A Semiparametric Transformation Approach to Estimating Usual Daily Intake Distributions

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

Working Paper 92-WP 99
September 1992

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

S.M. Nusser is an assistant professor of statistics, Iowa State University; A.L. Carriquiry is an assistant professor of statistics, Iowa State University; and W.A. Fuller is a distinguished professor of statistics, Iowa State University.

This research was supported in part by Research Agreement No. 58-9-032 between the Human Nutrition Information Service, USDA, and CARD.

ABSTRACT

The distribution of usual intakes of dietary components is important to individuals formulating food policy and to persons designing nutrition education programs. Usual intake of a dietary component for a person is the long run average of daily intakes of that component for that person. Because it is impossible to directly observe usual intake for an individual, it is necessary to develop an estimator of the distribution of usual intakes based on a sample of individuals with a small number of daily observations on each individual. Daily intake data for individuals are nonnegative and often very skewed. Also, there is large day–to–day variation relative to the individual–to–individual variation and the within–individual variance is correlated with the individual means. We suggest a methodology for estimating usual intake distributions that allows for varying degrees of departure from normality and recognizes the measurement error associated with daily dietary intakes. The estimation method contains four steps. First, the original data are standardized by adjusting for weekday and interview sequence effects. Second, the daily intake data are transformed to normality using a combination of power and grafted polynomial transformations. Third, using a normal components–of–variance model, the distribution of usual intakes is constructed for the transformed data. Finally, a transformation of normal usual intakes to the original scale is defined. The approach works well for a set of dietary components selected from the 1985–1986 Continuing Survey of Food Intakes by Individuals data. The selected components display a range of distributional shapes.

KEY WORDS: Measurement error models, nutritional status, Continuing Survey of Food Intakes by Individuals, density estimation.

1. INTRODUCTION

The United States Department of Agriculture has been responsible for conducting periodic surveys to estimate food consumption patterns of household and individuals in the United States since 1936. Because dietary intake data from these surveys are used to formulate food–assistance programs, consumer education and food regulatory activities, it is crucial that appropriate methodologies be used in the analysis of these data. However, inappropriate assumptions of normality and failure to recognize the measurement error inherent in mean observed daily intakes as an indicator of the usual daily intake (i.e., the normal or long–run average daily intake) often occur in the analysis of dietary intake data (Lörstad, 1971; Hegsted, 1972, 1982; National Research Council, 1986). This article outlines a methodology which recognizes that usual intake distributions are typically nonnormal and provides an appropriate estimate of the usual intake distribution from daily dietary intake data.

In evaluating the adequacy of diets, it is recognized that an individual who has a low intake of a given dietary component on one day is not necessarily deficient or at risk of deficiency for the dietary component under consideration. It is low intake over a sufficiently long period of time that produces dietary inadequacy. A dietary deficiency exists when the usual daily intake of the dietary component is less than the appropriate dietary standard, where usual intake is the long run average of daily intakes. The same concepts apply to excessive intakes.

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 intraindividual variability and, hence, is greater than the variance of the usual intakes. Other parameters of the distribution of mean intakes may differ from the parameters of the distribution of usual intakes. Because of these problems, using the mean

intake distribution as an estimate of the usual intake distribution can lead to erroneous inferences regarding nutritional status. For example, if the mean daily intake distribution is used to estimate the proportion of the population whose usual daily intakes fall below an intake level indicative of dietary deficiency, the overdispersion of the mean intake distribution relative to the usual intake distribution will lead to an inflated estimate of the proportion of individuals at risk for dietary inadequacy.

Two alternative approaches to estimating the usual intake distribution are to (a) model the data in the original scale, or (b) transform the observed intakes to normality. Recent research by Nusser et al. (1990) on estimating usual intake distributions uses the first approach. In that research, a measurement error model is hypothesized 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 intraindividual 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.

The second approach involves transforming the daily intakes so that the transformed values follow a normal distribution. The National Research Council (1986) recommends this approach and suggests power transformations. However, preliminary investigations using the data from the 1985–1986 Continuing Survey of Food Intakes by Individuals (CSFII) indicate that simple power transformations do not consistently produce transformed data that are normally distributed. In the case of the CSFII data described in Section 3, the three parameters of the model,

$$\text{Normal Score for Daily Intake} = \beta(\text{Daily Intake}^{\gamma} + \xi)\,,$$

were estimated in an attempt to transform the data to normality. The method of fitting described by Lin and Vonesh (1989) was used. Of the dietary components tested (calcium, energy, iron, protein, vitamin A, and vitamin C), only the transformed intake values for calcium and energy were consistent with the hypothesis of normality.

Because the three–parameter power transformation approach was not suitable for these data, a semiparametric transformation for dietary intake data was developed. The first step in the process is to fit a grafted cubic equation to a power of the original data to transform the observed daily intakes to normality. This fitting can be considered a semiparametric version of the Lin and Vonesh (1989) procedure. It is also related to the spline approach to 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 develop a predictor for the transformed usual daily intake for each individual. An inverse transformation is estimated for transforming normal usual intakes back to the original scale. The inverse transformation of the fitted normal distribution defines the distribution of usual intakes. Inferences concerning usual daily intakes can be made in the transformed space or in the original scale. Alternatively, the inverse transformation can be used to produce a set of pseudo usual intakes in the original scale and the pseudo usual intakes can then be used to estimate the distribution of usual intakes.

This article describes the transformation approach to analyzing dietary intake data. The approach was developed with the objective of producing an algorithm suitable for computer implementation and application to a large number of dietary components. To illustrate the approach, data from the 1985–1986 CSFII are analyzed using the proposed methodology.

2. THE TRANSFORMATION APPROACH

2.1 Overview

The transformation approach described below contains three parts. These are transforming the observed intakes to normality, estimating the parameters of the normal distribution of transformed usual intakes, and developing the transformation that carries the normal usual intakes into the original scale. The parameters of the normal usual intake distribution and the transformation of normal usual intakes to the original scale define the distribution of usual intakes in the original scale.

2.2 Transforming the Observed Data to Normality

The first step in the procedure is to develop the transformation to normality. Preliminary analyses established that no simple power transformation was applicable for all dietary components. Therefore, the transformation is specified as a grafted cubic applied to a power of the original observations. The grafted polynomial transformation described below is restricted to have continuous first and second derivatives. The number of join points is chosen so that the transformed values are approximately $N(0, 1)$ random variables.

A power transformation of the observed intakes is used as a starting point for the grafted cubic transformation to normality for two reasons. First, the grafted polynomial transformation required to obtain normal intakes from the power–transformed observed intakes will be much flatter and thus require fewer join points. Second, extrapolation for extreme intake values is likely to be more accurate for a power of the original data.

Let Y_{ij} denote the observed intake of a dietary component for individual i on day j, where $i = 1, 2, ..., n$ individuals and $j = 1, 2, ..., r$ days. Assume that the individuals are independent, and for each individual, daily intakes are independent. Let α be the selected power of the transformation and let Y_{ij}^{α} represent the power transformed data.

Let $\hat{F}$ denote the empirical cumulative distribution function constructed from the nr Y^{α}_{ij} values. By connecting the midpoints of the rises in the steps defined by $\hat{F}$, a continuous piecewise linear estimate $\tilde{F}$ of the true cumulative distribution function F is constructed. More formally, let $Y^{\alpha}_{(1)}, \ldots, Y^{\alpha}_{(m)}$ be the $m \leq nr$ ordered distinct observed daily intake values. Then $\tilde{F}$ is defined by

$$\tilde{F}(Y^{\alpha}) = \begin{cases} 2^{-1}\hat{F}(Y^{\alpha}_{(1)}) & \text{when } Y^{\alpha} < Y^{\alpha}_{(1)} \\ \hat{F}(Y^{\alpha}_{(t)}) + 2^{-1}[\hat{F}(Y^{\alpha}_{(t-1)}) - \hat{F}(Y^{\alpha}_{(t)})] & \text{when } Y^{\alpha}_{(t)} \leq Y^{\alpha} < Y^{\alpha}_{(t+1)} \\ & \quad \text{for } t = 1, \ldots, m \\ \hat{F}(Y^{\alpha}_{(m)}) + 2^{-1}[1 - \hat{F}(Y^{\alpha}_{(m)})] & \text{when } Y^{\alpha} > Y^{\alpha}_{(m)} \end{cases}$$

This approach was chosen because it produces a continuous piecewise linear estimate of F which yields approximately the same mean value as that computed with the empirical cumulative distribution function $\hat{F}$. The approach also accommodates data with sampling weights and repeated observations. Let

$$\tilde{X}_{ij} = \Phi^{-1}(\tilde{F}[Y^{\alpha}_{ij}]) , \tag{1}$$

where $\Phi(\cdot)$ is the normal cumulative distribution function.

