A Technical Guide to C-SIDE

Software for Intake Distribution Estimation

Kevin W. Dodd

Technical Report 96-TR 32

Dietary Assessment Research Series Report 9

September 1996

Department of Statistics and

Center for Agricultural and Rural Development

(CARD)

Iowa State University

Kevin W. Dodd is a graduate research associate in the Department of Statistics, Iowa State University

Funding support for the development of the C-SIDE software came from Cooperative Agreement No. 58-3198-2-006 between Iowa State University and the Agricultural Research Service, U.S. Department of Agriculture.

Preface

The computer software package C-SIDE estimates usual intake distributions for nutrients and foods using statistical methodology developed at Iowa State University. The manual that accompanies the software, *A User's Guide to C-SIDE* (Department of Statistics and CARD 1996), explains how to use C-SIDE to obtain estimates of usual intake distributions. This document provides a detailed description of the statistical methodology itself.

Several papers describing the theory and application of the methodology have appeared in the statistical literature, including Nusser et al. (1996a) and (1996b), but this document is the first to fully describe every step of the method, including the derivation of many results mentioned only briefly in other sources. The notation in this document may differ slightly from that used in other sources. The algorithms documented in this report are those implemented by C-SIDE Version 1.0.

Todd Krueger helped with the documentation of the algorithms in Chapter 5, and Wayne Fuller provided invaluable editing support. Judy Shafer and Sherrie Martinez performed the initial typesetting and formatting.

Contents

List of Tables

A TECHNICAL GUIDE TO C-SIDE

Introduction

The U.S. Department of Agriculture conducts surveys to assess the dietary adequacy of the population. An important concept in analyzing data from these surveys is that of *usual intake*, defined as the long-run average daily intake of a dietary component. To estimate distributions of usual intake, surveys collect daily intake measurements on individuals for a small number of days. Due to the small number of observations per individual, the distribution of individual mean intakes performs poorly as an estimate of the distribution of usual intakes. This is because the variance of the mean of a few daily intakes contains a sizable amount of within-individual variation. Assuming that daily intakes of a dietary component for an individual measure the individual's usual intake with error, the problem of estimating the distribution of usual intakes can be thought of as the problem of estimating the distribution of a random variable that is observed subject to measurement error. Once an estimator of the usual intake distribution is obtained, it yields estimates of usual intake moments, usual intake quantiles, and the proportion of the population with usual intake below a specified level.

Several characteristics of dietary intake data make statistical analysis difficult. Intake data are nonnegative, and the distributions of both daily intakes and individual mean intakes are often highly skewed. Nuisance effects are often present in the data; daily consumption patterns differ according to day-of-week and month of year. Within-individual variances may vary across individuals, suggesting that the measurement error variance is not constant. Nusser et al. (1996a) propose a method that combines power transformations and nonparametric regression splines to estimate usual intake distributions of dietary components such as nutrients, which are consumed on a daily basis. For infrequently consumed dietary components such as foods, Nusser et al. (1996b) extend the method to settings in which the data arise from a mixture of consumers and nonconsumers. The single-valued nonconsumer distribution is always zero and the consumer distribution is continuous,

but not necessarily normal. The computer software package C-SIDE was developed to implement the methodology detailed in this document.

The method for estimating usual intake distributions for nutrients has several steps. Daily intake data are adjusted for nuisance effects, then the intake data for each sample day are adjusted to have common mean and variance. Chapter 1 describe these preliminary adjustments. The adjusted daily intake data are transformed to normality as described in Chapter 2. The transformed observed intake data are assumed to follow a measurement error model, and the normal distribution methods of Sections 3.1 and 3.2 are used to estimate the parameters of the model. A transformation that carries the normal usual intake distribution back to the original scale is detailed in Section 3.3. The back transformation of the fitted normal distribution adjusts for the bias associated with a nonlinear transformation. The back transformation is used to define the distribution of usual intakes in the original scale, and is used in conjunction with the estimated measurement error model to obtain the estimates (Section 3.4) and associated standard errors (Chapter 4) for quantiles and cumulative distribution function values of the usual intake distribution. The estimated density function of usual intake is obtained from the back transformation as described in Section 3.5.

The method for estimating usual intake distributions for foods in C-SIDE requires an individual's usual intake to be unrelated to the individual's probability of consumption. A test for correlation between intake and probability of consumption is described in Section 5.1. Under the independence assumption, the usual intake for an individual is modeled as the individual's usual intake on days that the food is consumed multiplied by the individual's probability of consuming the food on any given day. The method for estimating usual intake distributions for nutrients is applied to the positive food intakes to estimate a consumption day usual intake distribution for the population. The estimation of the distribution of the probability of consumption is described in Section 5.2. The joint distribution of consumption day usual intakes and consumption probabilities is used to derive the usual intake distributions over all days for consumers and for the entire population. Section 5.3 gives the derivation and describes the estimation of moments, percentiles, and density functions for food intake distributions.

