Estimation of the Distribution of Usual Intakes for Selected Dietary Components

by George E. Battese,
Sarah M. Nusser, and Wayne A. Fuller

Technical Report 88-TR5

October 1988

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

George E. Battese is senior lecturer, Department of Econometrics, University of New England, Armedale, NSW, Australia, and was visiting professor of Statistics at Iowa State University at the time of this research. Sarah M. Nusser and Wayne A. Fuller are graduate assistant and distinguished professor, respectively, Department of Statistics, Iowa State University.

This research was partly supported by the Human Nutrition Information Service, United States Department of Agriculture, under USDA, Food and Nutrition Service Cooperative Agreement Number 58-3198-6-60.

TABLE OF CONTENTS

TABLES

FIGURES

ESTIMATION OF THE DISTRIBUTION OF USUAL INTAKES FOR SELECTED DIETARY COMPONENTS

by

G. E. Battese, S. M. Nusser, and W. A. Fuller
Iowa State University

INTRODUCTION

The U.S. Department of Agriculture (USDA) has been responsible for conducting periodic surveys to estimate food consumption patterns of households and/or individuals in the United States for over 50 years. Data from these surveys have had a significant impact on the formulation of food-assistance programs, on consumer education and on food regulatory activities.

In recent years, there has been interest in estimating the proportion of the population that has insufficient intake or excessive intake of certain dietary components. Different approaches have been suggested for the estimation of this proportion. In all approaches, it is necessary to analyze data on dietary intakes for a sample of individuals. Also, all approaches recognize that an individual who has a low intake of a given dietary component on one day is not necessarily deficient (or at risk of being deficient) so far as that dietary component is concerned. It is low intake over a sufficiently long period of time that produces a dietary deficiency. A dietary deficiency exists when the "usual" (i.e., normal or long-run average) intake of the dietary component is less than the appropriate dietary standard.

In this paper, our focus of attention is the usual intake of selected dietary components. We consider the estimation of the cumulative distribution function of usual intake using a sample of individuals for whom several daily observations on dietary intake have been obtained. The data on selected dietary components are from the 1985-86 Continuing Survey of Food Intakes by Individuals (CSFII).

THE CSFII DATA

During 1977-78, the USDA conducted its latest Nationwide Food Consumption Survey, in which food intakes of sample individuals (at home and away from home) were ascertained for three consecutive days. On the day of the interview, a sample respondent was asked to report his or her food intake during the previous day and then to record intakes for the day of the interview and for the day following the interview. After this survey was conducted, it was recommended by the National Research Council (1986) that food intakes be obtained for non-consecutive days over an extended period of time so that the normal consumption patterns of individuals may be better estimated.

In 1985 the USDA conducted a Continuing Survey of Food Intakes by Individuals. The survey collected daily dietary intakes for women between 19 and 50 years of age and their pre-school children. Intakes were 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. Data for subsequent days were collected by telephone whenever possible. The sample was a multi-stage stratified area probability sample from the 48 conterminous states. The primary sampling units were area segments, and the probability of

selection of area segments was proportional to the number of housing units in the segments as reported by the Bureau of the Census. Of the 1,459 women who agreed to participate and provided the first one-day dietary intakes, 71 percent completed at least four days, 63 percent completed at least five days, and 47 percent completed all six days.

In this paper we analyze a data set containing four days of dietary intakes for 785 women aged between 23 and 50 years who were responsible for meal planning within the household and who were not pregnant or lactating during the survey period. The four days of data consisted of the first one-day 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. Empirical results are presented for intakes of the five dietary components: calcium, energy, iron, protein and vitamin C.

PRELIMINARY ANALYSES

Since the survey data are for different days of the week and different months of the year, the effects of these as sources of variability are investigated. Let Y_{ij} represent the intake of a given dietary component for individual i for the j-th reporting day. We consider the linear regression model

$$Y_{ij} = \alpha + \beta_k + \gamma_m + \epsilon_{ij} , \qquad (1)$$

where

$\beta_k = 1$ if the (i,j)-th observation was collected in the k-th month, k=1, 2, ..., 12;

$= 0$ otherwise;

$\gamma_m = 1$ if the (i,j)-th observation was collected on the m-th day of the week, m=1, 2, ..., 7 ;

$= 0$ otherwise;

and ϵ_{ij} is the error in the regression equation.

In this model, month effects are significant at the five-percent level for all five dietary components. Weekday effects are significant for energy and protein, but not for calcium, iron or vitamin C.

Using data not adjusted for month or weekday effects, the average intake for the first-day of interview was significantly greater than the average intakes for the remaining three days for the five dietary components. The average intakes for the second, third and fourth days are not significantly different. There are no significant differences among the average intakes with respect to interview sequence, after accounting for the month and weekday effects. This conditional result is not particularly meaningful, however, because month effects and time-of-interview effects are confounded to a considerable extent in these four-day data. We conclude that interview sequence effects or month effects or both are present in the data.

The variances of intakes within individuals and among individuals (i.e., intra-individual and inter-individual variances) are estimated from the simple analysis of variance described in Table 1. In that table,

$$\bar{Y}_{i.} = 4^{-1} \sum_{j=1}^{4} Y_{ij} \qquad (2)$$

Table 1. Analysis of variance for observed individual intakes

Source	d.f.	S.S.	EMS
Individuals	784	$\sum_{i=1}^{n} 4(\bar{Y}_{i.} - \bar{Y}_{..})^2$	$\sigma_w^2 + 4\sigma_b^2$
Days/individual	2355	$\sum_{i=1}^{n} \sum_{j=1}^{4} (Y_{ij} - \bar{Y}_{i.})^2$	σ_w^2

and

$$\bar{Y}_{..} = n^{-1} \sum_{i=1}^{n} \bar{Y}_{i.} \qquad (3)$$

are the average of the four daily intakes for the i-th individual and the average of all observations on all sample individuals, respectively; σ_w^2 is the intra-individual variance; σ_b^2 is the inter-individual variance; and $n=785$. The ratios of the estimated intra-individual and inter-individual variances are presented in the first column of Table 2. The estimates from the four-day CSFII data are of similar magnitude to those reported by Sempos et al. (1985) for studies on adult women in two different years. The averages of the two ratio estimates reported by Sempos et al. (1985) are given in the last column of Table 2.

Other summary statistics for the daily intake of the five dietary components are presented in Table 3. Also presented in Table 3 are the Recommended Dietary Allowances (RDAs) for the United States, as reported in the latest RDA publication (see National Research Council 1980). The estimates for the mean daily intakes are the simple averages over all sample individuals. The estimates for the standard deviation, skewness

Table 2. Estimates for the ratio of intra-individual to inter-individual variances of daily intakes of dietary components

Dietary component	Estimates for σ_w^2/σ_b^2	
	This study	Sempos et al. (1985)
Calcium	1.8	1.1
Energy	2.0	1.6
Iron	2.5	2.6
Protein	2.9	2.1
Vitamin C	2.4	2.4

Table 3. Summary statistics for the distribution of the four-day average intakes of dietary components

Dietary component	RDA	Mean	Standard deviation	Skewness	Kurtosis
Calcium (mg)	800	579	281	1.16	2.41
Energy (kcal)	2,000[a]	1,493	487	0.61	0.82
Iron (mg)	18	10.0	3.68	1.24	3.51
Protein (g)	44	59.6	19.6	0.77	2.37
Vitamin C (mg)	60	75.2	49.6	1.37	2.62

[a]The value for energy is the mean energy requirement as stated in the latest RDA report [see National Research Council (1980, p. 23)].

and kurtosis are obtained from the daily intakes adjusted for month and weekday effects, according to the specifications of the analysis-of-variance model (1). All sample statistics were obtained using the SAS software (see SAS 1985). The standard deviation, skewness and kurtosis

statistics of Table 3 are

$$\hat{s}_{\epsilon} \equiv [(n-1)^{-1} \sum_{i=1}^{n} \hat{\epsilon}_i^2]^{1/2},$$

$$\frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} (\hat{\epsilon}_i / \hat{s}_{\epsilon})^3 , \qquad (4)$$

and

$$\frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} (\hat{\epsilon}_i / \hat{s}_{\epsilon})^4 - \frac{3(n-1)^2}{(n-2)(n-3)} ,$$

respectively, where $\hat{\epsilon}_i$ is the average of the four estimated errors of model (1) for the i-th individual.

The standard deviation is a measure of the dispersion of the intakes about the mean intake for the population of individuals. The skewness statistic measures the lack of symmetry of the distribution of intakes about the mean. A symmetrical distribution has a skewness measure equal to zero. If small values of intake are near the mean while large values are much greater than the mean, then the larger values have a greater contribution to the third moment and result in positive skewness. The kurtosis statistic measures the extent to which values tend to occur distant from the mean. The kurtosis measure for the normal distribution is zero. A positive kurtosis measure indicates that a distribution tends to have "fatter tails" than the normal distribution.

In Table 3, the estimated mean intakes of calcium and iron are less than the corresponding RDAs by about 0.8 and 2.2 standard deviations, respectively, whereas for protein and vitamin C the mean intakes are

greater than the corresponding RDAs by about 0.8 and 0.3 standard deviations, respectively. The estimated mean intake of energy is about one standard deviation less than the mean of energy requirements of 2,000 kcal from the latest RDA report.

The skewness estimates presented in Table 3 indicate that the distributions of daily intakes are skewed to the right (positively skewed) for all five dietary components. The energy and protein intakes are the least skewed. The kurtosis values indicate that the distributions of the average reported intakes for the five dietary components tend to have fatter tails than the normal distribution.

Of the five dietary components we consider in this analysis, three are discussed in Appendix A of the National Research Council (1986) report. The three variables common to that appendix and our study are iron, protein and vitamin C. The National Research Council (1986, p.114) reports that the mean intakes for females in the Nationwide Food Consumption Survey (NFCS) conducted in 1977-78 were 10.8 mg, 65.6 g and 72.6 mg, for iron, protein and vitamin C, respectively. These values were based on three-day intakes over consecutive days for about 2,400 women. The estimated mean intakes for protein for our CSFII data are significantly less than those based on the NFCS data. The estimated standard deviations of all three dietary components in the CSFII data are significantly less than those for the NFCS data.

The characteristics of the distribution of daily intakes of individuals were investigated using the sample variances, the sample third moments and the sample means of the reported intakes for given

individuals. Let the sample variance of daily intakes for the i-th individual be denoted by

$$S_i^2 = \frac{1}{r-1} \sum_{j=1}^{r} (Y_{ij} - \bar{Y}_{i.})^2 , \qquad (5)$$

where r is the number of observations per individual ($r=4$ in our study). Plots of the sample standard deviation, S_i, against the average intake, $\bar{Y}_{i.}$, for the sample individuals are presented in Figures 1 through 5 for the five dietary components. It is evident that for all variables, the average of the sample standard deviations tends to increase as the average intake increases. These plots suggest that the true standard deviation of individual intakes is a linear function of the mean intake.

Let the sample third moment of the individual dietary intakes be denoted by

$$M_{3i} = \frac{r}{(r-1)(r-2)} \sum_{j=1}^{r} (Y_{ij} - \bar{Y}_{i.})^3 . \qquad (6)$$

Plots of the cube root of the sample third moment against the average intake for the five dietary components are presented in Figures 6 through 10. In these plots some individuals have negative sample third moments and others have positive third sample moments. However, the plots suggest that the cube root of the sample third moments is positively correlated with the average intake. A more detailed discussion of models for the moments of individual daily intakes is given in a subsequent section.

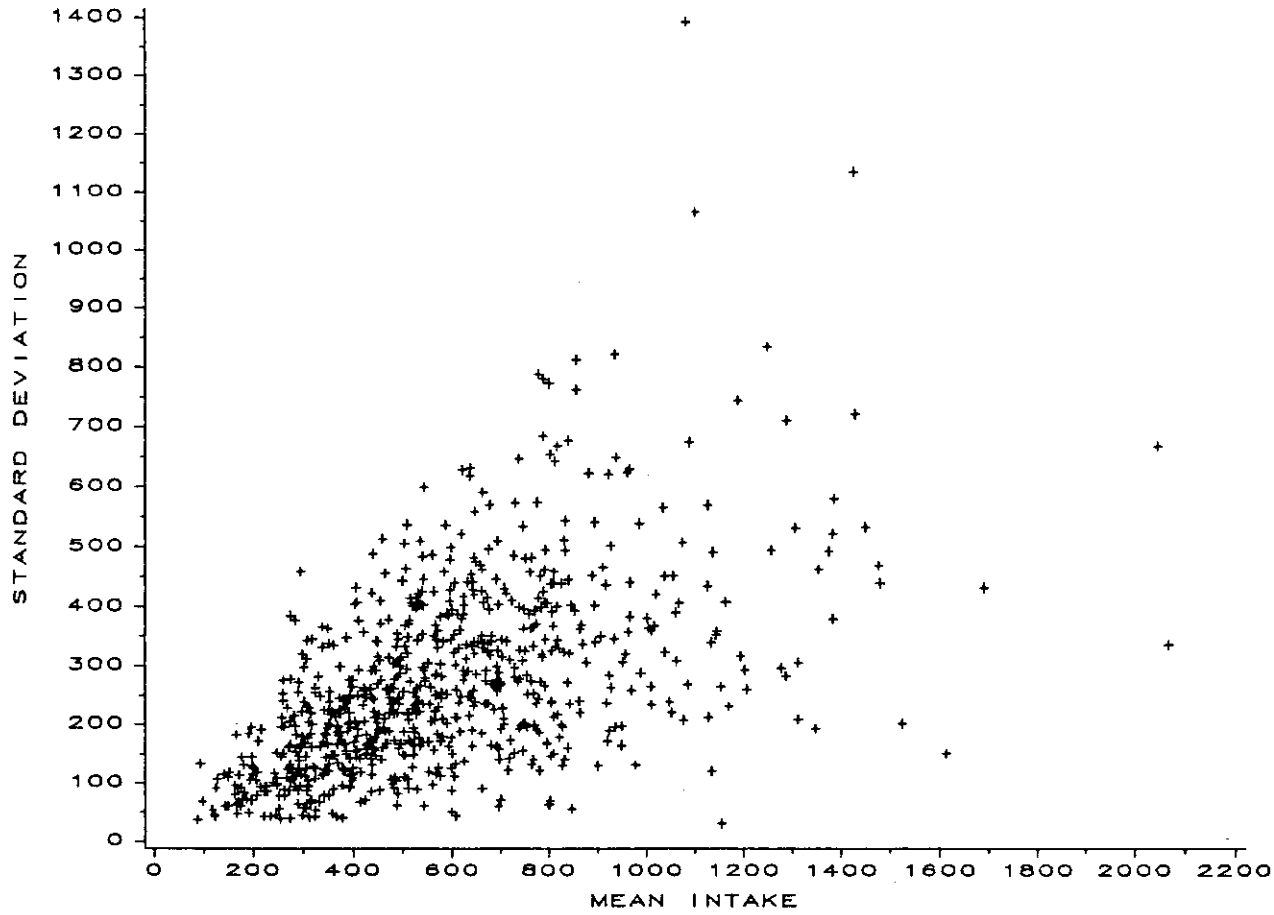

Figure 1. Plot of the sample standard deviations of daily intakes of calcium against the average intakes for women 23-50 years old.

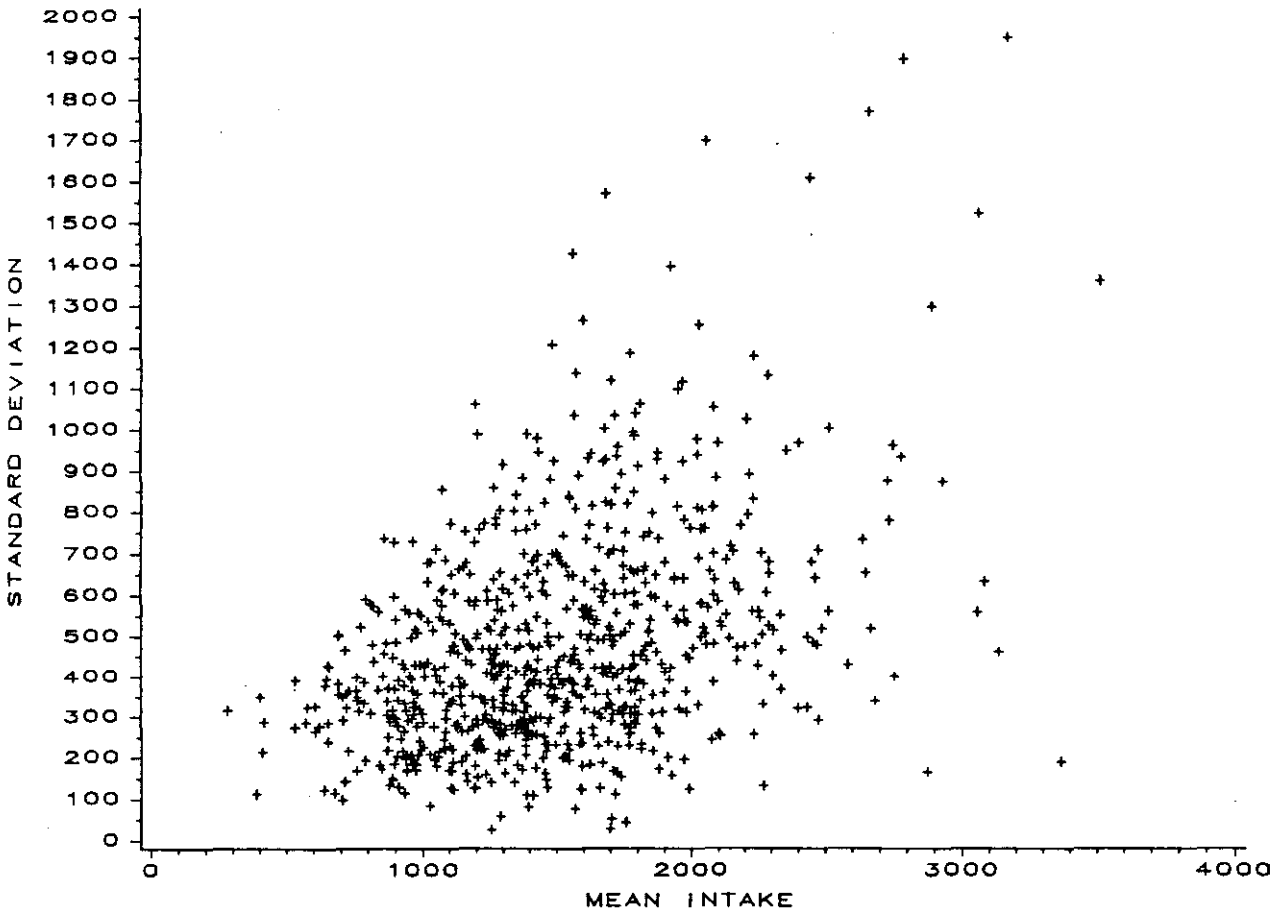

Figure 2. Plot of the sample standard deviations of daily intakes of energy against the average intakes for women 23-50 years old.