The coefficients, β_{ℓ}, of the regression equation

$$\tilde{X}_{ij} = \sum_{\ell=1}^{k} \zeta_{\ell}(Y^{\alpha}_{ij})\beta_{\ell} + e_{ij}$$

are estimated, where $\zeta_{\ell}(Y^{\alpha}_{ij})$, $\ell = 1, 2, \ldots, k$, are regression variables that define a function that is locally cubic, has continuous first and second derivatives, and is linear at

the beginning and end of the range of the data. Let X_{ij} be the transformed variables defined by

$$X_{ij} = \sum_{\ell=1}^{k} \zeta_\ell(Y_{ij}^{\alpha})\hat{\beta}_\ell ,$$

where $\hat{\beta}_\ell$ are the estimated regression coefficients.

Although the X_{ij} are approximately normal variables, they may not exhibit homogeneous intraindividual variance. A test of homogeneity can be constructed by regressing the standard deviations on the means and testing whether the slope from this regression is equal to zero. For the dietary components in the food intake data, the initial transformation produced homogeneous intraindividual variance as well as a normal distribution for the transformed observations.

2.3 Parameter Estimation in Normal Space

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

$$X_{ij} = x_i + u_{ij} ,$$

where

$$x_i \sim \mathrm{NI}(\mu_x, \sigma_x^2) , \qquad u_{ij} \sim \mathrm{NI}(0, \sigma_u^2) , \tag{2}$$

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 . Note that the transformed observed daily intakes X_{ij} from the transformation described in Section 2.2. have $\mu_x = 0$. This model implies that the X_{ij} are $N(0, \sigma_x^2 + \sigma_u^2)$ variates, and that the individual means

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

are independent random variables from a $N(0, \sigma^2_{\bar{X}})$ distribution, where

$$\sigma^2_{\bar{X}} = \sigma^2_x + r^{-1}\sigma^2_u .$$

Estimators for the moments are

$$\hat{\mu}_x = n^{-1} \sum_{i=1}^{n} \bar{X}_{i.} , \qquad \hat{\sigma}^2_{\bar{X}} = (n-1)^{-1} \sum_{i=1}^{n} (\bar{X}_{i.} - \hat{\mu}_x)^2 ,$$

$$\hat{\sigma}^2_u = [n(r-1)]^{-1} \sum_{i=1}^{n} (X_{ij} - \bar{X}_{i.})^2 , \qquad \hat{\sigma}^2_x = \hat{\sigma}^2_{\bar{X}} - r^{-1}\hat{\sigma}^2_u .$$

Let the assumptions of model (2) hold and let μ_x, σ^2_x and σ^2_u be known. Then the best linear unbiased predictor of x_i is

$$\tilde{x}_i = \mu_x + \sigma^{-2}_{\bar{X}}\sigma^2_x(\bar{X}_{i.} - \mu_x) ,$$

where $\mu_x = 0$, and the variance of the prediction error is

$$\mathrm{Var}(\tilde{x}_i - x_i) = \sigma^2_x - \sigma^{-4}_{\bar{X}}\sigma^2_x .$$

The unconditional variance of $\tilde{x}_i$ is

$$\mathrm{Var}(\tilde{x}_i) = \sigma^{-4}_{\bar{X}}\sigma^2_x . \qquad (3)$$

If the objective is to predict a single value of x_i, then $\tilde{x}_i$ is optimal with respect to mean square error. However, our objective is to obtain a set of values whose distribution is close to that of the true x_i, where true x_i has variance σ_x^2. It is clear from (3) that the variance of $\tilde{x}_i$ is less than σ_x^2. Predicted values of x_i with unconditional variance σ_x^2 can be obtained by using the predictor

$$\ddot{x}_i = \mu_x + \sigma_{\bar{X}}^{-1}\sigma_x(\bar{X}_{i.} - \mu_x) . \tag{4}$$

An analogous adjustment for empirical Bayes estimation was suggested by Louis (1984), given that the objective of prediction is to obtain estimates whose empirical cumulative distribution function is close to the true distribution function.

To implement the procedure of (4), the means calculated from the X_{ij} for each individual i and the estimates of μ_x, σ_x^2 and $\sigma_{\bar{X}}^2$ are inserted into (4) in the appropriate places. The resulting $\ddot{x}_i$ are called normal pseudo usual daily intakes.

2.4 The Transformation for Usual Intakes

An individual's usual intake is the expected value of that individual's daily intakes. That is,

$$y_i = E\{Y_{ij}|i\} ,$$

where y_i is the usual intake for individual i. In the transformed scale, x_i is the expected value of X_{ij} for individual i. Let g denote the transformation taking the original observed intakes to normality; i.e.

$$X_{ij} = g(Y_{ij}) .$$

Because the transformation g is nonlinear, $y_i \neq g^{-1}(x_i)$. Therefore, it is necessary to develop the transformation that carries x_i into y_i. Denote the desired transformation by h. The transformation h is constructed by adding to the inverse of the nonlinear transformation g an approximation for the bias necessary for transforming mean values.

An approximation for the transformation h is developed as follows. Let g^{-1} represent the inverse of g. Using a Taylor series approximation for $g^{-1}(x_i + u_{ij})$,

$$y_i = E\{Y_{ij}|i\}$$

$$= E\{g^{-1}(x_i + u_{ij})|i\}$$

$$\doteq g^{-1}(x_i) + 2^{-1}\frac{\partial^2 g^{-1}(x_i)}{\partial x^2}\sigma_u^2 .$$

To obtain an approximation for the second derivative of g^{-1}, consider a particular $\ddot{x}$ and the three points $[\ddot{x}_i - \sigma_u, g^{-1}(\ddot{x}_i - \sigma_u)]$, $[\ddot{x}_i, g^{-1}(\ddot{x}_i)]$, and $[\ddot{x}_i + \sigma_u, g^{-1}(\ddot{x}_i + \sigma_u)]$. A quadratic can be fit to these three points to furnish a local approximation to $g^{-1}(x)$. Thus, we can write

$$g^{-1}(x) \cong a_i x^2 + b_i x + c_i$$

for x near to $\ddot{x}_i$, where (a_i, b_i, c_i) is such that the quadratic passes through the three points. Furthermore, the second derivative of the approximation is $2a_i$ and the approximate y_i value for $x = \ddot{x}_i$ is

$$\ddot{y}_i = g^{-1}(\ddot{x}_i) + a_i\sigma_u^2$$

$$= 0.5[g^{-1}(\ddot{x}_i - \sigma_u) + g^{-1}(\ddot{x}_i + \sigma_u)] \,. \qquad (5)$$

Note that, given the power transformation and the grafted polynomial transformation, a value of $g^{-1}(x)$ can be obtained for any x through iterative numerical techniques. Thus, an approximate usual intake $\ddot{y}$ can be generated for any $\ddot{x}$ using (5).

The function $h(x)$ can be approximated by fitting a grafted cubic to the $(\ddot{y}_i, \ddot{x}_i)$ pairs in the same way that the function carrying the power transformed data to normality was estimated with the $(Y_{ij}^{\alpha}, X_{ij})$ pairs. This smoothed inverse transformation is called the mean transformation.

3. APPLICATION TO CSFII DATA

The procedures described in Section 2 were applied to a subset of the data from the 1985–1986 Continuing Survey of Food Intakes by Individuals (CSFII) conducted by the Human Nutrition Information Service of the U.S. Department of Agriculture. Daily dietary intakes were collected from women between 19 and 50 years of age and from the preschool children of the women. The design called for daily intakes to be obtained at approximate two–month intervals over the period of one year (April 1985 to March 1986). Data for the first day were collected by personal interview and were based on a 24–hour recall. Data for subsequent days were based on 24–hour recall and were collected by telephone whenever possible. The sample was a multi–stage stratified area probability sample from the 48 coterminous states. The primary sampling units were area segments, and the probabilities of selection of area segments were proportional to the numbers of housing units in the segments as estimated by the Bureau of the Census. The sample was designed to be self weighting. Because of the high rate of nonresponse for the six–day sample, the USDA constructed a four–day data set for analyses. The four days of data 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 data available. Weights were developed to adjust for nonresponse, but the analyses of this paper are constructed on unweighted data.

