dark green vegetables, (b) apples, (c) eggs, and (d) beef

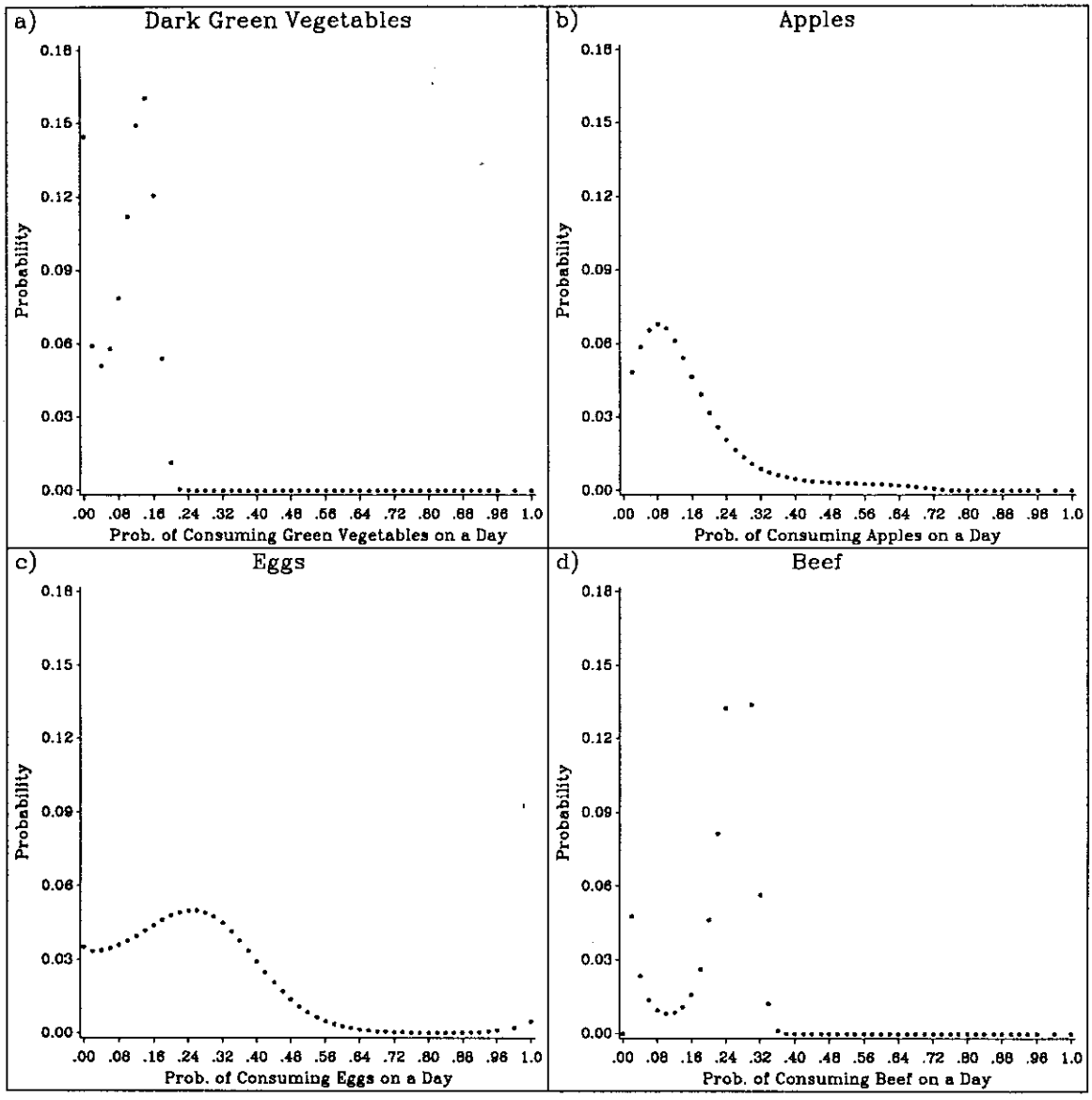

Figure 5. Estimated distribution of consumption probabilities for (a) dark green vegetables, (b) apples, (c) eggs, and (d) beef

Table 4. Estimated mean, standard deviation (SD), and skewness coefficient for usual intake distributions on consumption days only and on all days for consumers (in grams), and estimated proportion of nonconsumers

	Consumption Day Usual Intakes				Usual Intakes On All Days for Consumers		
Food Group	Mean	SD	Skewness	Percentage Nonconsumers	Mean	SD	Skewness
Dark Green Vegetables	72	52	1.3	14.5	1	7.4	2.2
Apples	132	21	0.1	29.7	21.1	18.3	2.5
Eggs	73	24	1.1	3.5	18.7	13.3	1.4
Beef	141	34	0.8	0.003	33.4	12.5	0.6

Source: USDA 1985 Continuing Survey of Food Intakes by Individuals (unweighted).