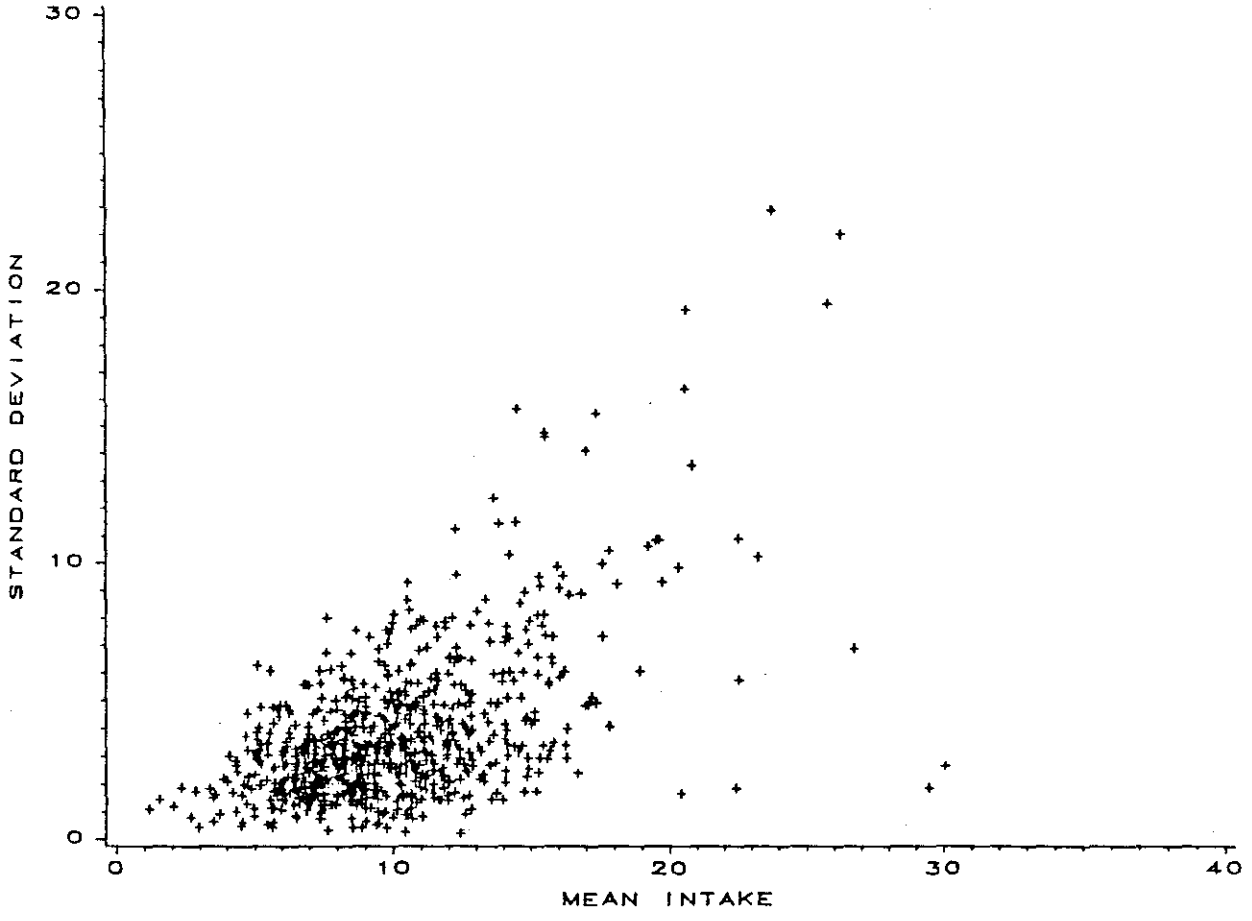

Figure 3. Plot of the sample standard deviations of daily intakes of iron against the average intakes for women 23-50 years old.

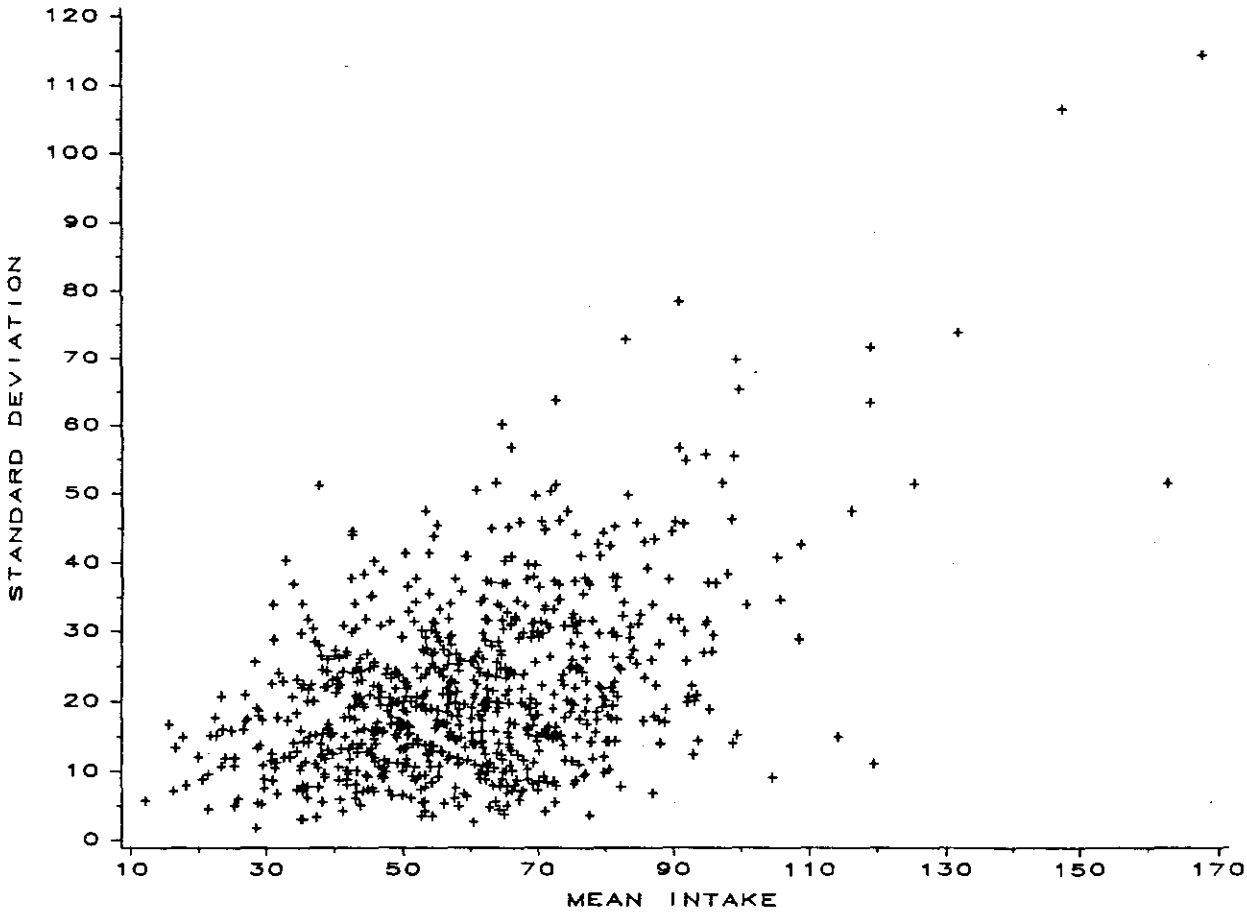

Figure 4. Plot of the sample standard deviations of daily intakes of protein against the average intakes for women 23-50 years old.

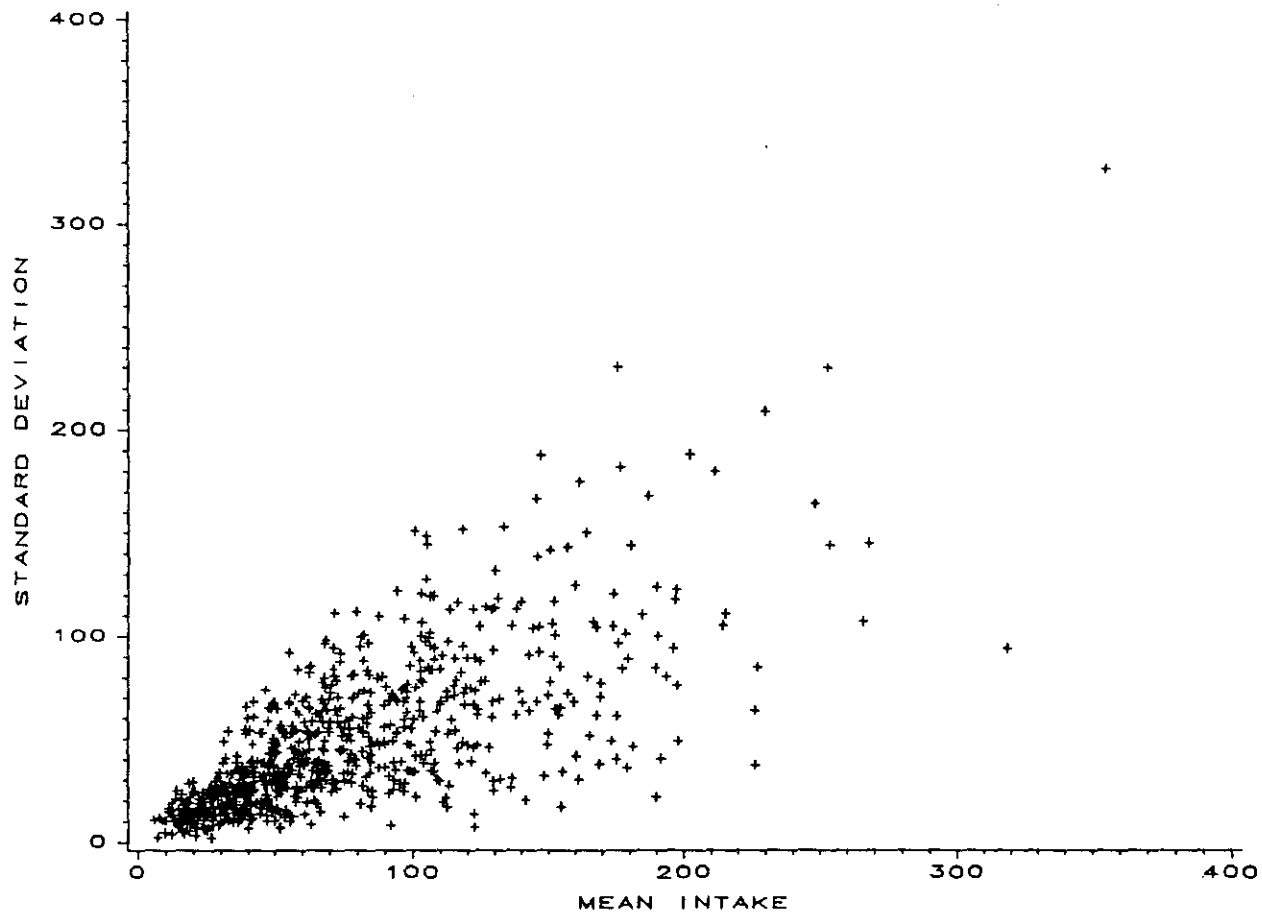

Figure 5. Plot of the sample standard deviations of daily intakes of vitamin C against the average intakes for women 23-50 years old.

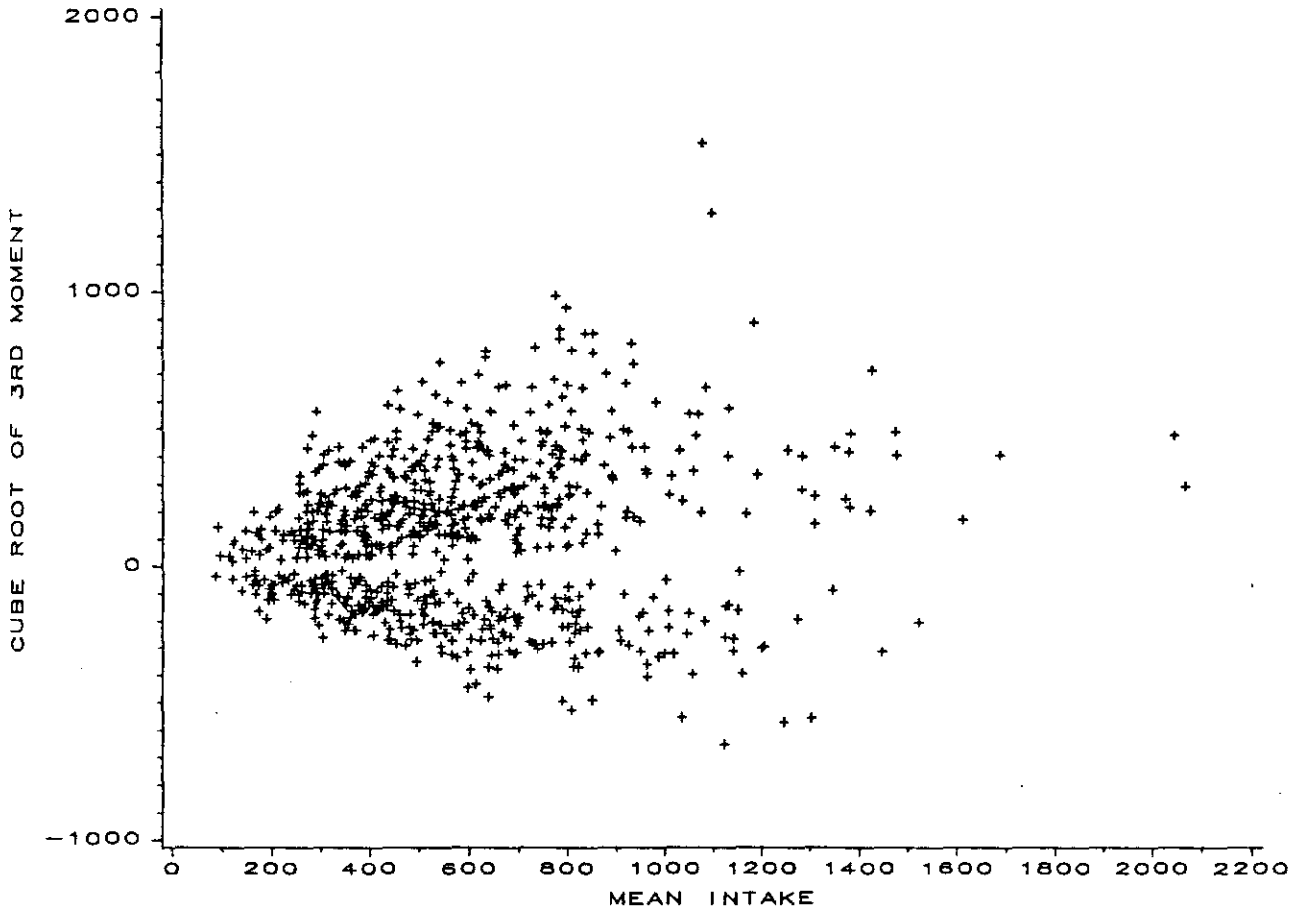

Figure 6. Plot of the cube roots of the sample third moment of daily intakes of calcium against the average intakes for women 23-50 years old.

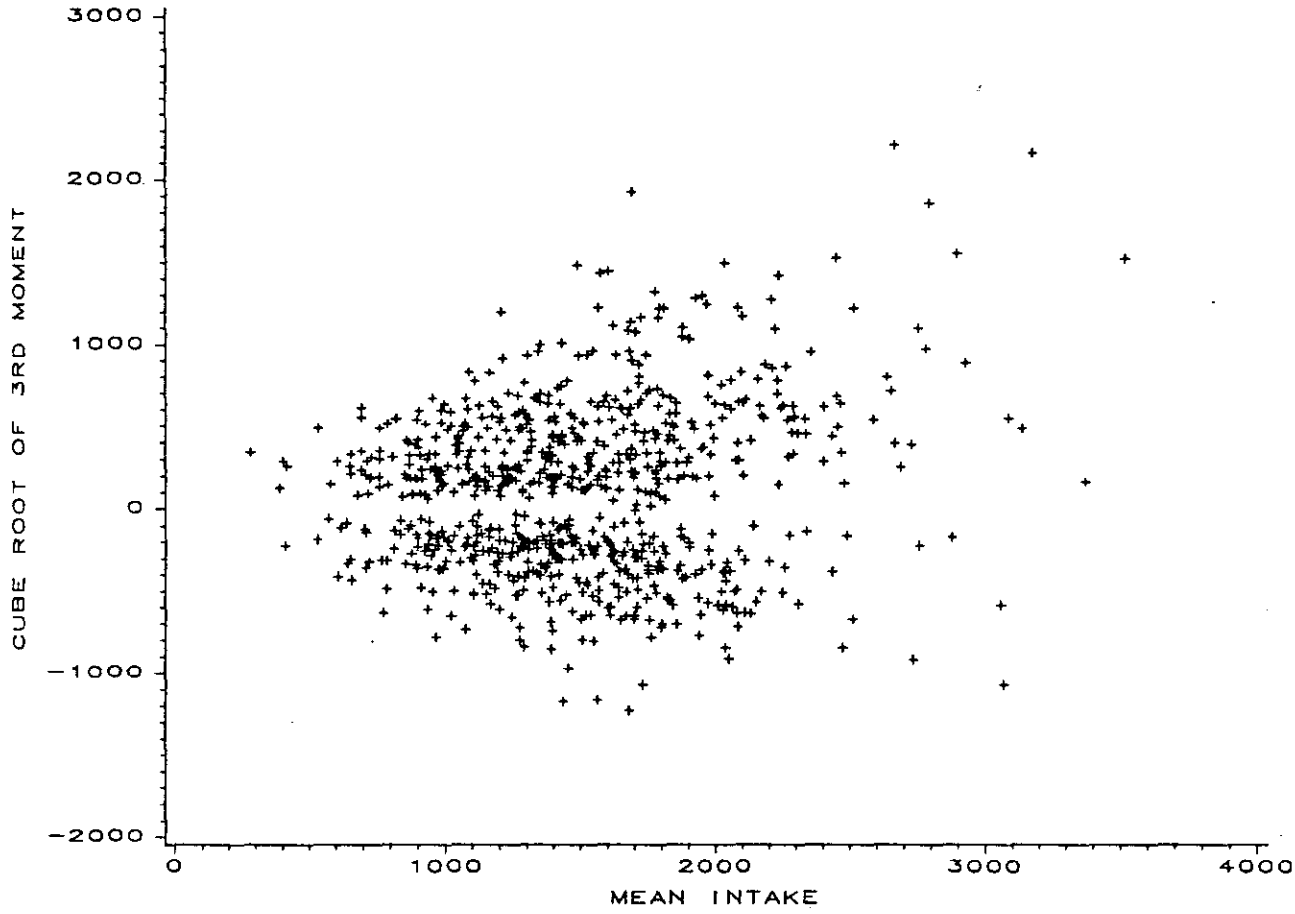

Figure 7. Plot of the cube roots of the sample third moment of daily intakes of energy against the average intakes for women 23-50 years old.