In this paper, we analyze a subset of the four–day data set containing dietary intakes for women between 23 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. There were 785 women who belonged to this category. 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 included in the analyses are calcium, energy, iron, protein, vitamin A and vitamin C. These components were selected because of their nutritional importance and because of their different distributional behaviors.

Most of the differences in distributional shapes for the different components are associated with the frequency of consumption for a dietary component. Dietary components that are consumed frequently, such as energy, tend to be more symmetrically distributed than those that are consumed sporadically. For example, there is a large variability in the amount of vitamin A in foods and vitamin A has a heavily right skewed intake distribution.

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. There were two interview methods, telephone and personal. Interview sequence refers to the order in which the daily data were obtained for sample individuals. There are four values for interview sequence, first interview, second interview, third interview and fourth interview.

The daily intake data were examined using least squares methods to determine whether weekday, month, interview method, and interview sequence effects were important. Preliminary analyses with weekday, month, interview method and interview sequence effects in the model indicated that month and interview sequence are confounded to a large degree. This is because the first interview was conducted at nearly the same point in time for all individuals. Hence, the month effects were deleted from the model, and for subsequent analyses, a model containing weekday, interview method and interview sequence as additive classification variables was used.

Interview method was not significant for any dietary component. Weekday effects were significant for energy ($p < 0.001$) and protein ($p < 0.01$) intakes. Contrasts indicated that the effect was primarily due to higher consumption on weekends for both dietary components. Weekday effects were not significant for calcium, iron or vitamin C. Sequence effects (confounded with month effects) were significant at the $\alpha = 0.001$ level for calcium, energy, iron and protein intakes. For all dietary components, a large proportion of the sequence variation was accounted for by a contrast of first interview day versus the intake for the other three days (92–99% of the sequence variation for calcium, energy, iron and protein and 78% for vitamin C). The mean intakes for the first interviews, conducted April through June, were consistently higher than mean intakes in other months.

Because of these results, we used data adjusted for weekday and interview sequence effects in the subsequent analyses. A ratio adjustment was used to insure that all adjusted intake values are nonnegative. The observed intake values were regressed on indicator variables representing the days of the week and interview sequence. The data adjusted for weekday and interview sequence are

$$\overset{*}{Y}_{ij} = \hat{Y}_{0ij}^{-1} \bar{Y}_{1.} Y_{0ij},$$

where Y_{0ij} is the original observed intake of individual i on day j, $\bar{Y}_{1.}$ is the mean of the original observed intakes for the first interview day, $\hat{Y}_{0ij}$ is the predicted intake from the regression, and $\overset{*}{Y}_{ij}$ is the ratio adjusted intake for individual i on day j.

To meet the assumptions of the basic model that the distributions are homogeneous across days, the ratio adjusted data were further modified using the following procedure. The intakes observed on the first survey day were ranked. For each remaining day, the data were also ranked. The data for each of the last three days were replaced by the first–day intake of the equivalent rank. The effect of this procedure is to produce smoothed data that are identically distributed on all four days. The first intake day was taken to be the "standard" because research workers in the field believe it to be the most accurate. The adjusted intakes are hereafter referred to as the observed daily intakes.

The among– and within–individual standard deviations for the observed intakes are presented in Table 1. These statistics indicate that there is considerable intraindividual variability relative to interindividual variability. The ratios of intra– to interindividual variances are similar to those noted for comparable data in National Research Council (1986). The skewness coefficient for the distribution indicates that for most components, an assumption of normality is unreasonable. In addition, plots shown in Figures 1 and 2 of

Table 1. Sample moments for data in the original scale.

Dietary Component	Mean	Among–Individual s.d.	Within–Individual s.d.	Skewness
Calcium	579.14	223.64	297.55	1.31
Energy	1492.97	382.73	507.25	0.73
Iron × 100	999.10	288.98	463.65	2.28
Protein × 10	595.35	140.43	238.78	1.17
Vitamin A ÷ 10	498.30	255.43	733.04	6.37
Vitamin C × 10	751.44	386.71	584.96	1.57

intraindividual standard deviations versus individual means for energy and vitamin C reveal that the intraindividual variances for these dietary components are related to the individual means.

The observed intakes for each dietary component were transformed to normality using the procedure described in Section 2.2. One one hundredth of the mean of the component was added to each observation before performing the power transformation. This was done because the components vitamin C and vitamin A had some daily observed values very close to zero and because the derivative of the power transformation is infinite at zero. Let Y_{ij} denote the observed intakes increased by one one hundredth of the sample mean.

The value of α was computed by using the 0.10, 0.50, and 0.90 values of the empirical distribution function. Let these vector of values be (w_1, w_2, w_3), and let the 0.10, 0.50 and 0.90 values of the standard normal distribution be (z_1, z_2, z_3). Then the value of $(\alpha, \beta_0, \beta_1)$ was chosen such that

$$\sum_{i=1}^{3} (z_i - \beta_0 - \beta_1 w_i^{\alpha})^2$$

is a minimum where the minimum is over the grid of values for α, $[1, (1.5)^{-1}, (2.0)^{-1}, \ldots, (10)^{-1}]$. This relatively simple estimation procedure was chosen so that it could be implemented automatically for future analyses of dietary components. Also, there is the second round to the transformation associated with the grafted polynomial. The inverses of the powers of the first round transformation are given in Table 2. The largest power of $(2.5)^{-1}$ was chosen for energy and the power for the two vitamins is the boundary power of $(10)^{-1}$.

The $\zeta_\ell(Y_{ij}^{\alpha})$ for the grafted polynomial were created in the following manner. Let n_p be the largest integer less than $(p-1)^{-1}(n-4)$. Let n_1 be the largest integer less

Table 2. Statistics for the transformations

Dietary Component	Inverse of power	Number of parameters	Original Anderson–Darling[a]	Mean Anderson–Darling[a]	t for standard deviation[b]
Calcium	4.5	3	0.24	0.25	–1.46
Energy	2.5	3	0.36	0.37	0.66
Iron	5.5	4	0.21	0.22	–0.03
Protein	3.0	4	0.20	0.63	0.66
Vitamin A	10.0	5	0.26	0.45	–0.35
Vitamin C	10.0	6	0.57	0.34	–0.88

(a) Reject at 10% level if Anderson–Darling statistic greater than 0.68.

(b) Reject null hypothesis of homogeneous variance at 5% level if $|t| > 1.96$.

than $2 + 0.5[n - (p-1)n_p]$, and let $n_p - n_1 - (p-1)n_p$. Let $A_1, A_2, \ldots, A_p$ be a set of points such that n_1 values of Y_{ij}^{α} are less than A_1, n_{p+1} values of Y_{ij}^{α} are greater than A_{p+1} and n_p values of Y_{ij}^{α} fall between A_i and A_{i+1} for $i = 1, 2, \ldots, p$. The $\zeta_\ell(Y_{ij}^{\alpha})$ are defined by

$$\zeta_1(Y_{ij}^{\alpha}) = 1 ,$$

$$\zeta_2(Y_{ij}^{\alpha}) = Y_{ij}^{\alpha} - (4n)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{4} Y_{ij}^{\alpha} ,$$

and

$$\zeta_\ell(Y_{ij}^{\alpha}) = G_\ell(Y_{ij}^{\alpha})$$

$$+ (A_{p-1} - A_p)^{-1}[(A_p - A_{\ell-2})G_{p+1}(Y_{ij}^{\alpha}) - (A_{p-1} - A_{\ell-2})G_{p+2}(Y_{ij}^{\alpha})] ,$$

for $\ell = 3, 4, \ldots, p$, where

$$G_\ell(Y_{ij}^\alpha) = 0 \qquad Y_{ij}^\alpha \le A_{\ell-2},$$

$$= (Y_{ij}^\alpha - A_{\ell-2})^3 \qquad Y_{ij}^\alpha > A_{\ell-2},$$

for $\ell = 3, 4, \ldots, p + 2$. Each of the $\zeta_\ell(Y_{ij}^\alpha)$ is a function that is: a) linear for $Y_{ij}^\alpha \le A_1$, b) linear for $Y_{ij}^\alpha \ge A_p$, and c) continuous with continuous first and second derivatives. Therefore, a linear combination of the functions will have the same properties. The fitted grafted polynomial function was constrained to be monotone and continuous with continuous first and second derivatives.