The consumption probability distribution for each food was estimated using the observed relative frequencies for consumption classes listed in Table 1. The distributions are presented in Figure 5, and the estimated proportion of nonconsumers are listed in Table 4. For apples, the mean of the distribution is 0.11, indicating that on average, apples are consumed about once every ten days by this population. Dark green vegetable and beef consumption tends to have sharply defined peaks, indicating modal consumption patterns. Our procedure for estimating consumption probability distributions proved to be very flexible. Estimated distributions for fruit and milk products (not presented here) reflected the flat and increasing shapes, respectively, expected for these foods based on the consumption patterns noted in Table 1.

Under the assumption that intake levels and the probability of consumption are independent, the consumer usual intake distribution for all days was estimated using equation (3). The estimated distributions for consumers are presented in Figure 4. The estimated mean, standard deviation, and skewness coefficient for the consumer usual intake

distributions are listed in Table 4. The results indicate that a variety of shapes can be estimated using this procedure.

Summary

We have developed a method for estimating the distribution of an unobservable random variable from data that are subject to considerable measurement error and that arise from a mixture of two populations, one of which has a single-valued distribution and the other having a continuous unimodal distribution. Although we motivate the methodology development with a specific problem in dietary assessment, the method is more broadly applicable. Mixture populations arise frequently in health studies and in reliability studies in industry, and noisy, nonnormal data are ubiquitous.

The method requires that at least two positive intakes be recorded for a subset of the subjects in order to estimate the variance components for the measurement error model. Thus, this procedure may not be applicable to foods that are so rarely consumed by the population under study that very few individuals report two consumption days.

The specific approach presented here is appropriate for food intakes when the probability that a subject consumes a food is unrelated to the usual intake on consumption days. Further work is planned to develop models that account for dependence between the probability of consumption and conditional usual intakes. Such an extension will permit the methods to be applied to a broader set of foods. In addition, research is under way to develop variance estimators for the estimated parameters of the usual intake distribution, and software is being written so that our methods can be readily implemented.

REFERENCES

Agresti, A. 1990. *Categorical Data Analysis.* New York: John Wiley & Sons.

Aickin, M. and C. Ritenbaugh 1991. "Estimation of the true distribution of vitamin A intake by the unmixing algorithm." *Communications in Statistics-Simulations,* 20:255-280.

Beaton, G. H., J. Milner, P. Corey 1979. "Sources of variance in 24-hour dietary recall data: implications for nutrition study design and interpretation." *American Journal of Clinical Nutrition* 32:2546-2559.

Carriquiry, A. L., H. H. Jensen, K. W. Dodd, S. M. Nusser, P. M. Guenther, and W. A. Fuller 1993. "Estimating usual intake distributions."

Emrich, L. J., D. Dennison, K. F. Dennison 1989. "Distributional shape of nutrition data." *Journal of American Dietetics Association* 89:665-670.

Graybill, F. A. 1976. *Theory and Application of the Linear Model.* Pacific Grove, California: Wadsworth & Brooks/Cole Advanced Books & Software.

Hegsted, D. M. 1972. "Problems in the use and interpretation of the recommended daily allowances." *Ecology of Food and Nutrition* 1:255-265.

Hegsted, D. M. 1982. "The classic approach - The USDA nationwide food consumption survey." *The American Journal of Clinical Nutrition* 35:1302-1305.

Jaynes, E. T. 1957. "Information theory and statistical mechanics." *Physics Review* 106:620-630.

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

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

Nusser, S. M. 1995. "Estimating usual intake distributions from 24-hour dietary intake data." *Proceedings of the Section on Epidemiology, American Statistical Association.* Alexandria, Virginia: American Statistical Association. pp. 49-58.

Nusser, S. M., A. L. Carriquiry, K. W. Dodd, and W. A. Fuller. "A semiparametric transformation approach to estimating usual daily intake distributions." Accepted with revision by *Journal of the American Statistical Association.*

Pao, E. M., K. H. Fleming, P. M. Guenther, and S. J. Mickle 1982. *Foods Commonly Eaten by Individuals: Amount Per Day and Per Eating Occasion.* HERR No. 44, U.S. Department of Agriculture, Human Nutrition Information Service, Hyattsville, MD.

Sempos, C. T., N. E. Johnson, E. L. Smith, C. Gilligan, C. 1985. "Effects of intraindividual and interindividual variation in repeated dietary records." *American Journal of Epidemiology* 121:120-130.

Schaller, S. 1993., "Estimating usual intake distributions for dietary components with many zero daily intakes." Unpublished Creative Component for M.S. degree, Ames, Iowa, USA: Department of Statistics, Iowa State University.

Shannon, C. E. 1948. "The mathematical theory of communication." *Bell System Technical Journal* 27:379-423.

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.

Zheng, Z. 1995. "Modified minimum chi-squared estimation of food consumption probability distributions." Unpublished Creative Component for M.S. degree, Ames, IA, USA: Department of Statistics, Iowa State University.