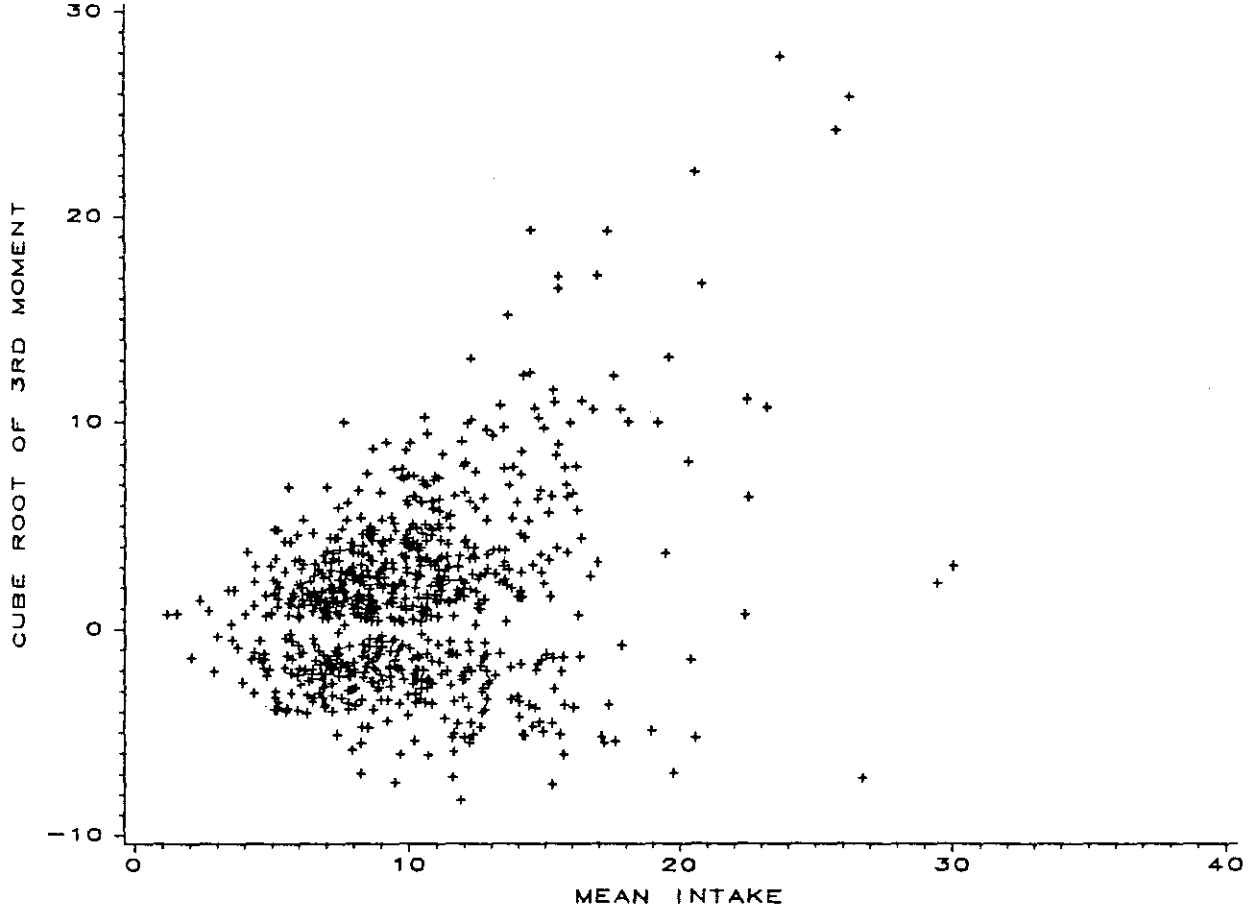

Figure 8. Plot of the cube roots of the sample third moment of daily intakes of iron against the average intakes for women 23-50 years old.

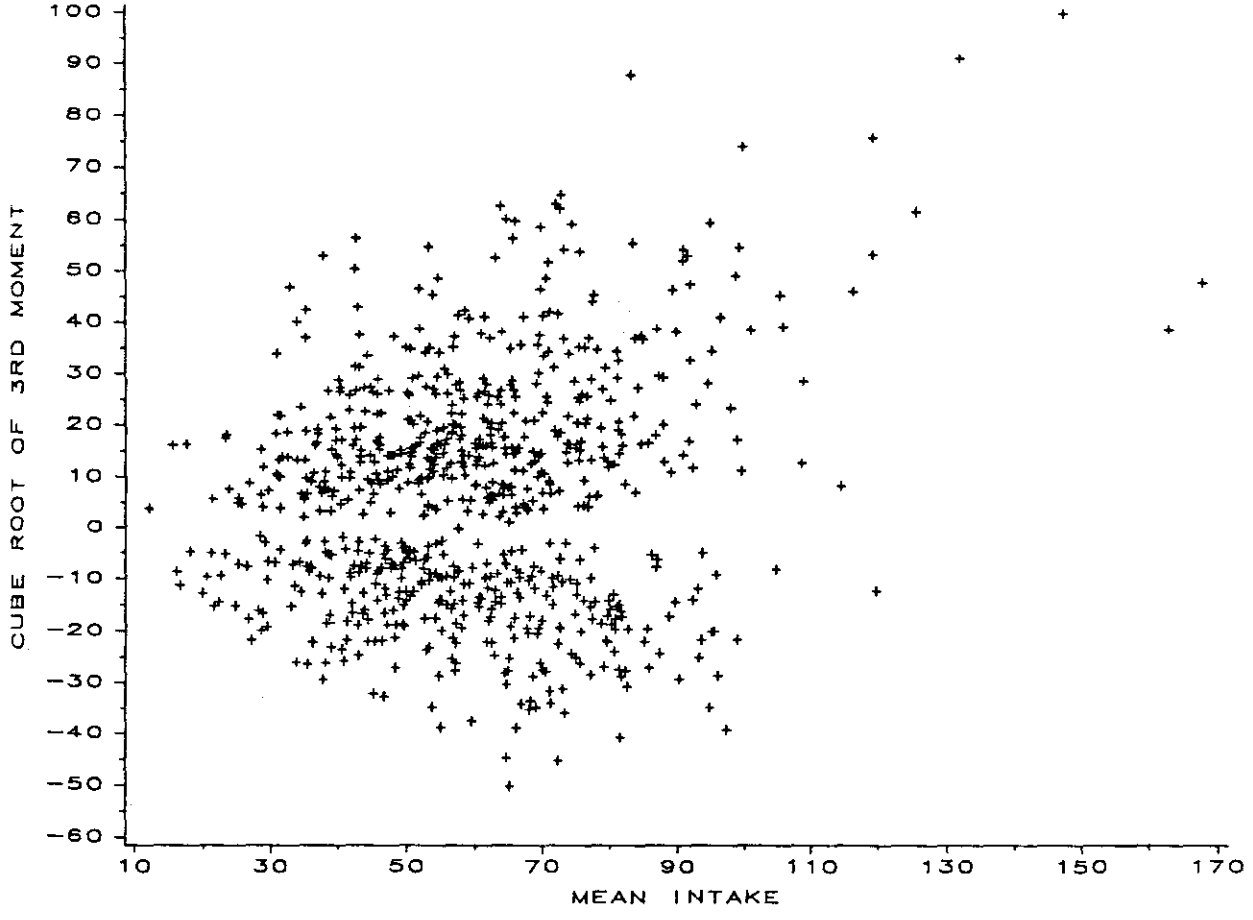

Figure 9. Plot of the cube roots of the sample third moment of daily intakes of protein against the average intakes for women 23-50 years old.

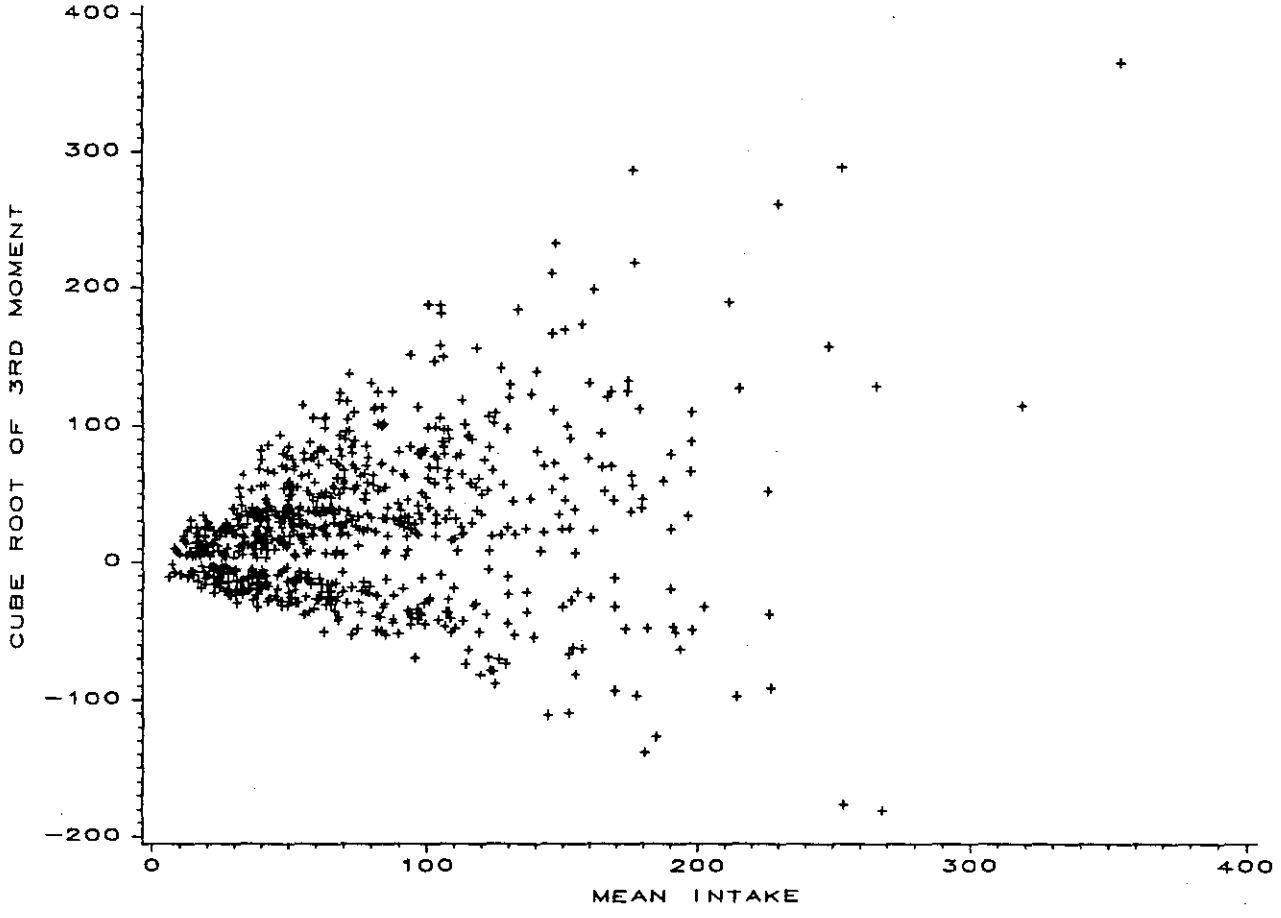

Figure 10. Plot of the cube roots of the sample third moment of daily intakes of vitamin C against the average intakes for women 23-50 years old.

USUAL INTAKE

The concept of the usual intake of a dietary component for an individual is crucial to our study. The usual intake for the i-th individual is defined to be the long-term average of the daily intakes and is denoted by y_i . That is, the usual intake is the conditional expectation of the daily intakes for individual i ,

$$y_i = E(Y_{ij}|i) \ .$$

One can think of usual intake for a given individual as the average of daily intakes where the average is over a sufficiently long period of time, such as a year.

The distribution of usual intake for individuals in the population is important for assessing the adequacy of intakes of a given dietary component. The procedure used to estimate the distribution of usual intake depends on the assumptions made about usual intakes and about the measurement errors, where the measurement error associated with the reported intake for the i-th individual on the j-th reporting day is $Y_{ij} - y_i$. The estimation or prediction of the usual intake for a given sample individual may also be of interest, but is not discusssed in detail in this report.

It is frequently suggested (e.g., National Research Council 1986, p.113) that intake data on dietary components be transformed by a logarithmic or power transformation and that statistical analyses be conducted on the transformed nutrient intakes. We prefer to analyze the original observations when estimating the distribution of usual intake.

A discussion of problems incurred by using transformations is presented in the Appendix.

DISTRIBUTION OF USUAL INTAKE

We assume that the cumulative distribution function of usual intake of a dietary component is of interest for a population of individuals. To estimate the distribution function, it is necessary to define a model for reported intakes in terms of usual intake and measurement errors. Suppose that a random sample of n individuals from the population is available and that r daily intakes are available for each individual. The additive decomposition associated with our definition of usual intake gives

$$Y_{ij} = y_i + e_{ij} , \quad j=1, 2, \ldots, r ; \quad i=1, 2, \ldots, n , \tag{7}$$

where $n=785$ and $r=4$ in our study. Under the definition of usual intake, the measurement errors, e_{ij} , $j=1, 2, \ldots, r$, have zero mean for all individuals, $i=1, 2, \ldots, n$.

We investigate alternative approaches to estimating the distribution function of usual intake using the gamma and Weibull distributions. We first define a model for the measurement errors and estimate the parameters of that model.

Model for Measurement Errors

We assume that the measurement errors, e_{ij} , in the model (7) are such that

$$E(e_{ij}^2 | i) = \alpha\, y_i^2 \ , \quad i=1, 2, \ldots, n \ , \tag{8}$$

$$E(e_{ij}^3 | i) = \gamma\, y_i^3 \ , \quad i=1, 2, \ldots, n \ , \tag{9}$$

and that the sixth moments exist. We also assume that the measurement errors for the i-th individual, $e_{i1}, e_{i2}, \ldots, e_{ir}$, are (conditionally) independent and that the measurement errors for different individuals, e_{ij} and $e_{i'j'}$, where $i \neq i'$, are independent.

Under the model specification (8)-(9), the standard deviations of the measurement errors and the cube roots of the third moments of the measurement errors are directly proportional to the usual intakes of individuals in the population. The model (8) for the variances of the measurement errors is consistent with the plots in Figures 1-5. The model (9) for the third moments of the measurement errors is consistent with the plots in Figures 6-10. As discussed below, these model assumptions appear to be appropriate for the five dietary components: calcium, energy, iron, protein and vitamin C.

Estimators for the parameters, α and γ, are

$$\hat{\alpha} = \Big[\sum_{i=1}^{n} (\bar{Y}_{i.}^2 - r^{-1} S_i^2) \Big]^{-1} \sum_{i=1}^{n} S_i^2 \tag{10}$$

and

$$\hat{\gamma} = \Big[\sum_{i=1}^{n} (\bar{Y}_{i.}^3 - 3r^{-1} \bar{Y}_{i.} S_i^2 + r^{-2} 2M_{3i}) \Big]^{-1} \sum_{i=1}^{n} M_{3i} \ , \tag{11}$$

where S_i^2 and M_{3i} are defined in (5) and (6), respectively. The estimators (10) and (11) are derived in the Appendix. Values of the

estimators, and their estimated standard errors, are presented in Table 4. All parameter estimates are significantly different from zero.

Table 4. Estimated parameters of the measurement error models for the dietary components

	Dietary Component				
Parameter	Calcium	Energy	Iron	Protein	Vitamin C
α	0.247 (0.013)	0.1246 (0.0054)	0.195 (0.012)	0.1708 (0.0076)	0.513 (0.028)
γ	0.123 (0.022)	0.0359 (0.0066)	0.161 (0.035)	0.0594 (0.0089)	0.496 (0.099)

The average daily intake for the i-th individual, $\bar{Y}_{i.}$, estimates the usual intake, y_i, with variance $r^{-1}\sigma_i^2$, where σ_i^2 is the (conditional) variance of the measurement errors for the i-th individual, i.e.,

$$\sigma_i^2 \equiv E(e_{ij}^2 | i) \ .$$

The sample standard deviation of the daily intakes for the i-th individual, S_i, although a biased estimator for the population standard deviation, σ_i, can be used to obtain a pooled estimator for y_i. Using Fisher's approximation, which is based on the normal distribution, the variance of S_i is approximated by $[2(r-1)]^{-1}\sigma_i^2$ (see Kendall and Stuart 1969, p. 371). Then, under variance model (8), we can write

$$\bar{Y}_{i.} = y_i + \bar{e}_{i.} ,$$

$$S_i = \lambda_2 y_i + v_i , \tag{12}$$

where

$$\begin{pmatrix} \bar{e}_{i.} \\ v_i \end{pmatrix} \sim \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} r^{-1}\sigma_i^2 & 0 \\ 0 & \tfrac{1}{2}(r-1)^{-1}\sigma_i^2 \end{pmatrix} \right) .$$

Under normality, the square of λ_2 is a constant multiple of α, where the multiple is a function of the number of observations on a given individual. If λ_2 is known, a pooled estimator for the usual intake, y_i, is

$$\hat{y}_i \equiv \frac{r\,\bar{Y}_{i.} + 2(r-1)\lambda_2 S_i}{r + 2(r-1)\lambda_2^2} . \tag{13}$$

The parameter, λ_2, in (12) can be estimated using the functionally-related option of EV CARP, a computer program for estimation of measurement error models (see Fuller 1987; Schnell and Fuller 1987). The residuals obtained from the measurement error fit of model (12) provide a check for the accuracy of the model (8). Let the residuals from the EV CARP fit be

$$\tilde{v}_i \equiv S_i - \tilde{\lambda}_2 \bar{Y}_{i.} , \quad i=1, 2, \ldots, n ,$$

where $\tilde{\lambda}_2$ is the EV CARP estimator for λ_2. The variance of the i-th residual is approximately $\sigma_i^2\{\tfrac{1}{2}(r-1)^{-1} + r^{-1}\lambda_2^2\}$. Let the predicted

usual intake (13) obtained by using $\tilde{\lambda}_2$ to estimate λ_2 be represented by $\tilde{y}_i$, i=1, 2, ..., n . The weighted residuals, $\tilde{y}_i^{-1}\tilde{v}_i$, have mean zero and common variance under the model. The plot of the weighted residuals, $\tilde{y}_i^{-1}\tilde{v}_i$, against the predicted usual intakes, $\tilde{y}_i$, i=1, 2, ..., n , should exhibit no systematic pattern if the model (8) adequately defines the variances of the measurement errors. These plots are presented in Figures 11 through 15 for the five dietary components. Since the weighted residuals in these figures are clustered around zero with no discernible pattern, it appears that the variance model (8) is adequate for the measurement errors associated with the reported daily intakes of the five dietary components.

A check for the adequacy of the model (9) for the third moments can be obtained by using the residuals from the measurement error fit of the cube root of the sample third moment, M_{3i} , on the sample means. Note that model (9) can be expressed as

$$[E(e_{ij}^3|i)]^{1/3} = \gamma^{1/3} y_i , \quad i=1, 2, \ldots, n .$$

Consider the model

$$M_{3i}^{1/3} = \lambda_3 y_i + w_i , \quad i=1, 2, \ldots, n , \tag{14}$$

where, under the assumptions of model (9), the variance of w_i exists. Let the residuals from the EV CARP fit of model (14) be

$$\tilde{w}_i = M_{3i}^{1/3} - \tilde{\lambda}_3 \tilde{Y}_{i.} , \quad i=1, 2, \ldots, n ,$$

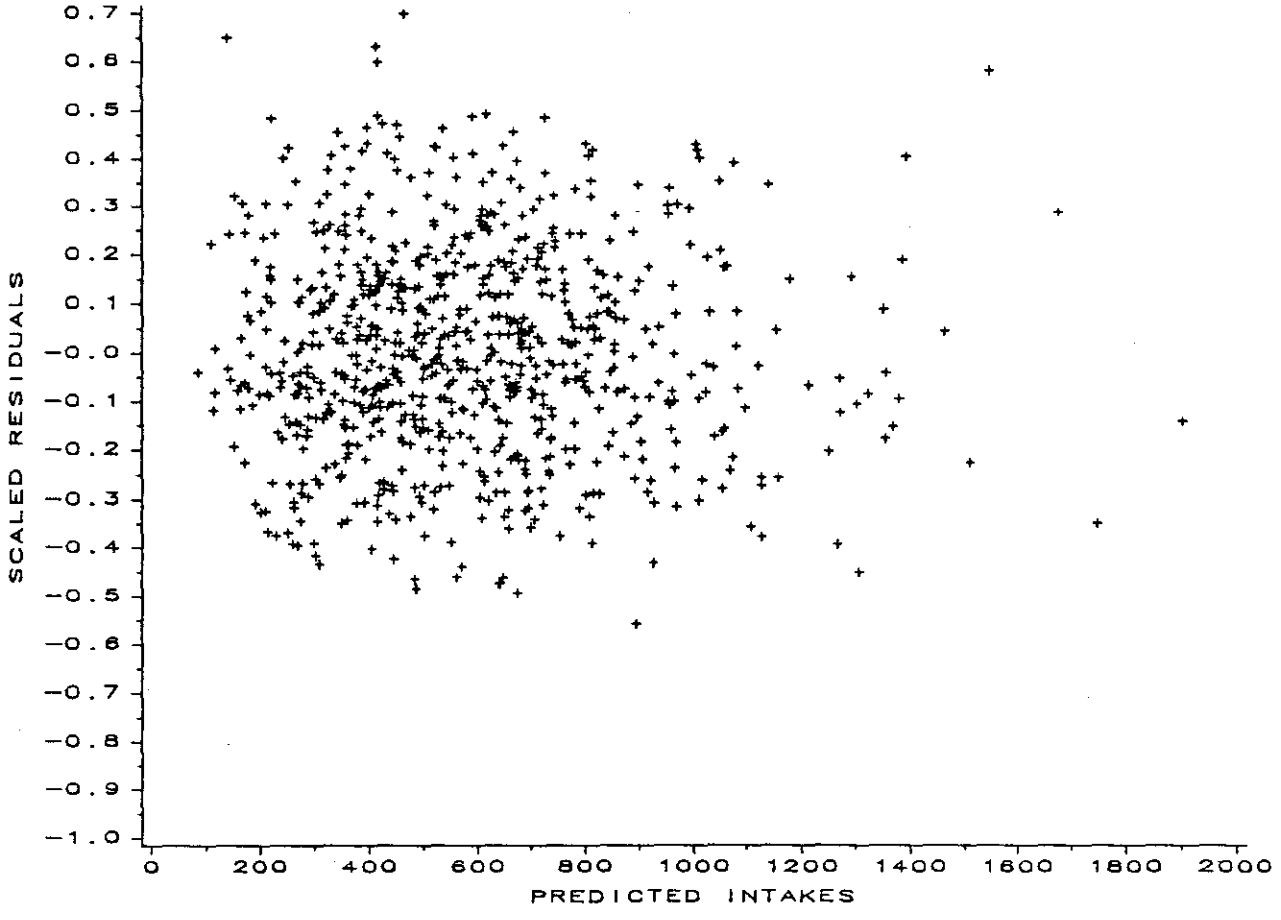

Figure 11. Plot of the weighted residuals for the variance model against the predicted usual intakes of calcium for women 23-50 years old.