The number of join points for the grafted cubic was chosen so that the Anderson–Darling test for normality was nonsignificant at the 10 percent level when applied to the transformed data. A minimum of three parameters was estimated for each component. The statistics for the transformation are given in Table 2. The minimum of three parameters was judged satisfactory for calcium and energy. For iron and protein, four parameters were included in the model. The model for vitamin A contained five parameters and the model for vitamin C contained six parameters.

Figure 3 contains a plot of the normal scores against the $(5.5)^{-1}$ power of the iron observations. It is clear that no simple power transformation would be adequate to transform the plot into a straight line. The estimated grafted cubic is the smooth line in the plot. The vertical dashed lines are the join points for the grafted polynomial. The fitted function is linear for the segments outside the exterior dashed lines. The dashed lines are spaced so that there is an equal number of observations in each segment. As a result, the middle segment is much narrower than the two adjacent segments.

A plot of the transformation from the original scale observed intakes to the normal observed intakes is presented in Figure 4 for vitamin C. The transformation for vitamin C differs considerably from a power function.

As an additional check on the transformation, the Anderson–Darling statistic was computed for the means of the four transformed observations. These statistics are given in the column headed "Mean Anderson–Darling" in Table 2. In no case is the statistic significant at the ten percent level.

To examine the intraindividual variances for the transformed data, the hypothesis of a zero slope for the regression of the intraindividual standard deviations on the individual means was tested. The results from these tests are presented in the last column of Table 2. In all cases, the hypothesis of zero slope is accepted. Plots of the intraindividual standard deviations versus the individual means for the transformed data were also constructed. There were no obvious deviations from homogeneous intraindividual variances in the plots. Figure 6 contains the plot for vitamin C.

The within and between variances for the transformed data are given in Table 3. In all cases, the sum of the within–individual and between–individual variances is close to one

Table 3. Sample moments for data in the normal scale.

Dietary Component	Among–Individuals Variance $\hat{\sigma}_x^2$	Within–Individual Variance $\hat{\sigma}_u^2$
Calcium	0.358	0.639
Energy	0.362	0.634
Iron	0.314	0.684
Protein	0.273	0.725
Vitamin A	0.226	0.767
Vitamin C	0.308	0.683

because the transformed data have mean zero and variance one. The within variance exceeds the among variance for all dietary components. The ratio of within to among is smallest for energy with a value of 1.75 and is largest for vitamin A with a ratio of 3.39. The ratios of within to among of Table 3 are larger than the corresponding ratios computed from the standard deviations in normal scale of Table 1. This is because in the original scale, the data are skewed and the individual standard deviations are positively correlated with the individual means.

The mean transformation was computed in three steps. First, an individual normal usual intake was computed for each individual using the individual means and formula (4). Then, an individual usual intake in the original scale was computed for each normal usual intake using equation (5). Finally, a function was fit to the $(\ddot{x}_i, \ddot{y}_i)$ pairs. The function was of the same form as the original transformation. That is, the power of $\ddot{y}_i$, the number of parameters estimated for the grafted polynomial, and the join points were the same as for the original function.

Table 4 lists the sample mean, variance and skewness coefficient for the pseudo usual intakes where the pseudo usual intakes are in the original scale. The pseudo usual intakes are those defined by equation (5). Comparison of the statistics for pseudo usual intakes with the same statistics for the distribution of individual means (Table 5) indicates that the distribution of four–day means is not an appropriate estimate of the usual intake distribution. For all dietary components, the standard deviation and skewness coefficient is larger for the mean distribution than for the estimated usual intake distribution.

The estimated densities of usual intakes for energy and vitamin C are the solid lines in Figures 7 and 8. These densities were constructed by taking the derivative of the h transformation and multiplying this derivative by the normal ordinate for the usual intake density of the component in the normal scale. Thus, the density of usual intakes is

Table 4. Sample moments for the pseudo usual intakes.

Dietary Component	Mean	Standard Deviation	Skewness
Calcium	580.00	220.10	0.93
Energy	1493.77	383.10	0.40
Iron × 100	990.37	286.33	1.03
Protein × 10	592.88	141.14	0.45
Vitamin A ÷ 10	464.71	256.27	1.56
Vitamin C × 10	765.66	368.64	0.80

Table 5. Sample moments for four–day means.

Dietary Component	Mean	Standard Deviation	Skewness
Calcium	579.14	268.60	1.13
Energy	1492.97	459.14	0.50
Iron × 100	999.10	370.48	1.40
Protein × 10	595.35	184.32	0.73
Vitamin A ÷ 10	498.30	446.74	3.66
Vitamin C × 10	751.44	484.86	1.13

$$f_y(y) = \phi_x[h^{-1}(y)]\frac{\partial h^{-1}(y)}{\partial y},$$

where $\phi_x(\,)$ is the distribution of usual intakes in normal space.

Also in the figures are the estimated density for one–day intakes identified by the short dashed lines and the estimated density of the four–day means identified by long dashed lines. The densities for four–day means were estimated by estimating the transformation to normality in exactly the same way as described for the one–day intakes. The one–day intake distribution for energy is only mildly skewed. The energy usual intake

distribution is even less skewed. The estimated density of four–day means for energy has properties falling between those of the one–day density and the usual intake density.

The one–day intake distribution for vitamin C is very skewed. While the four–day distribution is less skewed, it is still of a nonnormal shape. The distribution of usual intakes for vitamin C is less skewed than the four–day distribution but it is definitely skewed.

Dietary intake data are often used to make inferences about the nutritional status of the population. Of particular interest are estimates on the prevalence of a dietary inadequacy, that is, the proportion of the population with usual nutrient intake below the appropriate standard. Therefore, in deriving an estimate of usual intake distributions, it is the lower and upper percentiles of the distribution, rather than its first two moments, which are of interest.

To illustrate the differences among the distributions, an arbitrary standard of 800 kcal was adopted for energy. Under the distribution of one–day intakes, 12.8% of the observations are below the standard. These percentages are 6.1 and 2.1 for the four–day mean and usual intake distributions, respectively. Thus, a large error would be made if the distribution of four–day means were used as an approximation to the distribution of usual intakes.

Table 6 contains some estimated percentiles for the dietary components. The percentiles were computed from the estimated mean 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 intake in normal space are zero and 0.3082, respectively. Therefore, the estimated 95% point in normal space is $0.5552 \times 1.645 = 0.9132$. Using the estimated h–transformation, the 95% point of the usual intake distribution in original space is 1,422.

The numbers in parentheses are approximate standard errors calculated using Taylor approximations. In normal space, the estimated quantile is

Table 6. Estimated percentiles for usual intakes

Component	Percentile						
	0.01	0.05	0.10	0.50	0.90	0.95	0.99
Calcium	201	278	327	550	867	977	1211
	(10)	(9)	(9)	(9)	(19)	(24)	(36)
Energy	719	911	1023	1469	1991	2153	2476
	(25)	(22)	(20)	(16)	(28)	(34)	(47)
Iron × 100	466	589	661	952	1363	1518	1871
	(17)	(15)	(14)	(12)	(27)	(36)	(60)
Protein × 10	306	378	419	581	781	848	985
	(11)	(9)	(8)	(6)	(13)	(16)	(24)
Vitamin A ÷ 10	119	167	202	403	803	976	1406
	(7)	(8)	(8)	(11)	(34)	(49)	(92)
Vitamin C × 10	162	253	327	726	1236	1422	1847
	(10)	(14)	(16)	(15)	(33)	(44)	(73)

$$Q_x(p) = \hat{\mu}_x + \Phi^{-1}(p)\hat{\sigma}_x ,$$

where $\hat{Q}_x(p)$ is the estimated quantile in normal space, $\hat{\sigma}_x^2$ is the estimated variance of usual intakes in normal space, $\Phi(\ )$ is the standard normal cumulative distribution function and $\hat{\mu}_x$ is the estimated mean of usual intakes in normal space. The estimated variance of $\hat{\sigma}_x^2$ can be computed using the estimated variances of the analysis of variance. Because $\hat{\mu}_x$ is independent of $\hat{\sigma}_x^2$, the estimated variance of $\hat{Q}_x(p)$ is

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

For the estimated five percent point of the vitamin C usual intake in normal space,

$$\hat{V}\{\hat{Q}_x(p)\} = (1.645)^2(0.004947) + 0.006101 = 0.001949 ,$$

where

$$\hat{V}\{\hat{\mu}_x\} = n^{-1}(\hat{\sigma}_x^2 + 0.25\hat{\sigma}_u^2) ,$$

$$\hat{V}\{\hat{\sigma}_x^2\} = (0.25)^2 2n^{-1}[(\hat{\sigma}_u^2 + 4\hat{\sigma}_x^2)^2 + 3^{-1}\hat{\sigma}_u^4] ,$$

and we have used simple random sampling variance formulas. The estimated variance of the estimated quantile in original space is

$$\hat{V}\{\hat{Q}_y(p)\} = \left[\frac{\partial h(x)}{\partial x}\right]^2 \hat{V}\{\hat{Q}_x(p)\} ,$$

where the h–transformation is treated as fixed. For the five percent point for vitamin C we have

$$\hat{V}\{\hat{Q}_y(p)\} = (31.55)^2(0.001949) = 1.940 .$$

4. FUTURE RESEARCH

Several extensions of the methodology are being developed. Since most data are based on complex sample surveys, it is important that the method be extended to include data with weights other than n^{-1}. This modification can be incorporated by including a weight term in the estimate of the empirical distribution function. See Francisco and Fuller (1991). The calculation of standard errors of estimates based on complex surveys also requires additional work. Replication variance estimation methods are being investigated. Many dietary intake surveys are based on observations from adjacent days.