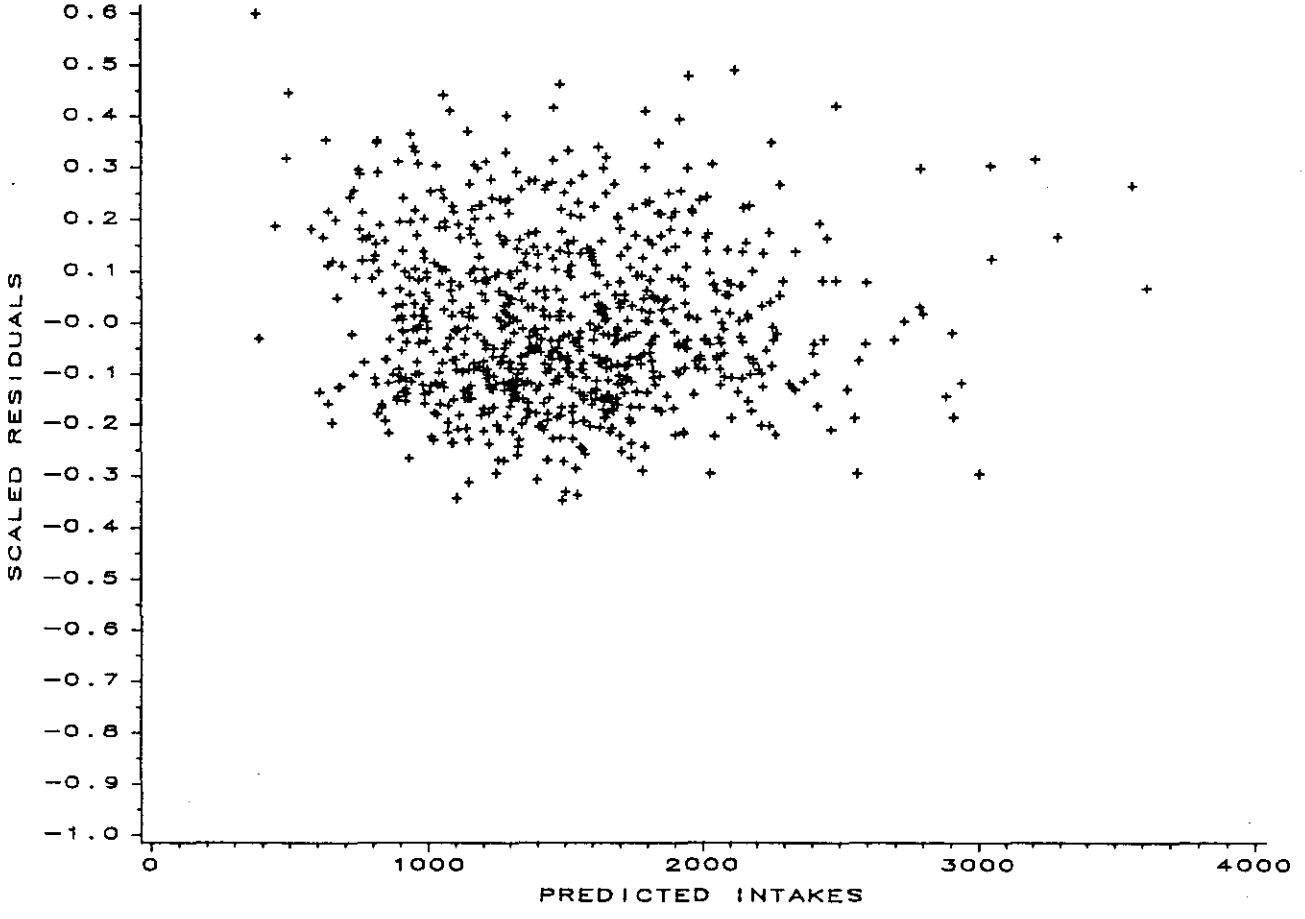

Figure 12. Plot of the weighted residuals for the variance model against the predicted usual intakes of energy for women 23-50 years old.

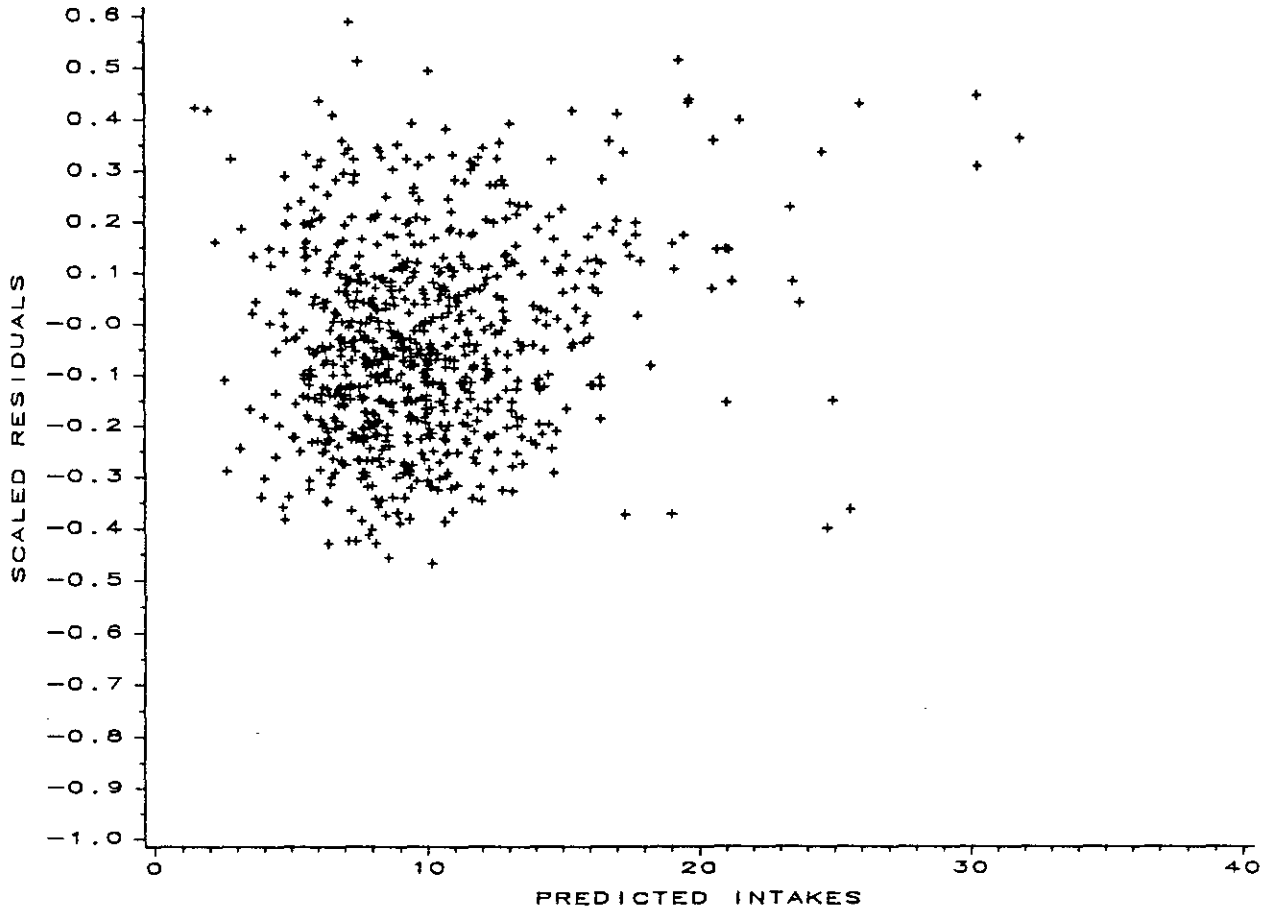

Figure 13. Plot of the weighted residuals for the variance model against the predicted usual intakes of iron for women 23-50 years old.

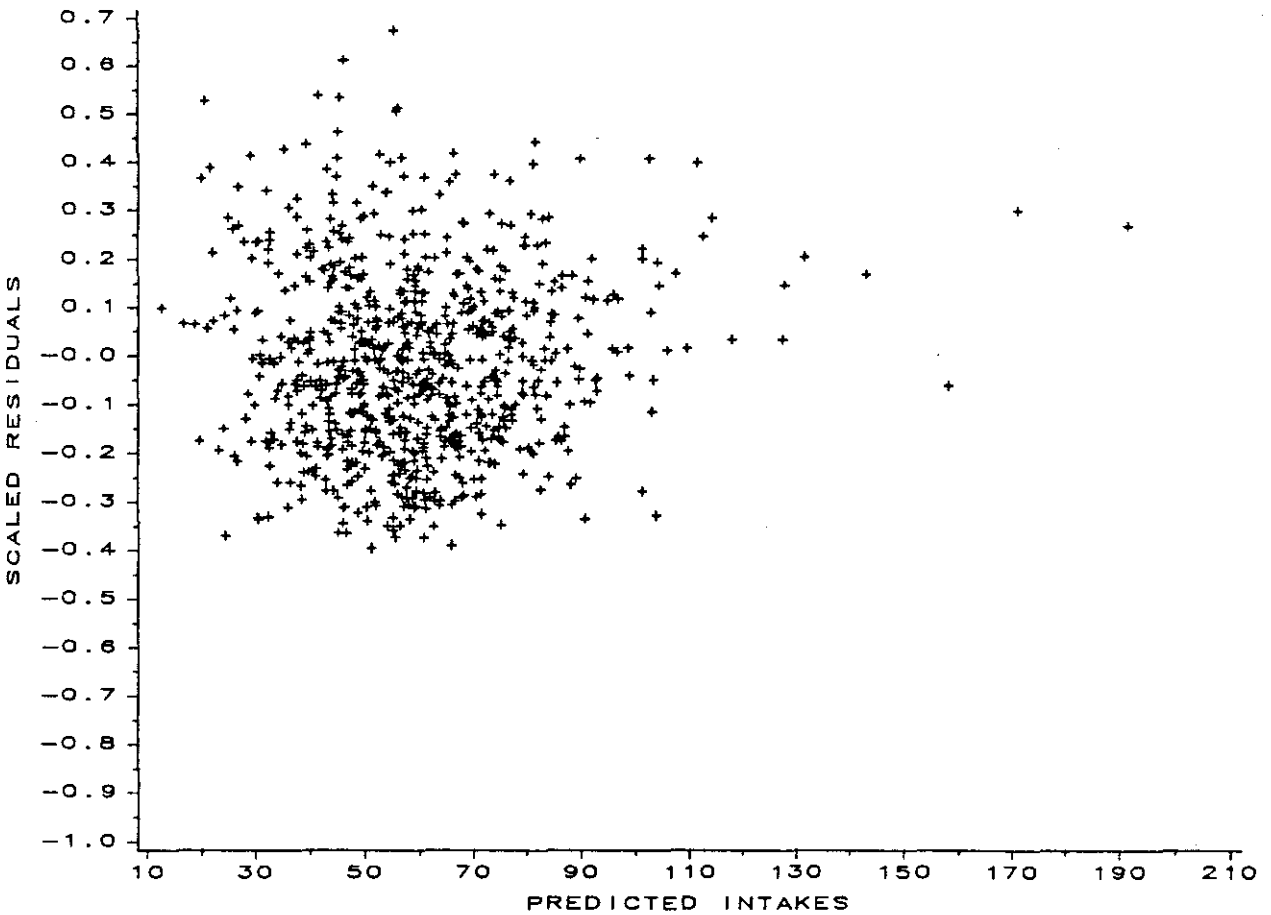

Figure 14. Plot of the weighted residuals for the variance model against the predicted usual intakes of protein for women 23-50 years old.

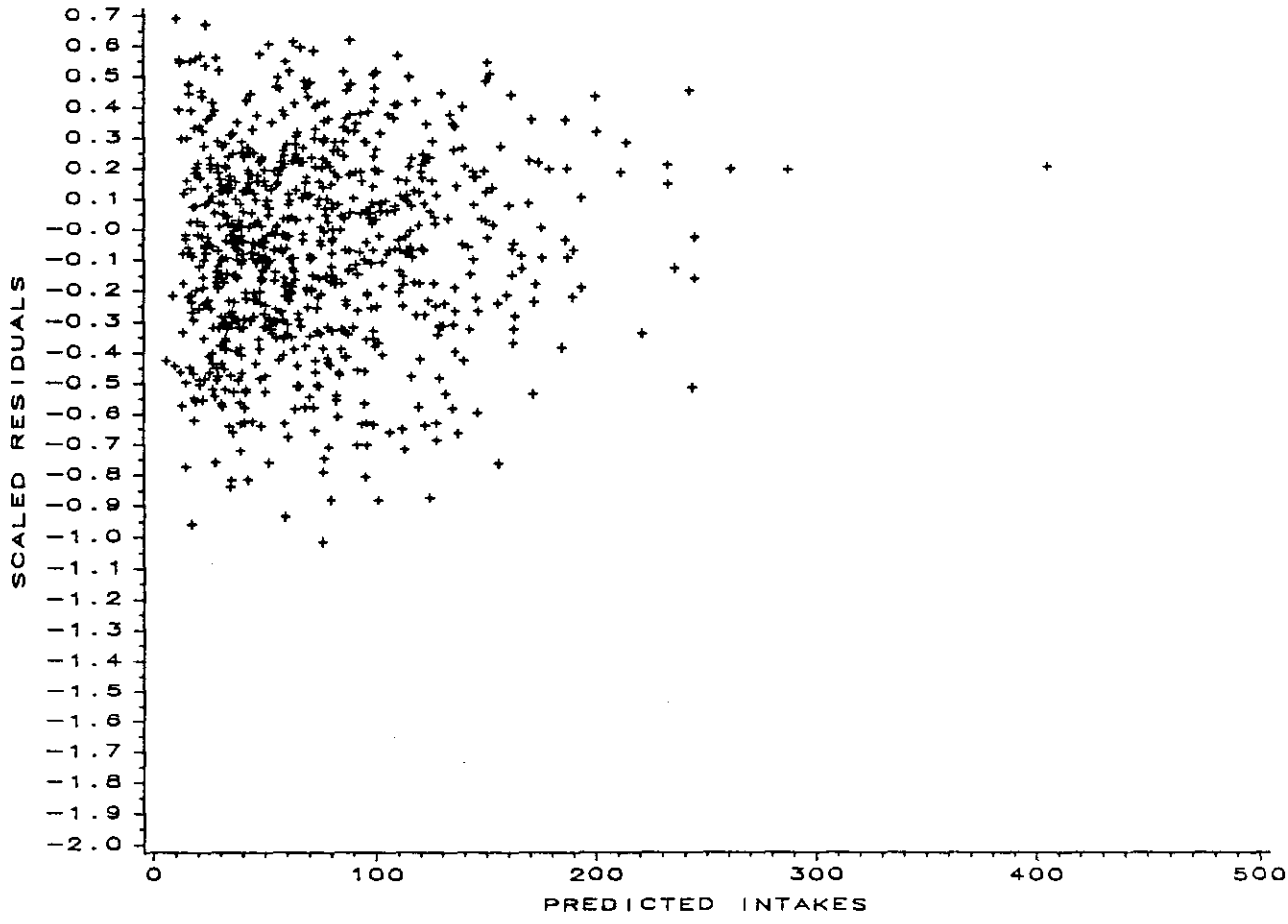

Figure 15. Plot of the weighted residuals for the variance model against the predicted usual intakes of vitamin C for women 23-50 years old.

where $\tilde{\lambda}_3$ is the EV CARP estimator for λ_3. The plot of the weighted residuals, $\tilde{y}_i^{-1}\tilde{w}_i$, against the predicted usual intakes, $\tilde{y}_i$, i=1, 2, ..., n, should exhibit no discernible pattern if the model (9) for the third moments adequately fits the data. Such plots indicated that the postulated model (9) is acceptable. The figures are not included in this paper because they are quite similar to Figures 11 through 15.

Moments of Usual Intake

We assume that the usual intakes, $y_1, y_2, \ldots, y_n$, are a random sample from a distribution with finite fourth moment. The mean of usual intake is represented by μ_y. The second, third and fourth central moments of usual intake are represented by μ_{2y}, μ_{3y} and μ_{4y}, where

$$\mu_{ky} \equiv E(y_i - \mu_y)^k , \quad k=2, 3, 4 .$$

The moments of the average of four daily intakes for a random sample of individuals can be expressed in terms of the moments of usual intake and the parameters of the measurement error model (8)-(9). These derivations are presented in the Appendix. Estimators for the moments of usual intake are also presented in the Appendix.

Estimates for the first four moments of usual intake are presented in Table 5. For example, the estimates for the cube root of the third moment, μ_{3y}, of usual intake are presented in the table. Taking roots results in estimators that are in the same units as the reported intakes. The sign of the cube root is the same as that for the estimate of the third moment. Also presented in Table 5 are estimates for the

Table 5. Estimates for the parameters of the distribution of usual intakes for dietary components

Dietary Component	Moment Parameters					
	μ_y	$\mu_{2y}^{1/2}$	$\mu_{3y}^{1/3}$	$\mu_{4y}^{1/4}$	Skewness $\beta_1^{1/2}$	Kurtosis $\beta_2 - 3$
Calcium (mg)	579.0	233.7	213.9	332.6	0.77	1.10
Energy (kcal)	1,493.2	403.0	245.9	541.9	0.23	0.27
Iron (mg)	10.0	2.88	2.80	4.43	0.91	2.58
Protein (g)	59.6	14.9	10.9	23.6	0.39	3.24
Vitamin C (mg)	75.2	39.2	33.6	50.7	0.63	-0.22

skewness and kurtosis parameters, $\beta_1^{1/2}$ and $\beta_2 - 3$, where $\beta_1^{1/2} = \mu_{3y}/\mu_{2y}^{3/2}$ and $\beta_2 = \mu_{4y}/\mu_{2y}^2$.