Such surveys require estimators that account for the correlation structure among observed daily intakes for an individual. Also, an extension of the method to multivariate data is under study.

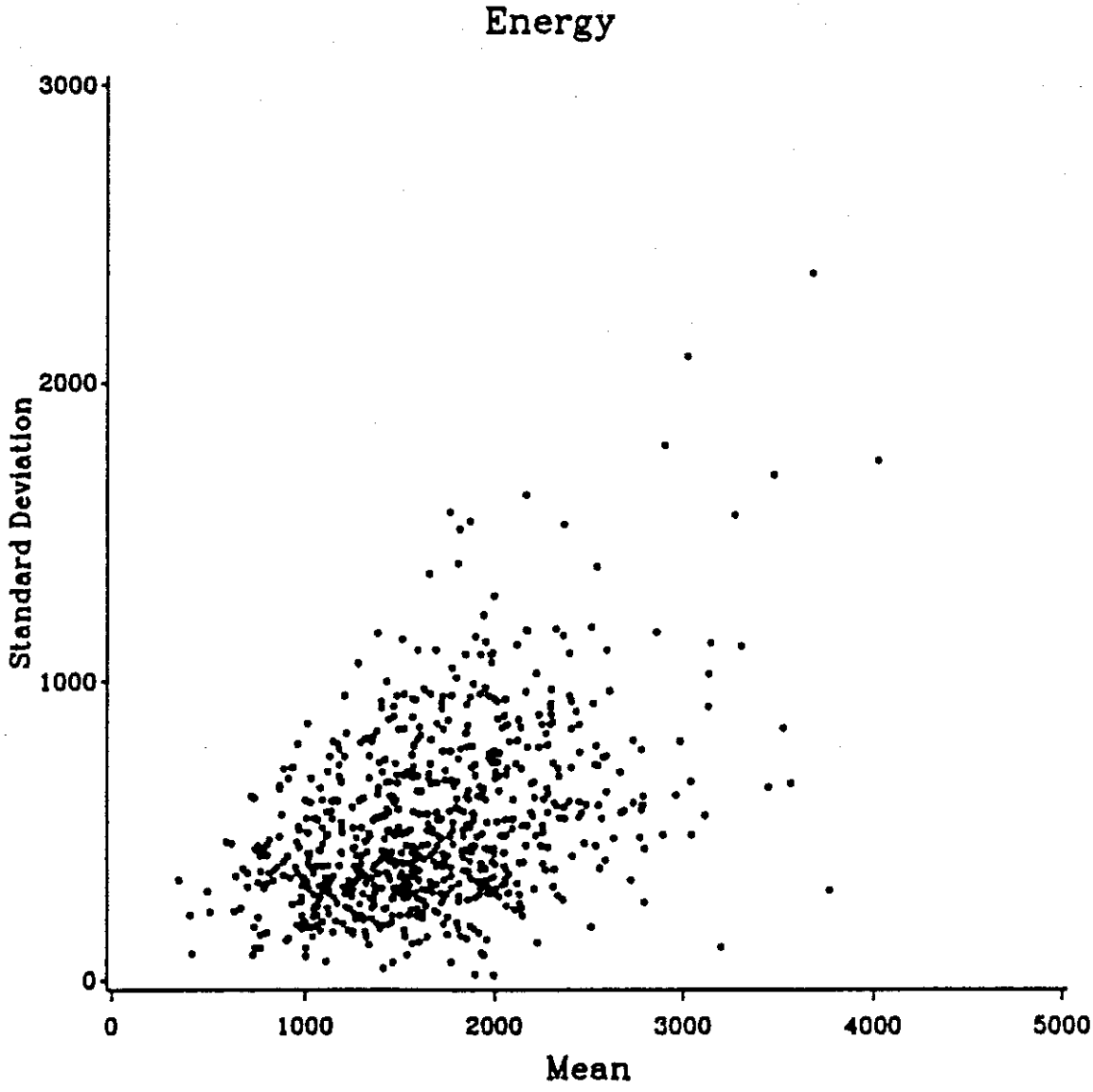

Figure 1. Plot of individual standard deviation of intakes against individual mean intakes for energy, computed from data in the original scale.

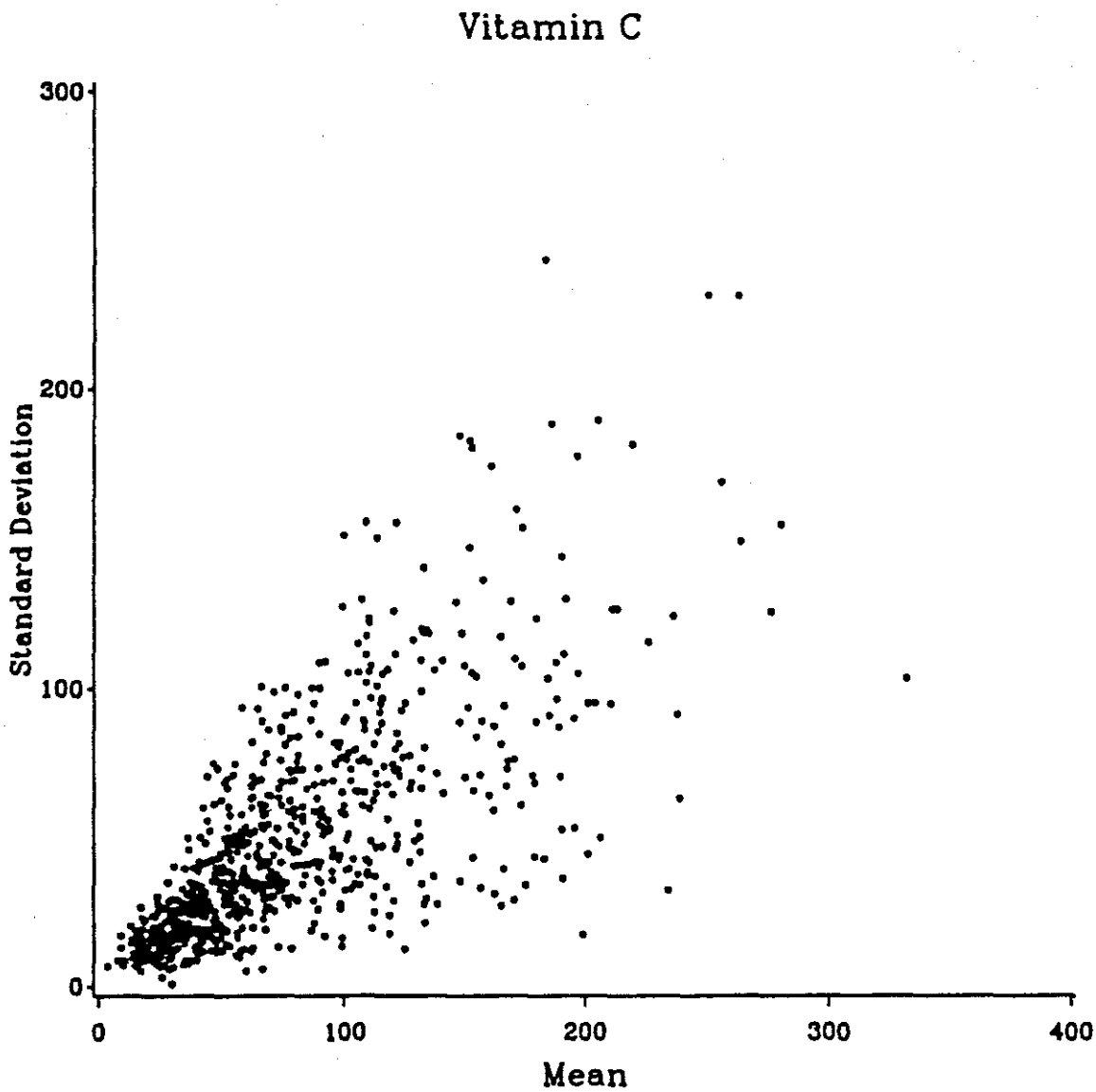

Figure 2. Plot of individual standard deviation of intakes against individual mean intakes for vitamin C, computed from data in the original scale.