The differences between the moments of Table 3 and those of Table 5 are worthy of note. The estimated variances of usual intake range from 70% of the estimated variance of the four-day mean for calcium to 58% of the estimated variance of the four-day mean for protein. Thus, the variability of daily intakes makes an important contribution to the total variability of four-day means. In all cases, the estimated skewness for the distribution of usual intake is less than the observed skewness for the four-day means.

Because the estimates of the third moments of usual intake are positive for all dietary components, the estimated distributions of

usual intakes are positively skewed. However, energy usual intake is basically symmetrical. The estimated kurtosis for usual intake is smaller than the estimated kurtosis of four-day means for all dietary components except protein. The distributions of usual intake for calcium, iron and protein appear to have fatter tails than the normal distribution. The kurtosis for energy and vitamin C differ little from that of the normal distribution.

Gamma Distributions

In this section, we assume that the usual intakes, $y_1, y_2, \ldots, y_n$, are a random sample from the two-parameter gamma distribution with density function

$$f(y) = \theta^{-\beta}\Gamma^{-1}(\beta)y^{\beta-1}e^{-y/\theta} \quad , \quad y > 0, \ \theta > 0, \ \beta > 0 \ ,$$

where β and θ are parameters to be estimated. Given that the shape parameter, β, is greater than one, the density function is unimodal and right-skewed with a value of zero at the origin. Furthermore, the first three moments of usual intake are

$$E(y_i) = \theta\beta \ ,$$

$$E(y_i - \theta\beta)^2 = \theta^2\beta$$

and

$$E(y_i - \theta\beta)^3 = 2\theta^3\beta \ .$$

In addition, we assume that the measurement errors for the i-th individual, $e_{i1}, e_{i2}, \ldots, e_{ir}$, are (conditionally) independent random variables, defined by

$$e_{ij} = Z_{ij} - E(Z_{ij}|i) \quad , \quad j=1, 2, \ldots, r \quad ,$$

where $\{Z_{i1}, Z_{i2}, \ldots, Z_{ir}\}$ is a random sample from the gamma distribution with parameters θ_{ei} and β_e. These assumptions imply that the gamma distributions associated with the measurement errors on different individuals have different scale parameters, θ_{ei}, $i=1, 2, \ldots, n$, but a common shape parameter, β_e. Given that the model is constrained to satisfy the moment properties defined by equations (8) and (9), it follows that $\theta_{ei} = \delta y_i$, $i=1, 2, \ldots, n$, where δ is a positive constant. Furthermore, the parameters δ and β_e are expressible in terms of α and γ by

$$\delta = (2\alpha)^{-1}\gamma$$

and

$$\beta_e = 4\alpha^3\gamma^{-2} \quad .$$

Thus, method-of-moments estimators for the common parameters of the distribution of the measurement errors are

$$\hat{\delta} = (2\hat{\alpha})^{-1}\hat{\gamma}$$

and

$$\hat{\beta}_e = 4\hat{\alpha}^3\hat{\gamma}^{-2}$$

where $\hat{\alpha}$ and $\hat{\gamma}$ are the method-of-moments estimators for α and γ derived in the Appendix. Values of the estimators for δ and β_e for the five dietary components are given in Table 6.

Table 6. Estimates for the parameters of the hypothesized gamma distribution for measurement errors for five dietary components

Dietary component	$\hat{\delta}$	$\hat{\beta}_e$
Calcium	0.249	3.969
Energy	0.144	6.021
Iron	0.412	1.151
Protein	0.174	5.644
Vitamin C	0.483	2.203

The method-of-moments estimators for the parameters of the distribution of usual intake are

$$\hat{\theta} = \hat{\mu}_y^{-1}\hat{\mu}_{2y}$$

and

$$\hat{\beta} = \hat{\mu}_{2y}^{-1}\hat{\mu}_y^2 ,$$

where $\hat{\mu}_y$ and $\hat{\mu}_{2y}$ are the estimators for the mean and variance of usual intake, defined in the Appendix. Estimates of the scale and shape

parameters for calcium, energy, iron, protein and vitamin C are presented in Table 7.

Table 7. Scale and shape parameter estimates for the gamma distribution of usual intake for five dietary components

Dietary component	$\hat{\theta}$	$\hat{\beta}$
Calcium	94.34	6.14
Energy	108.77	13.73
Iron	0.83	12.01
Protein	3.75	15.88
Vitamin C	20.47	3.67

To test the fit of these distributions, Monte Carlo methods were used to generate the distribution of individual four-day means from the estimated gamma distributions for usual intakes and measurement errors. For each nutrient, 100,000 usual intakes y_i were generated along with $r = 4$ measurement errors e_{ij} for each y_i according to the parameters of the respective estimated gamma distributions. The usual intake plus the average error for each intake,

$$y_i + \bar{e}_{i.} ,$$

were used to generate a cumulative distribution function (cdf) against which the empirical cdf for observed individual means could be compared. The hypothesized cdf for individual means was generated by

counting the number of generated observations contained in each of 1,000 intervals over the range of observed means. A chi-square goodness-of-fit statistic was used as the test statistic. Values of the test statistic are listed in Table 8 for each of the five dietary components. The chi-square goodness-of-fit statistics are significant for energy and protein. The assumptions that usual intake has a gamma distribution and the measurement errors are generated from gamma distributions are not satisfactory for energy and protein, but are satisfactory for calcium, iron and vitamin C.

Table 8. Goodness-of-fit statistics for the distribution of four-day mean intakes based on gamma distributions

Dietary component	χ^2 [a]
Calcium	35.39
Energy	47.69
Iron	25.01
Protein	46.59
Vitamin C	24.61

[a]For size, 0.05, the hypothesized distributional assumptions are rejected if the chi-square goodness-of-fit statistic, χ^2, exceeds 40.11, which is the 95-th percentile for the chi-square distribution with 27 degrees of freedom.

Plots comparing the empirical cdf of individual means with the hypothesized cdf based on gamma densities for each dietary component are shown in Figures 16 through 20. The plots for energy and protein indicate a poor fit to the hypothesized distribution in the neighborhood of the medians.

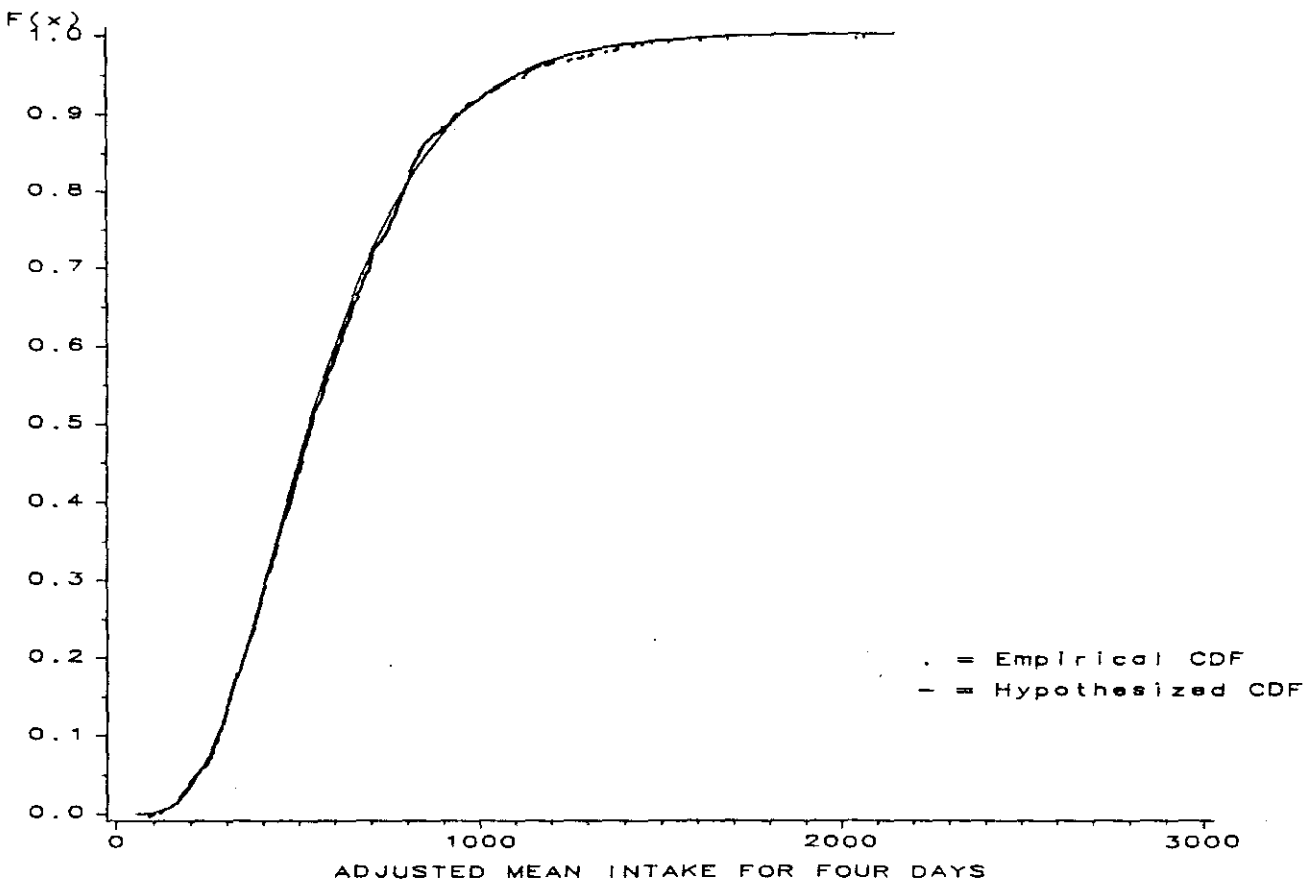

Figure 16. Plots comparing the empirical cdf of individual means with the hypothesized cdf based on gamma densities for calcium.

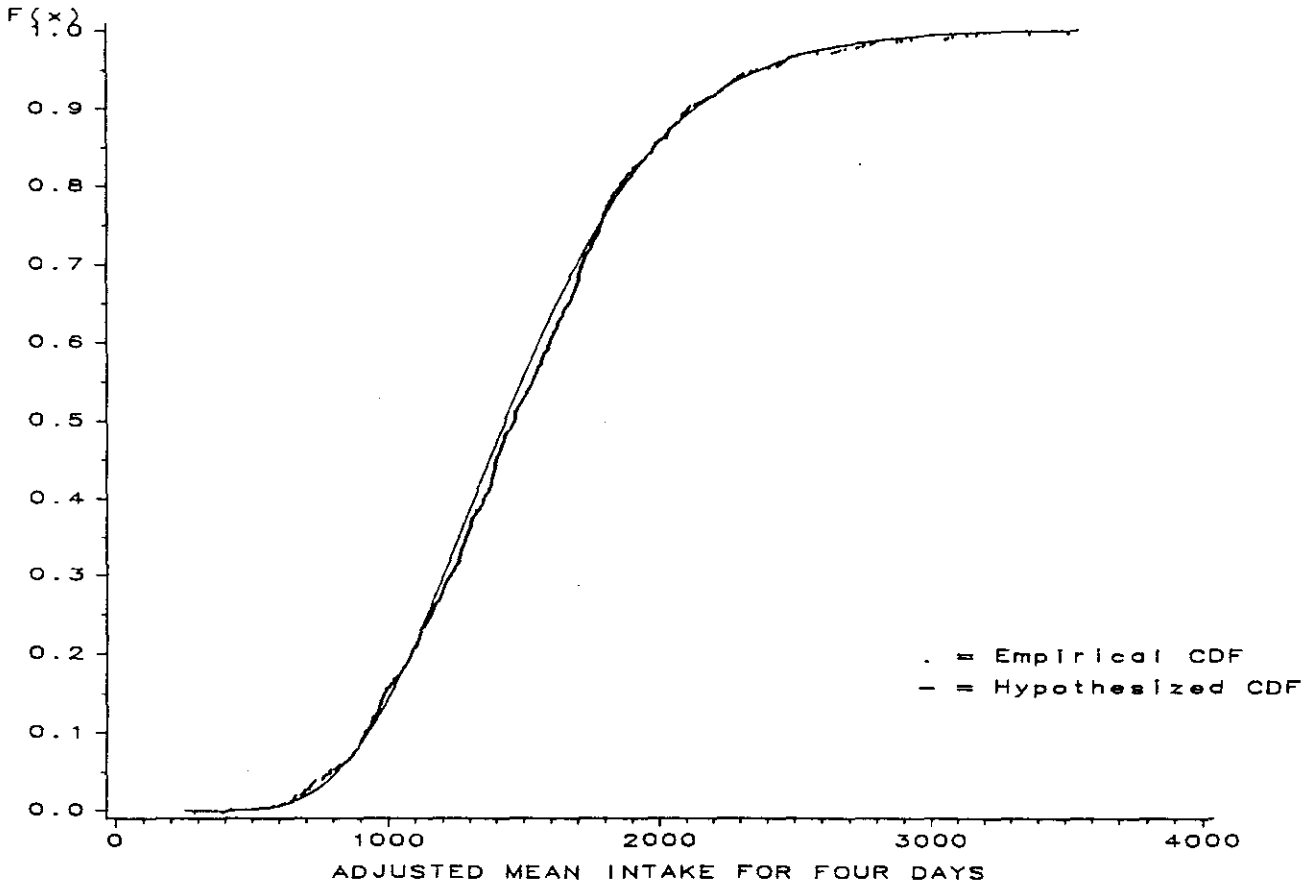

Figure 17. Plots comparing the empirical cdf of individual means with the hypothesized cdf based on gamma densities for energy.

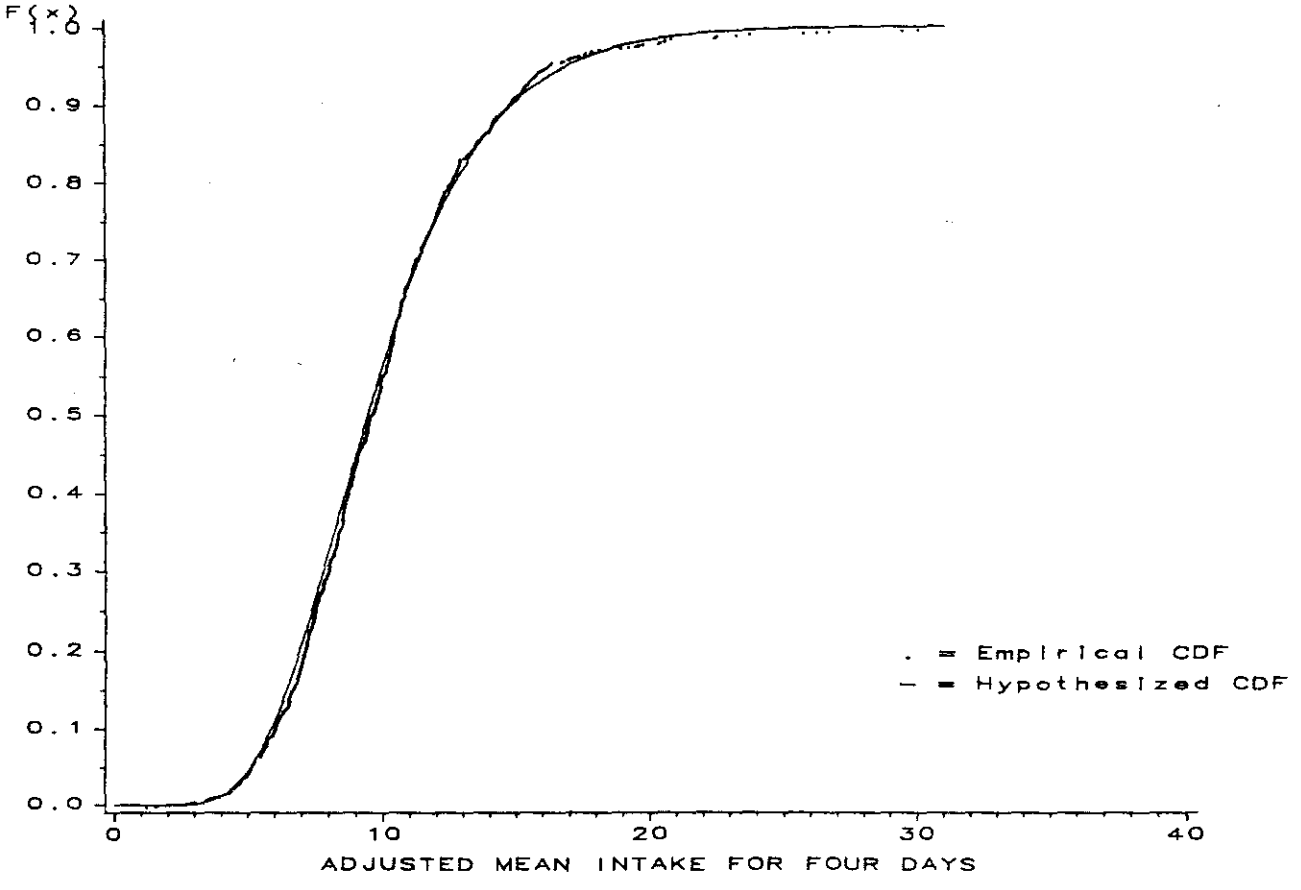

Figure 18. Plots comparing the empirical cdf of individual means with the hypothesized cdf based on gamma densities for iron.

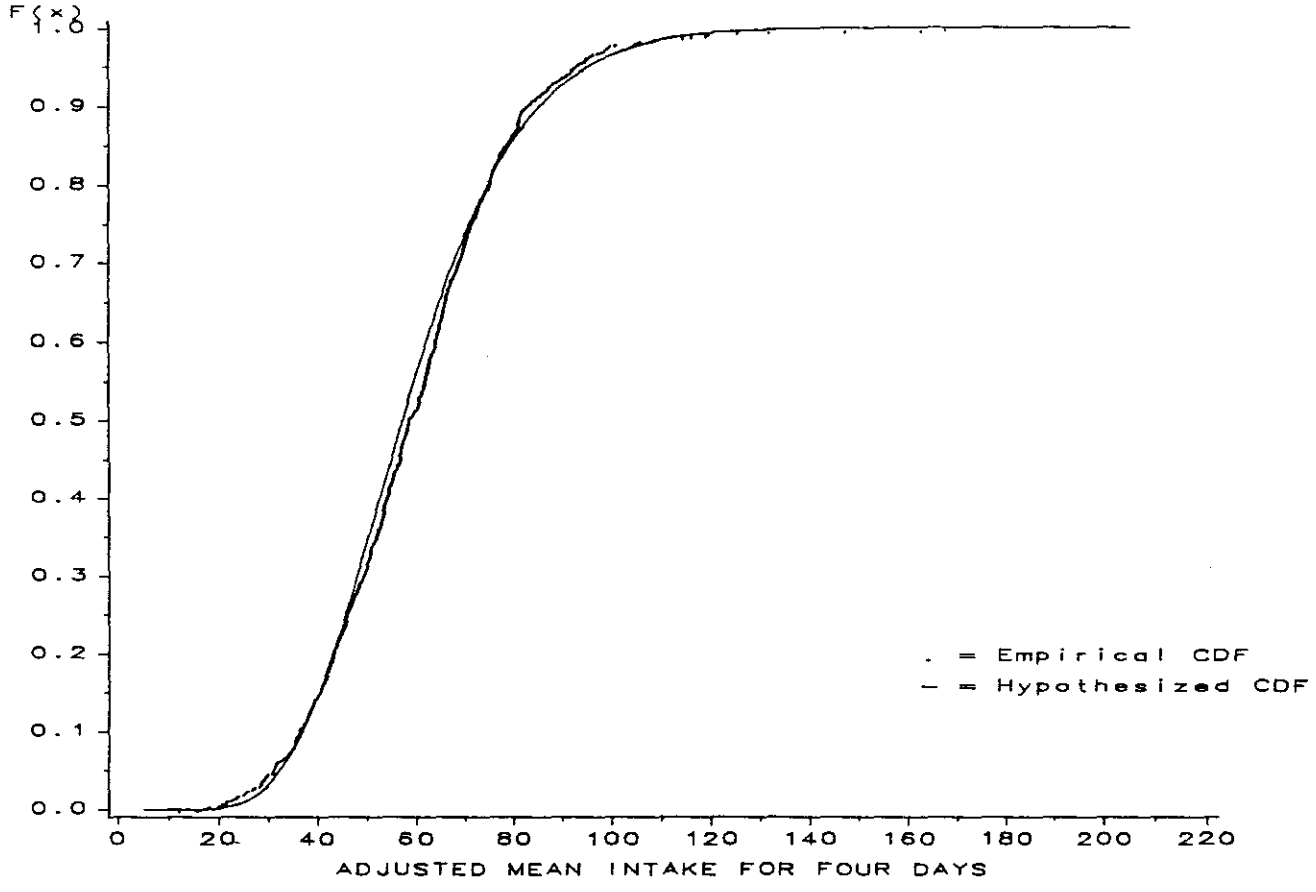

Figure 19. Plots comparing the empirical cdf of individual means with the hypothesized cdf based on gamma densities for protein.

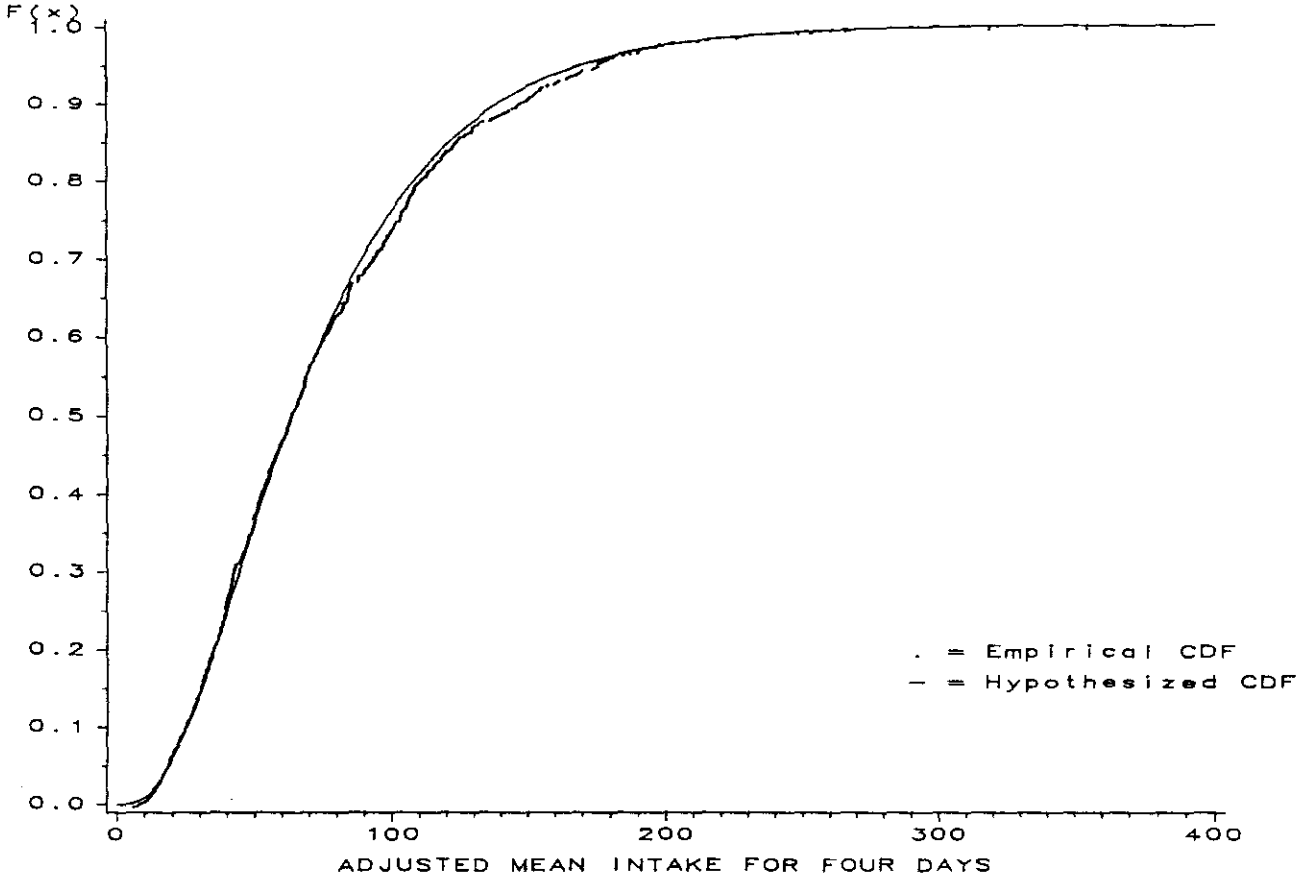

Figure 20. Plots comparing the empirical cdf of individual means with the hypothesized cdf based on gamma densities for vitamin C.

Because of the poor fit for energy and protein, alternative models were developed for those components. In order to make the fitted distribution more symmetric, it was assumed that the square of usual intake had gamma distribution. The parameters of the gamma distribution were estimated using the second and fourth moments of the original observations. The estimated parameters were $(\hat{\theta}_S, \hat{\beta}_S) = (667, 678.4, 3.58)$ and $(\hat{\theta}_S, \hat{\beta}_S) = (986.7, 3.82)$ for the square of energy and protein, respectively. The plots of the estimated cdf and empirical cdf are given in Figures 17a and 19a for energy and protein, respectively. The chi-square goodness-of-fit statistics for these models were 34.05 and 37.58, respectively, which are not significant at the five-percent level. Thus, the gamma distributions in the squares are acceptable for the usual intake of energy and protein.

Weibull Distributions

In this section, we assume that the usual intakes, $y_1, y_2, \ldots, y_n$, are a random sample from the two-parameter Weibull distribution with parameters τ and η, having density function

$$f(y) = \tau^{-\eta} \eta \, x^{\eta-1} e^{-(x/\tau)^{\eta}} , \quad x > 0 , \quad \eta > 0 , \quad \tau > 0 .$$

Given that the shape parameter, η, is greater than one, the density is right-skewed and unimodal with a value of zero at the origin. The first three moments of the Weibull distribution are

$$E(y_i) = \tau \, \Gamma(1 + \eta^{-1}) ,$$

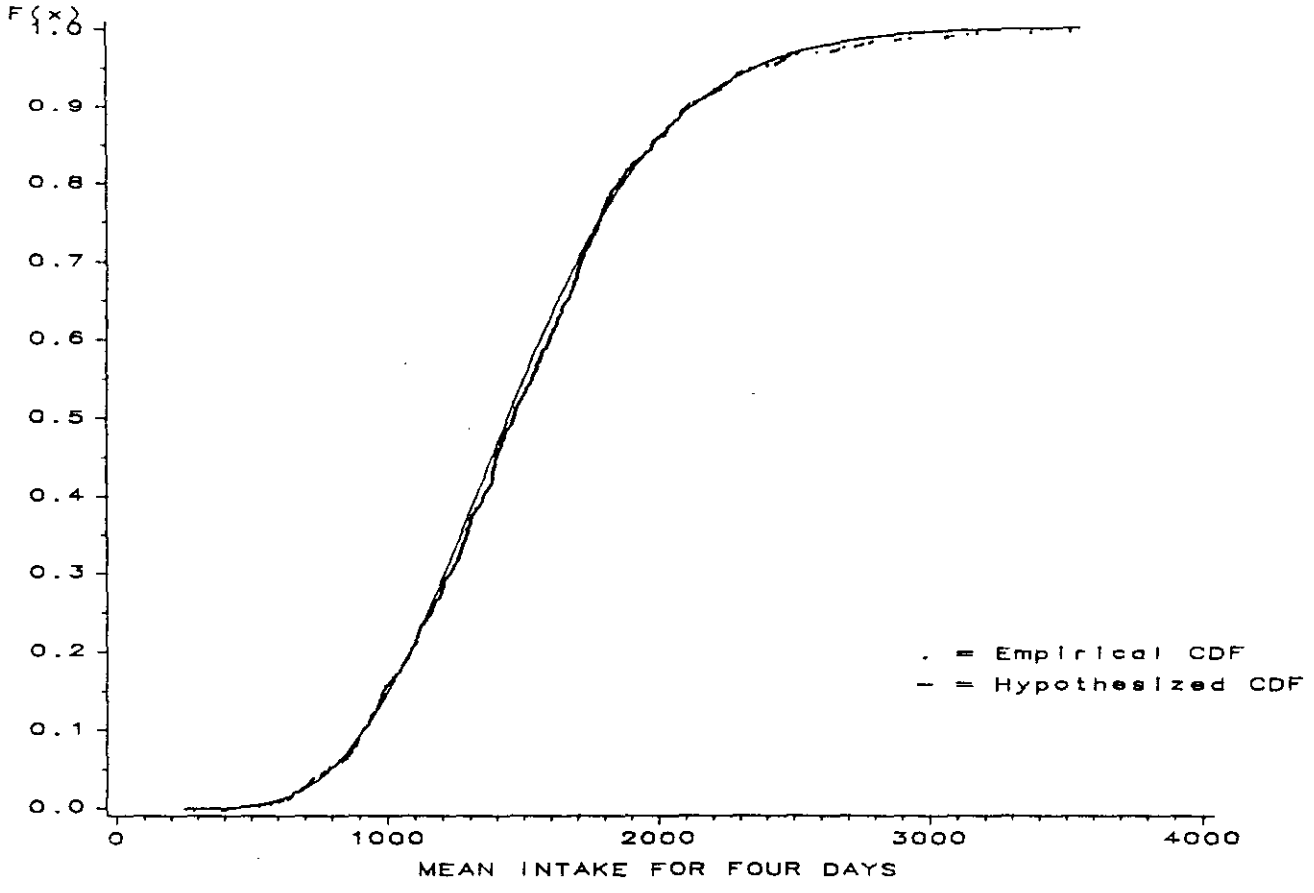

Figure 21. Plots comparing the empirical cdf of individual means with the hypothesized cdf based on the assumption that the square of usual intake of energy has gamma distribution.

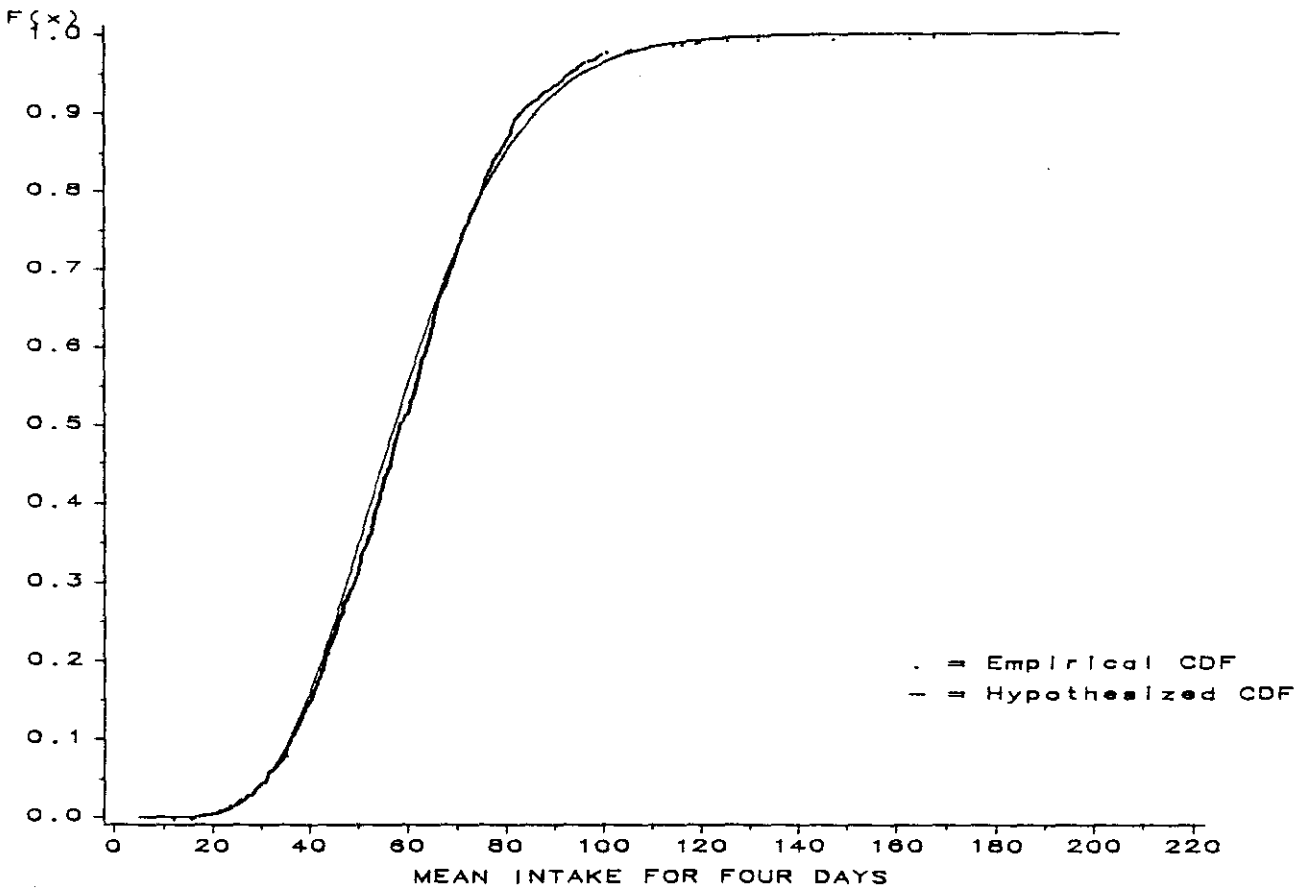

Figure 22. Plots comparing the empirical cdf of individual means with the hypothesized cdf based on the assumption that the square of usual intake of protein has gamma distribution.

$$E\{y_i - \tau\Gamma(1 + \eta^{-1})\}^2 = \tau^2[\Gamma(1 + 2\eta^{-1}) - \Gamma^2(1 + \eta^{-1})] ,$$

and

$$E\{y_i - \tau\Gamma(1+\eta^{-1})\}^3 = \tau^3[\Gamma(1+3\eta^{-1}) - 3\Gamma(1+2\eta^{-1})\Gamma(1+\eta^{-1}) + 2\Gamma^3(1+\eta^{-1})] .$$

In addition, the distribution of the measurement errors, $e_{i1}, e_{i2}, \ldots, e_{ir}$, for the i-th individual are assumed to be (conditionally) independent random variables, defined by

$$e_{ij} = W_{ij} - E(W_{ij}|i) , \quad j=1, 2, \ldots, r ,$$

where $\{W_{i1}, W_{i2}, \ldots, W_{ir}\}$ is a random sample from the Weibull distribution with parameters τ_{ei} and η_e. Given that the model is constrained to satisfy the moment properties of equations (8) and (9), it follows that the method-of-moments estimators for τ_{ei} and η_e are defined by

$$\hat{\tau}_{ei} = \hat{\delta}\bar{Y}_{i.} ,$$

$$\hat{\alpha} = \hat{\delta}^2[\Gamma(1 + 2\hat{\eta}_e^{-1}) - \Gamma^2(1 + \hat{\eta}_e^{-1})] ,$$

and

$$\hat{\gamma} = \hat{\delta}^3 [\Gamma(1 + 3\hat{\eta}_e^{-1}) - 3\Gamma(1 + 2\hat{\eta}_e^{-1})\Gamma(1 + \hat{\eta}_e^{-1}) + 2\Gamma^3(1 + \hat{\eta}_e^{-1})] ,$$

where $\hat{\alpha}$ and $\hat{\gamma}$ are the estimators for the parameters of the moments of the measurement errors (8)-(9). The IMSL iterative routine DNEQNF,

which uses the Levenburg-Marquadt algorithm and a finite-difference approximation to the Jacobian, was employed to generate the parameter estimates listed in Table 9. The method-of-moments estimators for the parameters τ and η are defined by the system of equations

$$\hat{\mu}_y = \hat{\tau}\Gamma(1 + \hat{\eta}^{-1})$$

and

$$\hat{\mu}_{2y} = \hat{\tau}^2[\Gamma(1 + 2\hat{\eta}^{-1}) - \Gamma^2(1 + \hat{\eta}^{-1})] \ .$$

The IMSL routine DNEQNF was used to construct the estimates of $\hat{\tau}$ and $\hat{\eta}$ by solving the nonlinear system of equations. The parameter estimates are listed in Table 10.

Table 9. Estimates for the parameters of the hypothesized Weibull distribution for measurement errors for five dietary components

Dietary components	$\hat{\delta}$	$\hat{\kappa}_e$
Calcium	0.845	1.560
Energy	0.675	1.757
Iron	0.472	1.048
Protein	0.776	1.726
Vitamin C	0.999	1.299

The cumulative distribution function for individual means, assuming the Weibull distributions apply, was generated by the same methods

Table 10. Scale and shape parameter estimates for the Weibull distribution of usual intake for five dietary components

Dietary components	$\hat{\tau}$	$\hat{\eta}$
Calcium	651.4	2.67
Energy	1643.4	4.17
Iron	11.0	3.88
Protein	65.2	4.52
Vitamin C	84.8	2.00

explained above for the gamma distributions. Results of the goodness-of-fit statistics are presented in Table 11. The chi-square goodness-of-fit statistics are significant at the five-percent level for calcium and iron. These results indicate that the assumptions that usual intake has Weibull distribution and the measurement errors are generated from Weibull distributions are not appropriate for calcium and iron.

Table 11. Goodness-of-fit statistics results for testing the distribution of four-day intakes based on Weibull distributions for five dietary components

Dietary components	χ^2
Calcium	42.85
Energy	32.52
Iron	43.07
Protein	26.09
Vitamin C	28.48

Figures 21 through 25 contain plots of the empirical cdf and Weibull-based cdf for the five variables. The plots indicate that the hypothesized distribution functions do not fit well in the tails in the cases of calcium, energy and iron.