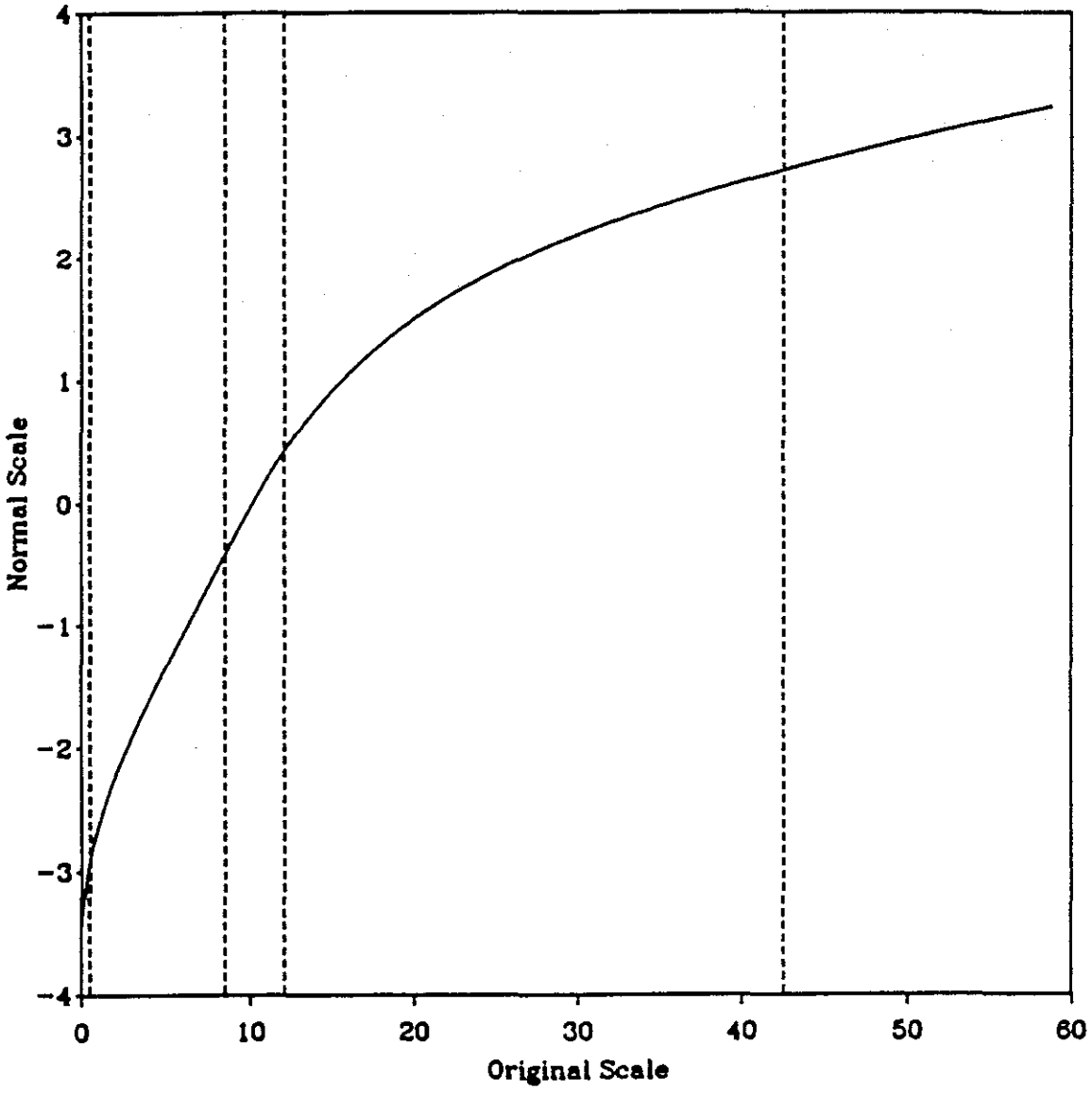

Figure 3. Plot of grafted polynomial for iron.

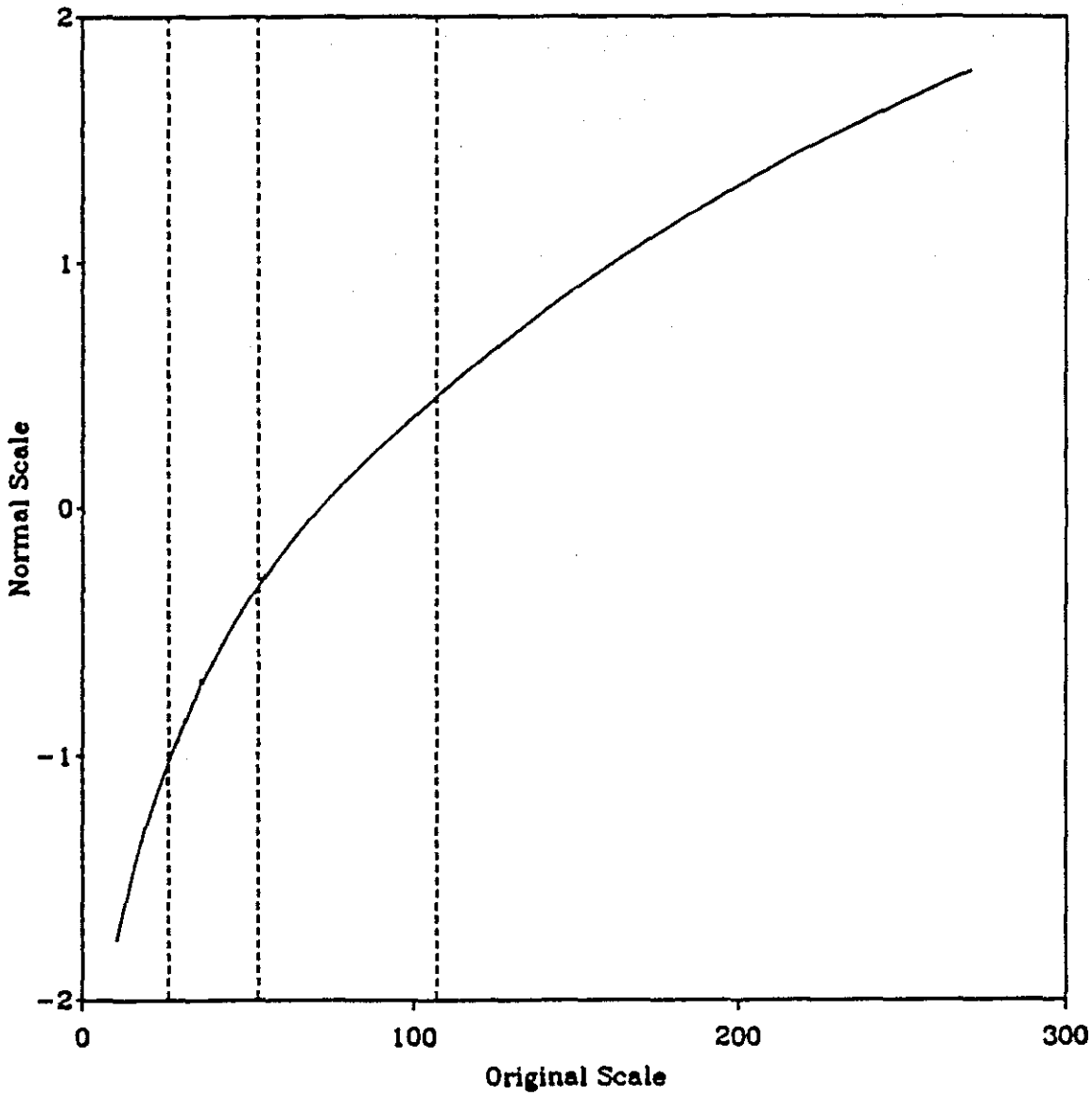

Figure 4. Plot of the g–transformation for vitamin C.