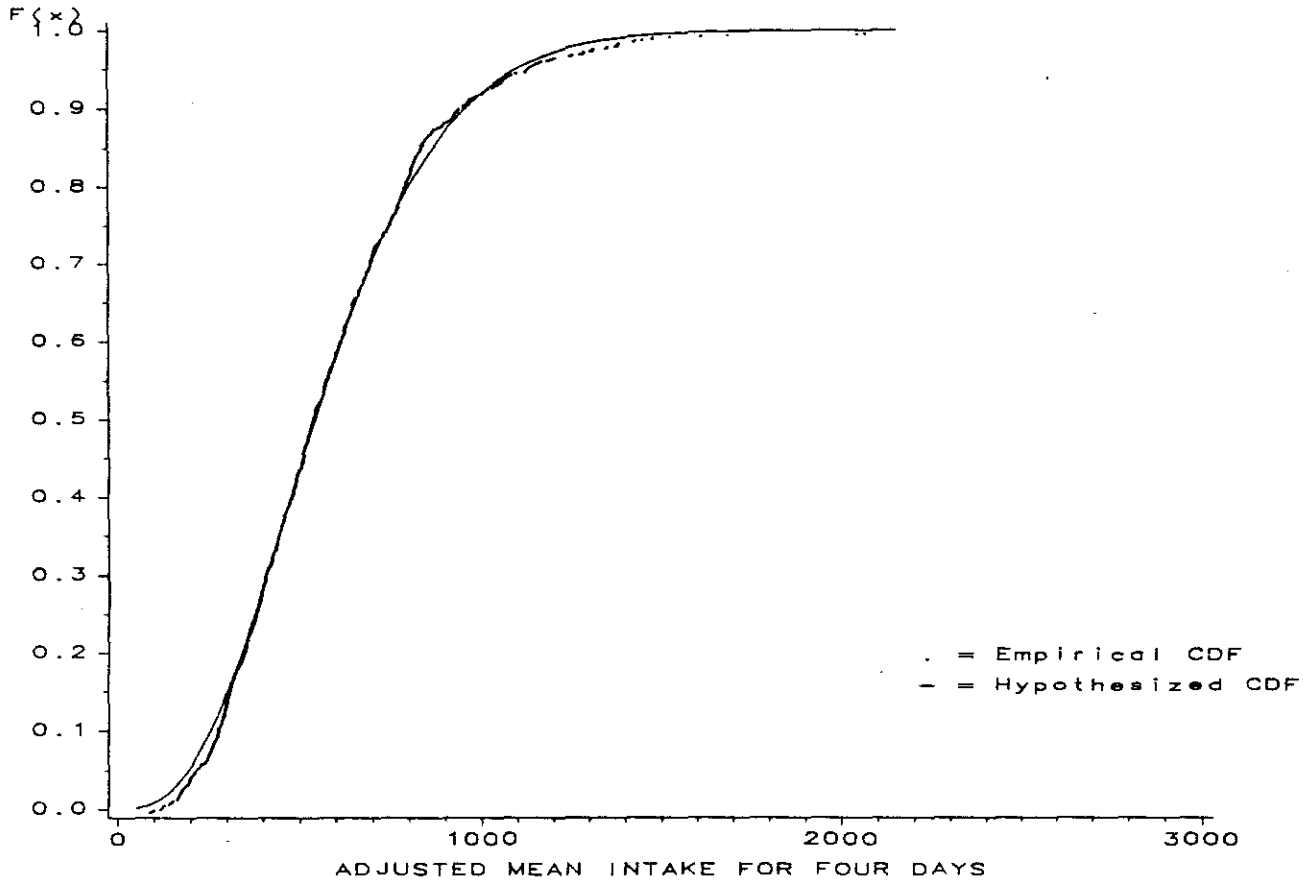

Figure 23. Plots comparing the empirical cdf of individual means with the hypothesized cdf based on Weibull densities for calcium.

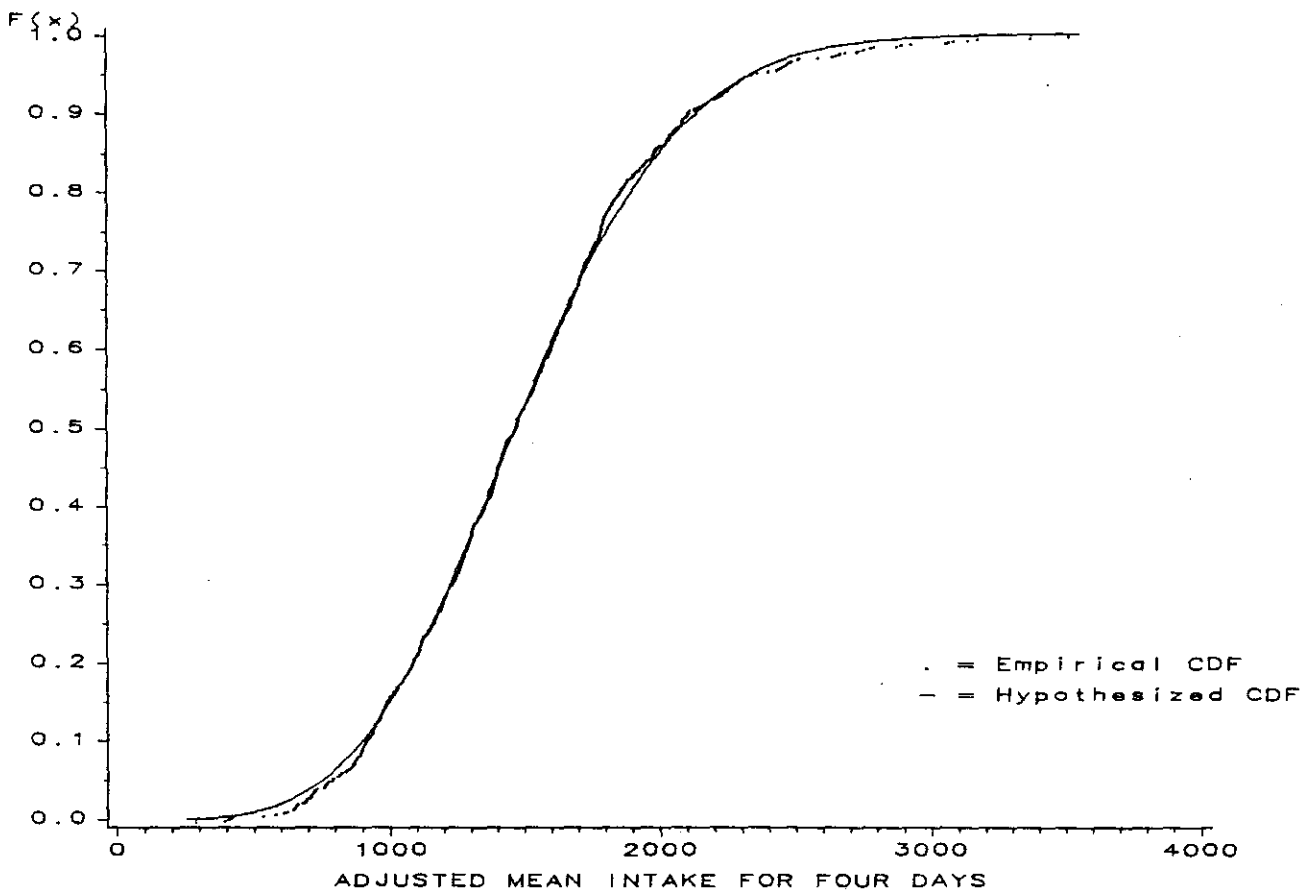

Figure 24. Plots comparing the empirical cdf of individual means with the hypothesized cdf based on Weibull densities for energy.

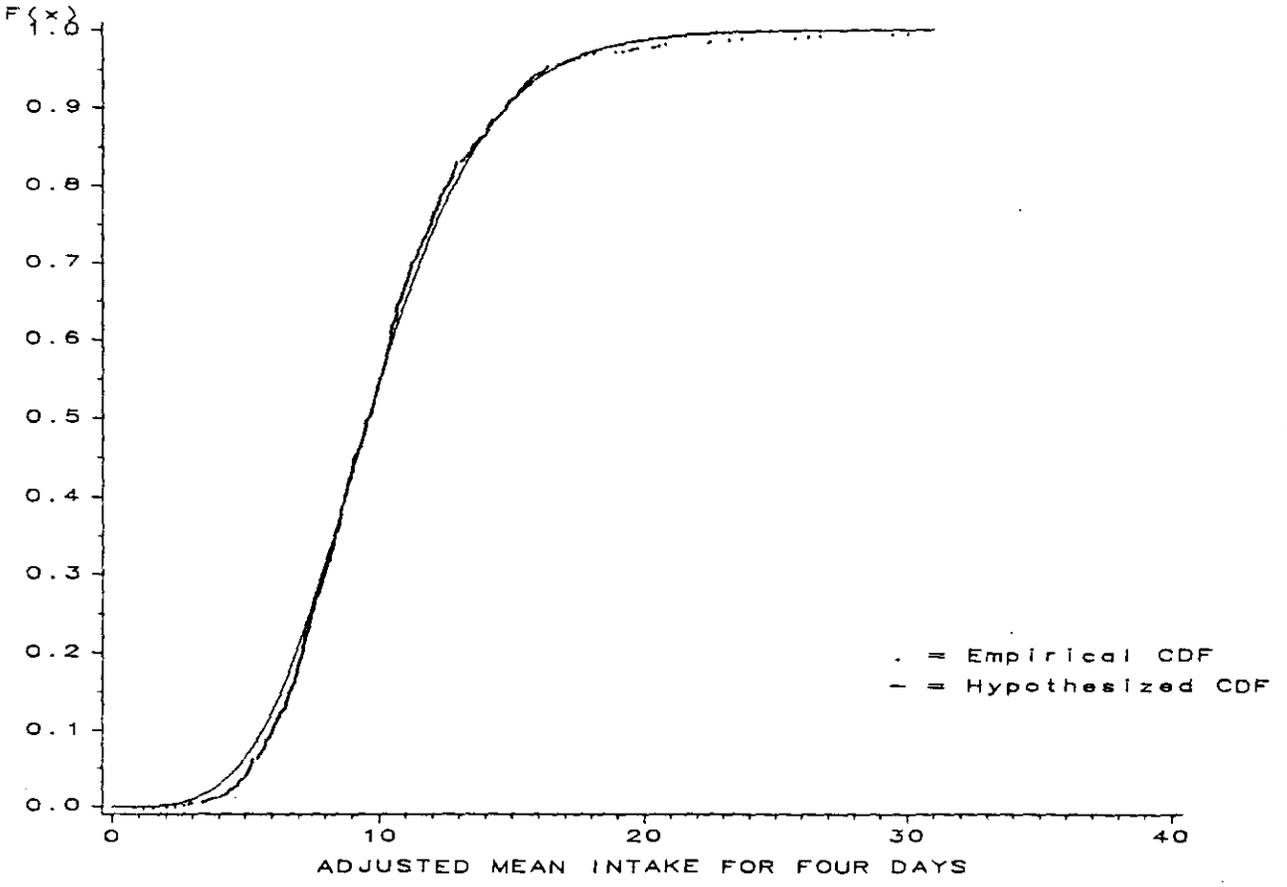

Figure 25. Plots comparing the empirical cdf of individual means with the hypothesized cdf based on Weibull densities for iron.

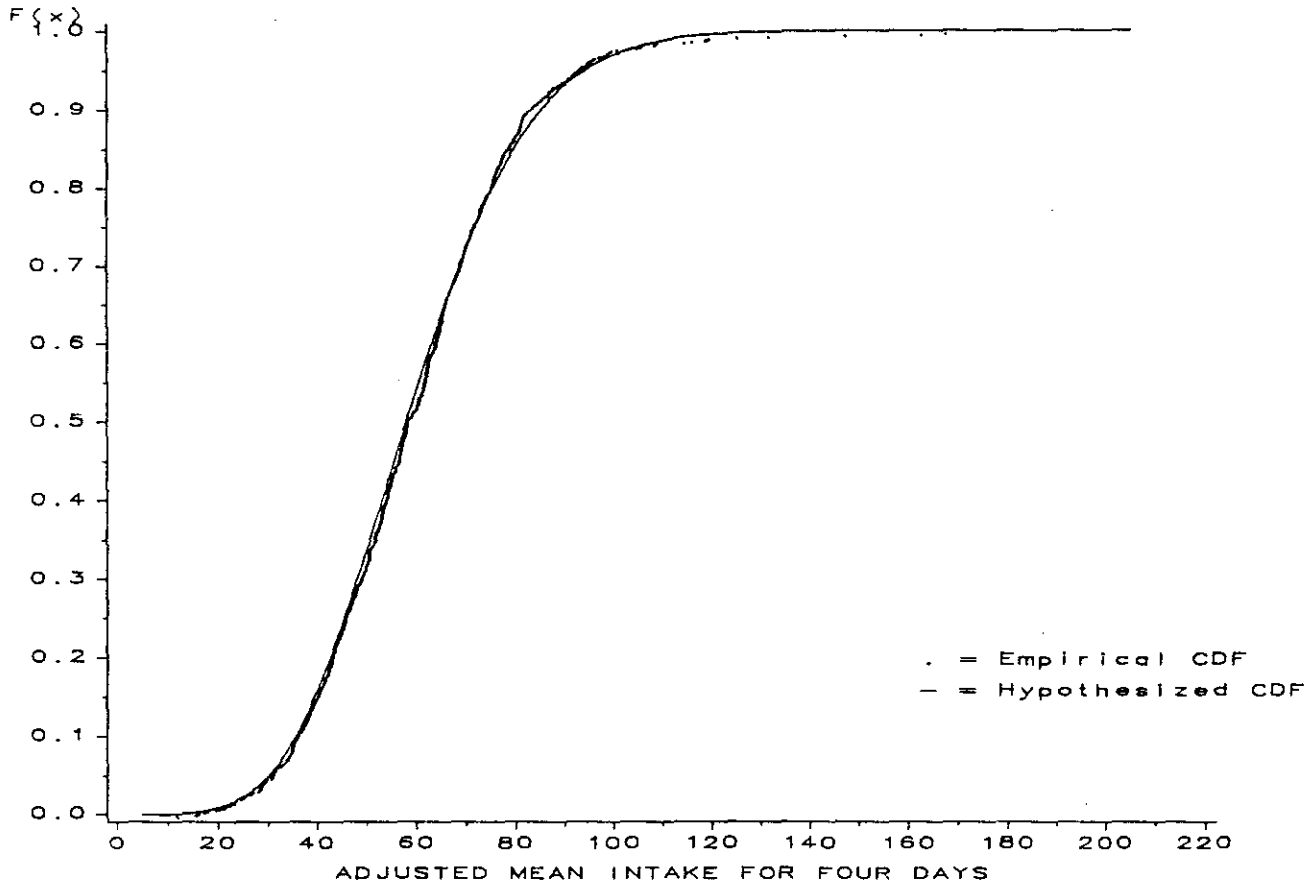

Figure 26. Plots comparing the empirical cdf of individual means with the hypothesized cdf based on Weibull densities for protein.

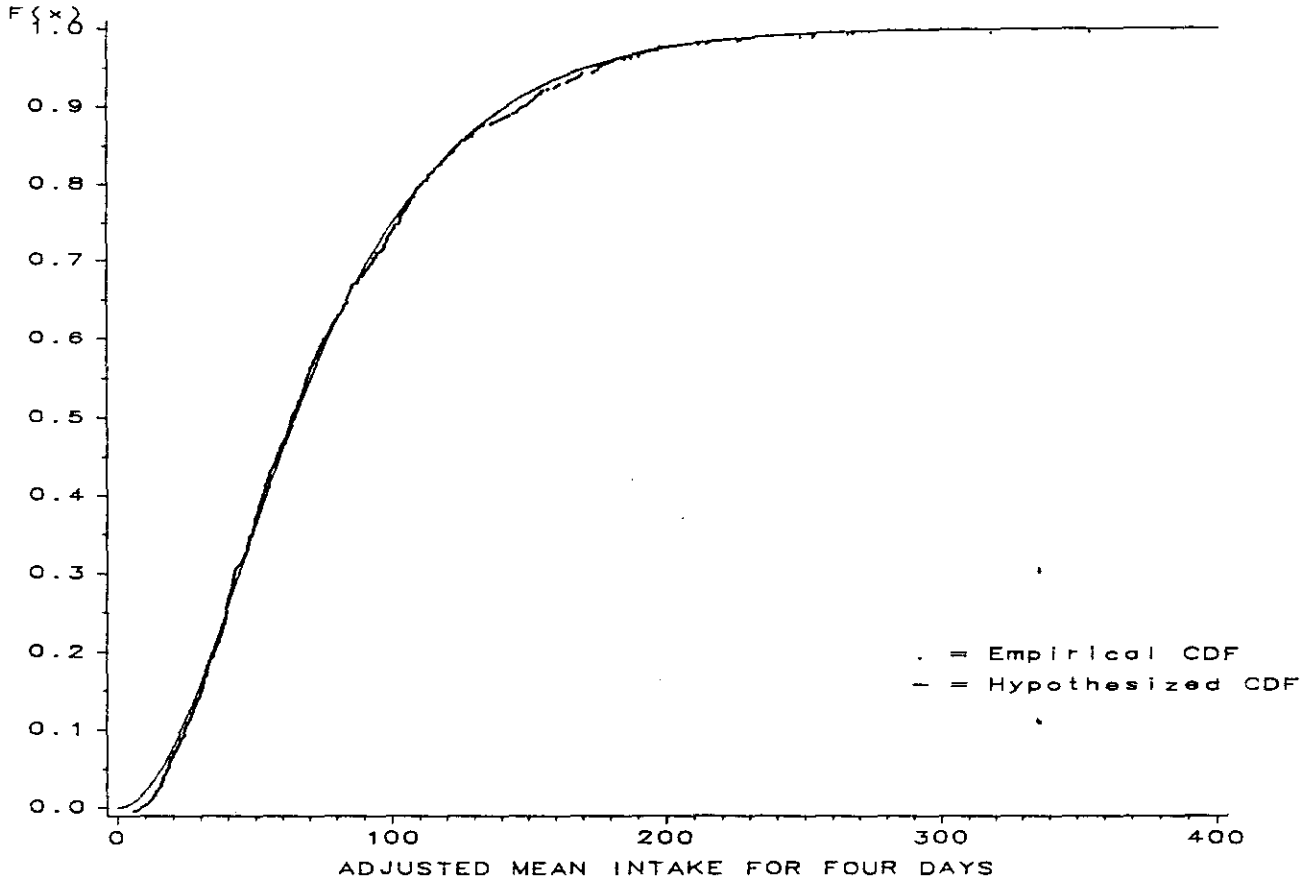

Figure 27. Plots comparing the empirical cdf of individual means with the hypothesized cdf based on Weibull densities for vitamin C.

CONCLUSIONS

The distribution of mean intakes is frequently used to approximate the distribution of usual intakes. However, errors in measurement lead to inflated variance in the distribution of mean intakes relative to the distribution of usual intakes. For the dietary components considered in this paper, the intra-individual (i.e., measurement) variance of individual daily intakes ranged from 64 percent to 74 percent of the total variance of the daily intakes. Clearly, using the distribution of individual mean intakes as a basis for inferences concerning the proportion of individuals with inadequate (or excessive) intakes can lead to serious errors.

To circumvent this problem, we offer a model that decomposes an observed individual intake into the individual's usual intake plus a measurement error, where the standard deviation and cube root of the third moment of the measurement errors for an individual are both linearly related to the usual intake of the individual. Based on data for calcium, energy, iron, protein and vitamin C intakes for women 23-50 years of age, the standard deviation and cube root of the third moment of the measurement errors are each satisfactorily approximated by a constant multiple of the usual daily intakes of the women.

The gamma distribution and the Weibull distribution were employed as models for the distribution of usual intake. On the basis of plots and of formal tests, acceptable distributions for usual intake are:

Calcium: gamma distribution;

Energy: gamma distribution, in squares;

Iron: gamma distribution;

Protein: Weibull distribution or gamma distribution, in squares;

Vitamin C: gamma distribution or Weibull distribution.

The methods developed in this paper provide an approach to estimating the distribution of usual intake that explicitly recognizes that daily intakes for an individual are only an approximation to the individual's usual intake. In addition, the estimation procedure recognizes the fact that the distribution of daily intakes is heavily skewed and that the distribution of usual intakes may also be skewed.

The proposed procedure for the estimation of the distribution of usual intakes for individuals requires a sample of daily intakes of individuals, with multiple daily intakes for at least a subset of the sampled individuals. The repeated sample of daily individual intakes should be spaced in time so that the assumption of conditional independence among the sample daily intakes for a given individual is acceptable. Some of the sampled individuals must provide a number of daily intake records per individual equal to the highest moment of the distribution of individual intake that is to be estimated.