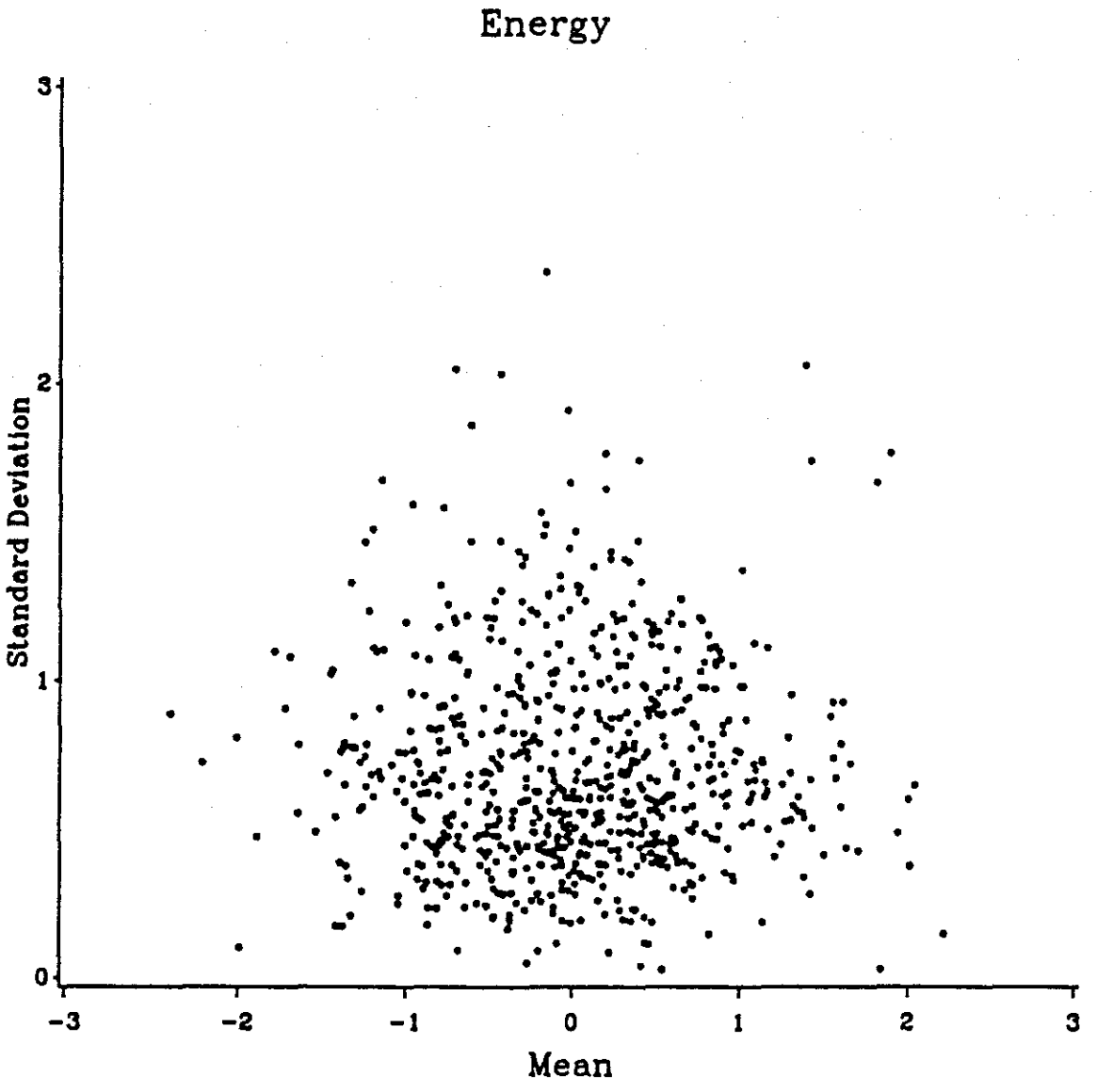

Figure 5. Plot of individual standard deviations against individual mean intake of energy, computed from the transformed data.

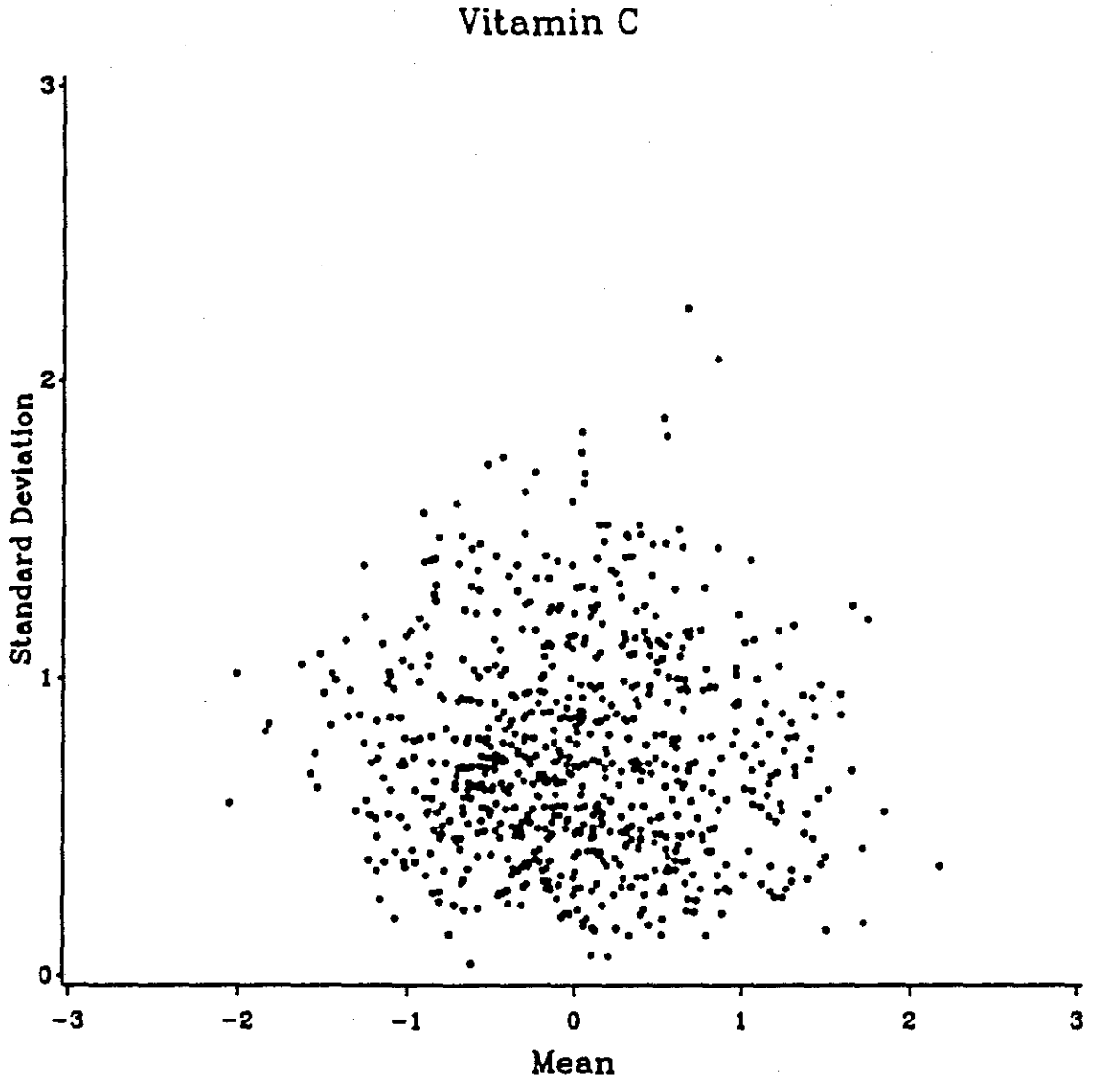

Figure 6. Plot of individual standard deviations against individual mean intakes of vitamin C, computed from the transformed data.