Although the methods of this paper appear to be relatively successful for the dietary components under consideration, there are several directions in which the methods can be extended. Neither the two-parameter gamma family nor the two-parameter Weibull family appear to be sufficiently broad to cover the intake distributions of all dietary components. Therefore, assuming the usual intake distribution to be a member of a wider class of distributions may be useful. The error models of this paper may not be adequate for other dietary components. In particular, preliminary analyses indicate that the model

for the third moment of the errors does not hold for vitamin A. Estimators for the usual intake moments that do not rely upon the error models are being considered for future research.

REFERENCES

Elderton, W. P. and N. L. Johnson 1969. Systems of frequency curves. Cambridge: At the University Press.

Fuller, W. A. 1987. Measurement error models. New York: Wiley.

Fuller, W. A., W. Kennedy, D. Schnell, G. Sullivan, and H. J. Park 1986. PC CARP. Ames, Iowa: Statistical Laboratory, Iowa State University.

Kendall, M. G. and A. Stuart 1969. The advanced theory of statistics, Vol. 1. London: Griffin.

National Research Council 1980. Recommended dietary allowances, Ninth Edition. Washington, D.C.: National Academy of Sciences.

National Research Council 1986. Nutrient adequacy: Assessment using food consumption surveys. Washington, D.C.: National Academy Press.

SAS Institute Inc. 1982. SAS user's guide: Basics, 1982 Edition. Cary, North Carolina:

Schnell, D. and W. A. Fuller 1987. EV CARP. Ames, Iowa: Statistical Laboratory, Iowa State University.

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

APPENDIX

Transformation of Intakes

Given any transformation of the observed intakes, $g(Y_{ij})$, it is always possible to define the decomposition,

$$g(Y_{ij}) = \mu_g + \alpha_i + \epsilon_{ij} ,$$

where $\alpha_i \sim (0, \sigma_\alpha^2)$ is uncorrelated with $\epsilon_{ij} \sim (0, \sigma_\epsilon^2)$ and α_i is the individual effect. Transformations are typically chosen so that $g(Y_{ij})$ is approximately normally distributed. For ϵ_{ij} with positive variance and a nontrivial function, $g(\cdot)$, a problem arises in that the expectation of $g(Y_{ij})$ for an individual is not equal to $g(y_i)$. That is,

$$E\{g(Y_{ij})|i\} \neq g(y_i) .$$

For example, consider the square-root transformation,

$$g(Y_{ij}) = Y_{ij}^{1/2} .$$

Then the usual intake for individual i is

$$y_i \equiv E(Y_{ij}|i)$$

$$= E[(\mu_g + \alpha_i + \epsilon_{ij})^2|i]$$

$$= (\mu_g + \alpha_i)^2 + \sigma_\epsilon^2 \ .$$

Hence,

$$y_i^{1/2} = [(\mu_g + \alpha_i)^2 + \sigma_\epsilon^2]^{1/2}$$

$$> \mu_g + \alpha_i \ .$$

Thus, if the distribution of usual intake is to be estimated, then original observations need to be directly analyzed. The analysis of a transformation of intakes yields estimates for the long-run average of the transformed values. In other words, our analysis is based on the postulate that it is the average daily intake of a dietary component that is of interest, not the average of (say) the logarithm of daily intake.

Moments of the Usual Intakes

We express the first four moments of the usual intakes, y_i , $i=1, 2, \ldots, n$, in terms of moments of $\bar{Y}_{i.} \equiv y_i + \bar{e}_{i.}$ and the parameters, α and γ , of the model (8)-(9). We will make repeated use of the following lemma.

Lemma 1. Let $X_1, X_2, \ldots, X_t$ be a random sample of size t on the random variable, X , where X has finite fourth moment. Then

$$E(\bar{X}) = \mu_X \ ,$$

$$E\Big(\sum_{i=1}^{t} (X_i - \bar{X})^2\Big) = (t-1)\mu_2 \ ,$$

$$E\{\sum_{i=1}^{t}(X_i - \bar{X})^3\} = (t-1)(t-2)t^{-1}\mu_3 ,$$

and

$$E\{\sum_{i=1}^{t}(X_i - \bar{X})^4\} = (t-1)t^{-2}\{(t^2 - 3t + 3)\mu_4 + 3(2t - 3)\mu_2^2\} ,$$

where $\bar{X} \equiv t^{-1}\Sigma_{i=1}^{t}X_i$ and $\mu_k \equiv E\{(X - \mu_X)^k\}$, $k=2, 3, 4$.

Proof. The results are obtained by straightforward algebra, using the decomposition,

$$X_i - \bar{X} = (X_i - \mu_X) - (\bar{X} - \mu_X) , \quad i=1, 2, \ldots, t .$$

Now

$$(X_i - \bar{X})^2 = (X_i - \mu_X)^2 - 2(X_i - \mu_X)(\bar{X} - \mu_X) + (\bar{X} - \mu_X)^2 .$$

Therefore,

$$E(X_i - \bar{X})^2 = \mu_2 - 2(t^{-1})\mu_2 + (t^{-1})\mu_2$$

$$= (t^{-1})(t-1)\mu_2 .$$

Similarly, from

$$(X_i - \bar{X})^3 = (X_i - \mu_X)^3 - 3(X_i - \mu_X)^2(\bar{X} - \mu_X)$$

$$+ 3(X_i - \mu_X)(\bar{X} - \mu_X)^2 - (\bar{X} - \mu_X)^3 ,$$

it follows that

$$E(X_i - \bar{X})^3 = \mu_3 - 3(t^{-1})\mu_3 + 3(t^{-2})\mu_3 - (t^{-2})\mu_3$$

$$= (t^{-2})(t^2 - 3t + 2)\mu_3$$

$$= (t^{-2})(t-2)(t-1)\mu_3 \ .$$

Finally,

$$(X_i - \bar{X})^4 = (X_i - \mu_X)^4 - 4(X_i - \mu_X)^3(\bar{X} - \mu_X) + 6(X_i - \mu_X)^2(\bar{X} - \mu_X)^2$$

$$- 4(X_i - \mu_X)(\bar{X} - \mu_X)^3 + (\bar{X} - \mu_X)^4$$

implies that

$$E(X_i - \bar{X})^4 = \mu_4 - 4(t^{-1})\mu_4 + 6(t^{-2})[\mu_4 + (t-1)(\mu_2)^2]$$

$$- 4(t^{-3})[\mu_4 + 3(t-1)(\mu_2)^2] + (t^{-3})[\mu_4 + 3(t-1)(\mu_2)^2]$$

$$= (t^{-3})\{(t^3 - 4t^2 + 6t - 3)\mu_4 + [6t(t-1) - 9(t-1)](\mu_2)^2\}$$

$$= (t^{-3})(t-1)\{(t^2 - 3t + 3)\mu_4 + 3(2t - 3)(\mu_2)^2\} \ . \qquad []$$

Given the assumptions of the model (7)-(9) it follows from Lemma 1 that

$$E\{\sum_{i=1}^{n}(\bar{Y}_{i.} - \bar{Y}_{..})^2\} = (n-1)\mu_{2\bar{Y}} \ ,$$

$$E\{\sum_{i=1}^{n}(\bar{Y}_{i.} - \bar{Y}_{..})^3\} = n^{-1}(n-1)(n-2)\mu_{3\bar{Y}}$$

and

$$E\{\sum_{i=1}^{n}(\bar{Y}_{i.} - \bar{Y}_{..})^4\} = n^{-2}(n-1)\{(n^2 - 3n + 3)\mu_{4\bar{Y}} + 3(2n - 3)(\mu_{2\bar{Y}})^2\} ,$$

where

$$\mu_{k\bar{Y}} \equiv E(\bar{Y}_{i.} - \mu_{\bar{Y}})^k , \quad k=2, 3, 4 .$$

Now

$$\bar{Y}_{i.} - \mu_{\bar{Y}} = (y_i - \mu_y) + \bar{e}_{i.}$$

and so

$$\mu_{2\bar{Y}} \equiv E(\bar{Y}_{i.} - \mu_{\bar{Y}})^2$$

$$= E\{(y_i - \mu_y)^2 + 2(y_i - \mu_y)\bar{e}_{i.} + \bar{e}_{i.}^2\}$$

$$= \mu_{2y} + r^{-1}E(\alpha y_i^2)$$

$$= \mu_{2y} + r^{-1}\alpha(\mu_{2y} + \mu_y^2) ,$$

i.e.,

$$\mu_{2\bar{Y}} = [1 + r^{-1}\alpha]\mu_{2y} + (r^{-1})\alpha\ \mu_y^2 .$$

Similarly,

$$\mu_{3\bar{Y}} \equiv E(\bar{Y}_{i.} - \mu_{\bar{Y}})^3$$

$$= E\{(y_i - \mu_y)^3 + 3(y_i - \mu_y)^2\bar{e}_{i.} + 3(y_i - \mu_y)\bar{e}_{i.}^2 + \bar{e}_{i.}^3\}$$

$$= \mu_{3y} + 3\,E\{(y_i - \mu_y)[(r^{-1})\alpha\, y_i^2]\} + E[(r^{-2})\gamma\, y_i^3]$$

$$= \mu_{3y} + 3(r^{-1})\alpha[\mu_{3y} + 2\mu_y\mu_{2y}] + (r^{-2})\gamma[\mu_{3y} + 3\mu_y\mu_{2y} + \mu_y^3] ,$$

hence it follows that

$$\mu_{3\bar{Y}} = [1 + 3(r^{-1})\alpha + (r^{-2})\gamma]\mu_{3y} + 3(r^{-2})\mu_y[2r\alpha + \gamma]\mu_{2y} + (r^{-2})\gamma\mu_y^3 .$$

Finally,

$$\mu_{4\bar{Y}} \equiv E(\bar{Y}_{i.} - \mu_{\bar{Y}})^4$$

$$= E\{(y_i - \mu_y)^4 + 4(y_i - \mu_y)^3\bar{e}_{i.} + 6(y_i - \mu_y)^2\bar{e}_{i.}^2 + 4(y_i - \mu_y)\bar{e}_{i.}^3 + \bar{e}_{i.}^4\}$$

$$= \mu_{4y} + 6E\{(y_i - \mu_y)^2(r^{-1})\alpha y_i^2\} + 4E\{(y_i - \mu_y)(r^{-2})\gamma y_i^3\} + E(\bar{e}_{i.}^4)$$

$$= \mu_{4y} + 6r^{-1}\alpha[\mu_{4y} + 2\mu_y\mu_{3y} + \mu_y^2\mu_{2y}]$$

$$+ 4r^{-2}\gamma[\mu_{4y} + 3\mu_y\mu_{3y} + 3\mu_y^2\mu_{2y}] + E(\bar{e}_{i.}^4) .$$

Now

$$E(\bar{e}_{i.}^4) = E\{r^{-3}[E(e_{ij}^4|i) + 3(r-1)E(e_{ij}^2 e_{ij'}^2|i)]\} ,$$

and

$$E(e_{ij}^2 e_{ij'}^2|i) = E(e_{ij}^2|i)E(e_{ij'}^2|i)$$

$$= (\alpha\, y_i^2)(\alpha\, y_i^2) \ .$$

Thus

$$E(\bar{e}_{i.}^4) = r^{-3}[E(e_{ij}^4) + 3(r-1)\alpha^2 E(y_i^4)] \ .$$

However, by Lemma 1, it follows that

$$E\{\sum_{j=1}^{r}(Y_{ij} - \bar{Y}_{i.})^4 | i\} = (r-1)r^{-2}\{(r^2 - 3r + 3)E(e_{ij}^4 | i)$$

$$+ 3(2r - 3)[E(e_{ij}^2 | i)]^2\}$$

and so

$$E\{\sum_{j=1}^{r}(Y_{ij} - \bar{Y}_{i.})^4\} = (r-1)r^{-2}\{(r^2 - 3r + 3)E(e_{ij}^4) + 3(2r - 3)E(\alpha y_i^2)^2\} \ .$$

Hence

$$E(e_{ij}^4) = \{r^2(r-1)^{-1}E[\sum_{j=1}^{r}(Y_{ij} - \bar{Y}_{i.})^4]$$

$$- 3(2r - 3)\alpha^2 E(y_i^4)\}(r^2 - 3r + 3)^{-1} \ .$$

Thus

$$E(\bar{e}_{i.}^4) = (r^{-3})\{[r^2(r-1)^{-1}E[\sum_{j=1}^{r}(Y_{ij} - \bar{Y}_{i.})^4] - 3(2r - 3)\alpha^2 E(y_i^4)]$$

$$\times (r^2 - 3r + 3)^{-1} + 3(r-1)\alpha^2 E(y_i^4)\}$$

$$= [r(r-1)(r^2 - 3r + 3)]^{-1}E[\sum_{j=1}^{r}(Y_{ij} - \bar{Y}_{i.})^4] + 3(r^{-3})(r-1-c)\alpha^2E(y_i^4)$$

where $c \equiv (2r - 3)(r^2 - 3r + 3)^{-1}$. But

$$E(y_i^4) = \mu_{4y} + 4\mu_y\mu_{3y} + 6\mu_y^2\mu_{2y} + \mu_y^4 .$$

Therefore,

$$\mu_{4\bar{Y}} = \mu_{4y} + 6(r^{-1})\alpha[\mu_{4y} + 2\mu_y\mu_{3y} + \mu_y^2\mu_{2y}]$$

$$+ 4(r^{-2})\gamma[\mu_{4y} + 3\mu_y\mu_{3y} + 3\mu_y^2\mu_{2y}]$$

$$+ [r(r-1)(r^2 - 3r + 3)]^{-1}E[\sum_{j=1}^{r}(Y_{ij} - \bar{Y}_{i.})^4]$$

$$+ 3(r^{-3})(r-1-c)\alpha^2[\mu_{4y} + 4\mu_y\mu_{3y} + 6\mu_y^2\mu_{2y} + \mu_y^4]$$

$$= \{1 + 6(r^{-1})\alpha + 4(r^{-2})\gamma + 3(r^{-3})(r-1-c)\alpha^2\}\mu_{4y}$$

$$+ \{12(r^{-3})\mu_y[r^2\alpha + r\gamma + (r-1-c)\alpha^2]\mu_{3y}$$

$$+ \{6(r^{-3})\mu_y^2[r^2\alpha + 2r\gamma + 3(r-1-c)\alpha^2]\}\mu_{2y} + 3(r^{-3})(r-1-c)\alpha^2\mu_y^4$$

$$+ [r(r-1)(r^2 - 3r + 3)]^{-1}E[\sum_{j=1}^{r}(Y_{ij} - \bar{Y}_{i.})^4] .$$

Estimators for Error Model Parameters and Usual Intake Moments

It is easily verified that unbiased estimators for $E(e_{ij}^2|i)$ and $E(e_{ij}^3|i)$ are

$$S_i^2 = (r-1)^{-1} \sum_{j=1}^{r} (Y_{ij} - \bar{Y}_{i.})^2$$

and

$$M_{3i} = r[(r-1)(r-2)]^{-1} \sum_{j=1}^{r} (Y_{ij} - \bar{Y}_{i.})^3 \text{ , respectively.}$$

Also, unbiased estimators for y_i^2 and y_i^3 are $[\bar{Y}_{i.}^2 - r^{-1}S_i^2]$ and $[\bar{Y}_{i.}^3 - 3r^{-1}\bar{Y}_{i.}S_i^2 + 2r^{-2}M_{3i}]$, respectively. Thus method-of-moments estimators for α and γ in the moment models (8) and (9) are

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} S_i^2}{\sum_{i=1}^{n} (\bar{Y}_{i.}^2 - r^{-1}S_i^2)}$$

and

$$\hat{\gamma} = \frac{\sum_{i=1}^{n} M_{3i}}{\sum_{i=1}^{n} [\bar{Y}_{i.}^3 - 3r^{-1}\bar{Y}_{i.}S_i^2 + r^{-2}2M_{3i}]} .$$

The values of the estimators, $\hat{\alpha}$ and $\hat{\gamma}$, and their estimated standard errors can be obtained using PC CARP, a computer program for survey sampling estimation. This program for the personal computer is described in Fuller, et al. (1986).

Method-of-moment estimators for the first four moments of the usual intakes, μ_y, μ_{2y}, μ_{3y} and μ_{4y}, are

$$\hat{\mu}_y = \bar{Y}_{..} ,$$

$$\hat{\mu}_{2y} = (\hat{\mu}_{2\bar{Y}} - r^{-1}\hat{\alpha}\,\hat{\mu}_y^2)[1 + r^{-1}\hat{\alpha}]^{-1} ,$$

$$\hat{\mu}_{3y} = (\hat{\mu}_{3\bar{Y}} - 3(r^{-2})\hat{\mu}_y(2r\hat{\alpha} + \hat{\gamma})\hat{\mu}_{2y} - (r^{-2})\hat{\gamma}\,\hat{\mu}_y^3)[1 + 3(r^{-1})\hat{\alpha} + (r^{-2})\hat{\gamma}]^{-1} ,$$

$$\hat{\mu}_{4y} = (\hat{\mu}_{4\bar{Y}} - [12(r^{-3})\hat{\mu}_y(r^2\hat{\alpha} + r\hat{\gamma} + (r-1-c)\hat{\alpha}^2)]\hat{\mu}_{3y}$$

$$- [6(r^{-3})\hat{\mu}_y^2(r^2\hat{\alpha} + 2r\hat{\gamma} + 3(r-1-c)\hat{\alpha}^2)]\hat{\mu}_{2y} - 3(r^{-3})(r-1-c)\hat{\alpha}^2\hat{\mu}_y^4$$

$$- \frac{1}{n}\sum_{i=1}^{n} M_{4i})[1 + 6(r^{-1})\hat{\alpha} + 4(r^{-2})\hat{\gamma} + 3(r^{-3})(r-1-c)\hat{\alpha}^2]^{-1} ,$$

where

$$\hat{\mu}_{2\bar{Y}} \equiv \frac{1}{n-1}\sum_{i=1}^{n} (\bar{Y}_{i.} - \bar{Y}_{..})^2 ,$$

$$\hat{\mu}_{3\bar{Y}} \equiv \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n} (\bar{Y}_{i.} - \bar{Y}_{..})^3$$

$$\hat{\mu}_{4\bar{Y}} \equiv (\frac{n^2}{n-1}\sum_{i=1}^{n} (\bar{Y}_{i.} - \bar{Y}_{..})^4 - 3(2n-3)(\hat{\mu}_{2\bar{Y}})^2)(n^2 - 3n + 3)^{-1} ,$$

and

$$M_{4i} \equiv \frac{1}{r(r-1)(r^2 - 3r + 3)}\sum_{j=1}^{r} (Y_{ij} - \bar{Y}_{i.})^4 .$$

Pearson Distributions

Karl Pearson introduced a class of distributions whose density functions, $f(\cdot)$, are characterized by the solution of the differential equation,

$$\frac{d\ \ell n\ f(x)}{dx} = \frac{x + a}{b_0 + b_1 x + b_2 x^2} ,$$

where the parameters, a, b_0, b_1 and b_2, are known functions of the first four moments of the random variable involved. The particular solution of the differential equation depends on the nature of the roots of the quadratic equation, $b_0 + b_1 x + b_2 x^2 = 0$. This depends on the value of the parameter, κ, called "the criterion," which is defined by

$$\kappa = b_1^2/(4b_0 b_2) .$$

Given the expression for the parameters of the quadratic equation, in terms of the first four moments (see Elderton and Johnson 1969, p.39) the criterion is

$$\kappa = \frac{\beta_1(\beta_2 + 3)^2}{4(2\beta_2 - 3\beta_1 - 6)(4\beta_2 - 3\beta_1)} ,$$

where $\beta_1 = \mu_3^2/\mu_2^3$ and $\beta_2 = \mu_4/\mu_2^2$.