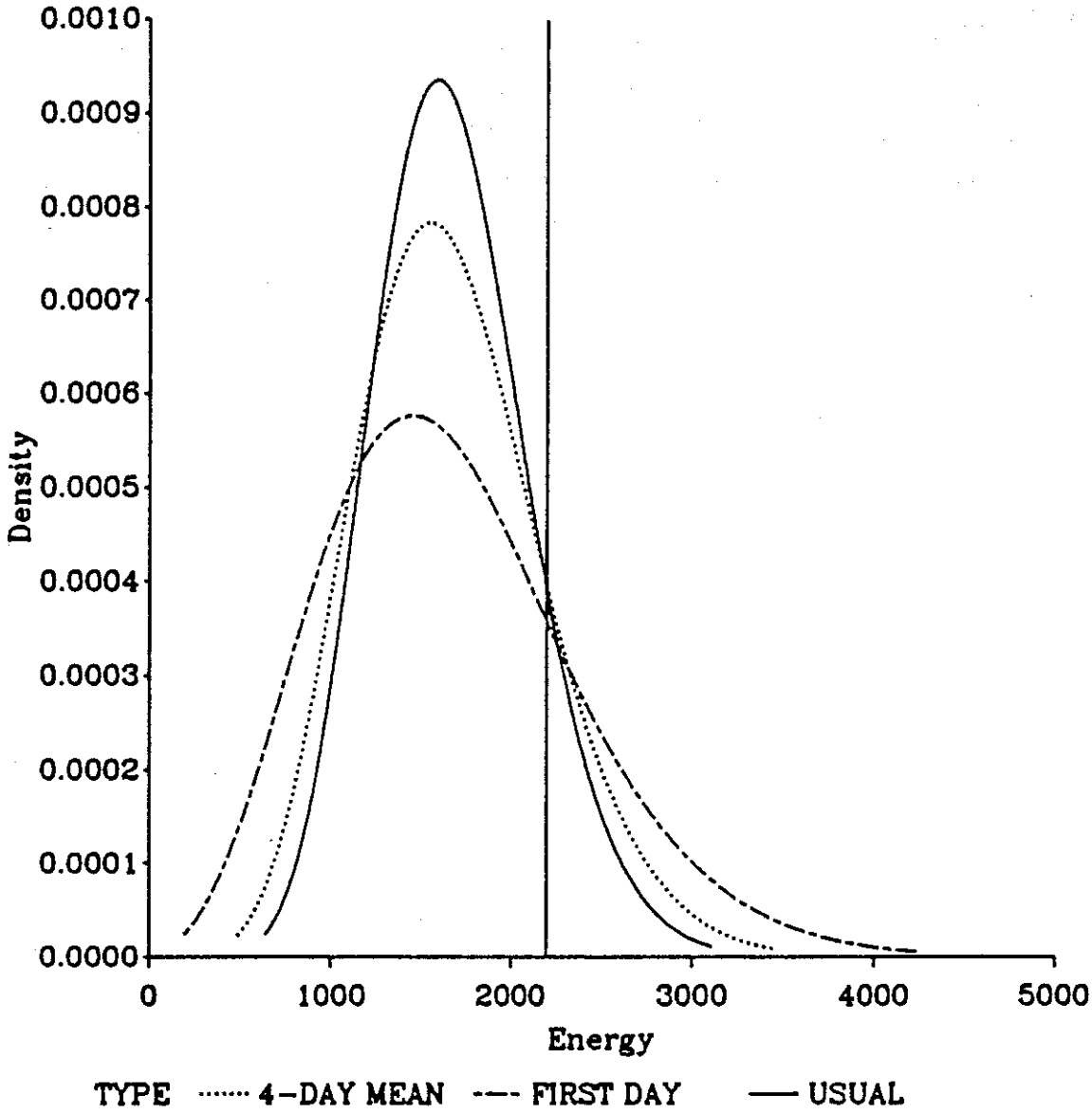

Figure 7. Empirical densities of usual intakes, four–day means, and one–day intakes for energy.

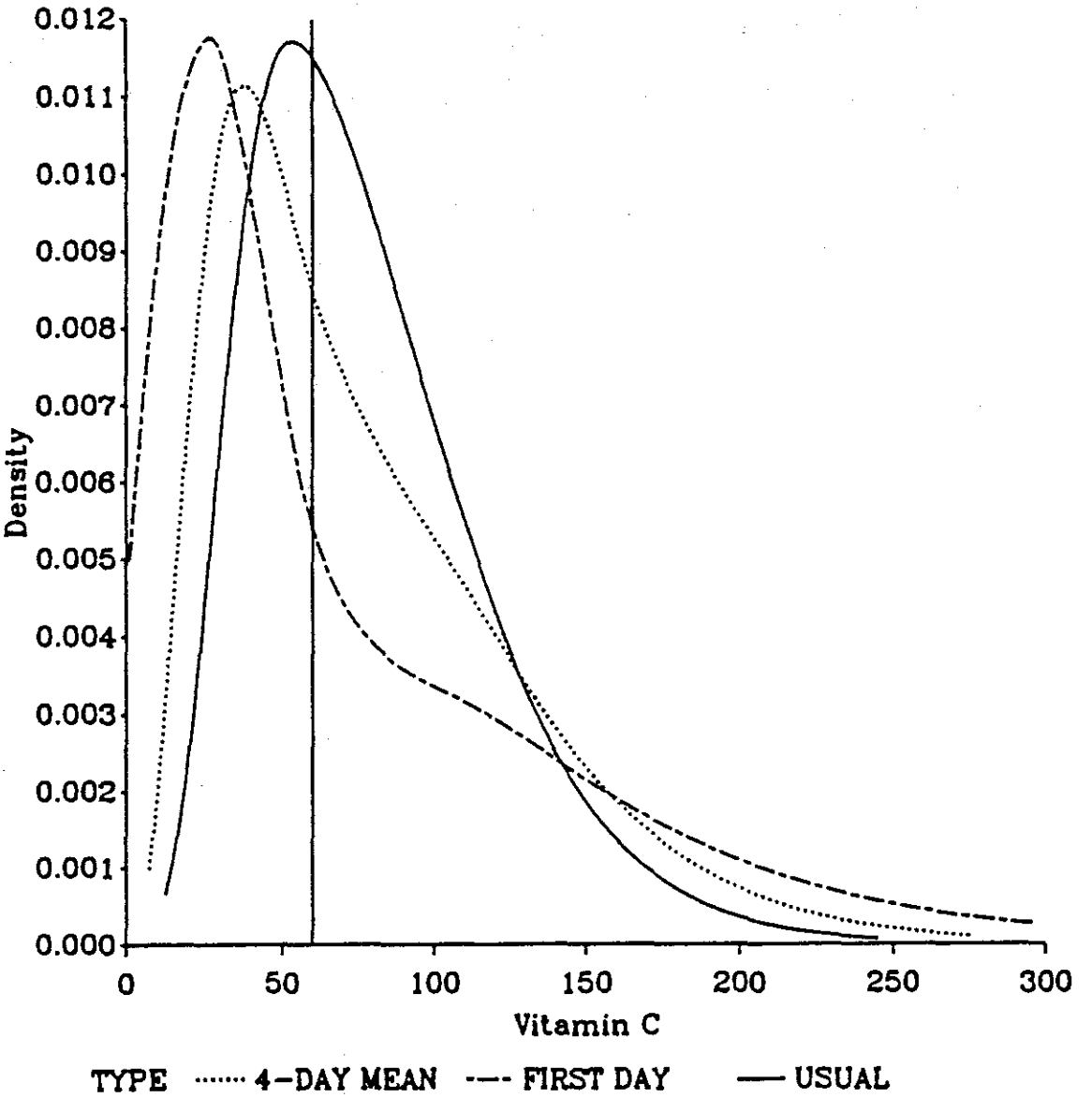

Figure 8. Empirical densities of usual intakes, four–day means and one–day intakes for vitamin C.

REFERENCES

Francisco, C. A., and W. A. Fuller (1991), "Quantile estimation with a complex survey design," The Annals of Statistics 19, 454–469.

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.

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.

Lörstad, H. H. (1971), "Recommended intake and its relation to nutrient deficiency," Nutrition Newsletter 9, 08–31.

Louis, T. A. (1984), "Estimating a population of parameter values using Bayes and empirical Bayes methods," Journal of the American Statistical Association 79, 393–398l.

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.

Stephens, M. A. (1974), "EDF statistics for goodness of fit and some comparisons," Journal of the American Statistical Association 69, 347–730.

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

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