Chapter 1

Preliminary Adjustments

1.1 The Data

The C-SIDE software analyzes a data set consisting of daily intake observations recorded on each of n individuals. Let the observations be denoted by $Y_{ij}^{(o)}$, $i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, k_i$, where k_i, $k_i \leq k$, denotes the number of observations for individual i. Each individual has a sampling weight $W_i,\ i = 1, 2, \ldots, n$. If the data are assumed to be equally weighted, each W_i is 1. Two sets of observation weights are created from the individual weights W_i as

$$\widetilde{W}_{ij} = W_i \ ,$$
$$\widetilde{w}_{ij} = k_i^{-1} W_i \ ,$$

for $\quad j = 1, 2, \ldots, k_i$. Both sets of weights are normalized to sum to unity.

$$W_{ij} \quad = \quad \left(\sum_{i=1}^{n} \sum_{j=1}^{k_i} \widetilde{W}_{ij} \right)^{-1} \widetilde{W}_{ij} \ , \tag{1.1}$$

$$w_{ij} \quad = \quad \left(\sum_{i=1}^{n} \sum_{j=1}^{k_i} \widetilde{w}_{ij} \right)^{-1} \widetilde{w}_{ij} \ . \tag{1.2}$$

1.2 Summary Statistics

C-SIDE applies several preliminary smoothing procedures to the intake data. After each smoothing procedure is complete, C-SIDE outputs some simple descriptive statistics for the

smoothed intake data denoted below by Y_{ij}. The weights W_{ij} from (1.1) are used to compute the quantities

$$\begin{aligned}
N &= \sum_{i=1}^{n} k_i \ , \\
\bar{Y} &= \left(\sum_{i=1}^{n} \sum_{j=1}^{k_i} W_{ij} \right)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{k_i} W_{ij} Y_{ij} \ , \\
s^2 &= \left(\sum_{i=1}^{n} \sum_{j=1}^{k_i} W_{ij} \right)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{k_i} W_{ij} (Y_{ij} - \bar{Y})^2, \\
s &= \sqrt{s^2} \ , \\
m_3 &= s^{-3} \left(\sum_{i=1}^{n} \sum_{j=1}^{k_i} W_{ij} \right) \sum_{i=1}^{n} \sum_{j=1}^{k_i} W_{ij} (Y_{ij} - \bar{Y})^3, \\
m_4 &= s^{-4} \left(\sum_{i=1}^{n} \sum_{j=1}^{k_i} W_{ij} \right) \sum_{i=1}^{n} \sum_{j=1}^{k_i} W_{ij} (Y_{ij} - \bar{Y})^4 - 3 \ ,
\end{aligned}$$

If the observations Y_{ij} are equally weighted, the following formulas are used.

$$\begin{aligned}
N &= \sum_{i=1}^{n} k_i \ , \\
\bar{Y} &= N^{-1} \sum_{i=1}^{n} \sum_{j=1}^{k_i} Y_{ij} \ , && (1.3) \\
s^2 &= N^{-1} \sum_{i=1}^{n} \sum_{j=1}^{k_i} (Y_{ij} - \bar{Y})^2, && (1.4) \\
s &= \sqrt{s^2} \ , \\
m_3 &= s^{-3} \left(N^2 - 3N + 2 \right)^{-1} N \sum_{i=1}^{n} \sum_{j=1}^{k_i} (Y_{ij} - \bar{Y})^3, && (1.5) \\
m_4 &= s^{-4} \left(N^3 - 6N^2 + 11N - 6 \right)^{-1} \left(N^2 + N \right) \sum_{i=1}^{n} \sum_{j=1}^{k_i} (Y_{ij} - \bar{Y})^4 \\
&\quad -3 \left(N^2 - 5N + 6 \right)^{-1} (N-1)^2 \ , && (1.6)
\end{aligned}$$

The quantities m_3 and m_4 are estimates of skewness and kurtosis, respectively. The formulas (1.5) and (1.6) are found in Fisher (1963).

1.3 Normal Scores

C-SIDE uses a variant of the Blom (1958) normal scores in the construction of semiparametric normality transformations. For a set of N ordered observations $Y_{(1)}, Y_{(2)}, \ldots, Y_{(N)}$, where $Y_{(1)} \leq Y_{(2)} \leq \ldots \leq Y_{(N)}$, a set of normal scores is constructed by computing

$$Z_{(i)} = \Phi^{-1}\left(\frac{i - \frac{3}{8}}{N + \frac{1}{4}}\right) \tag{1.7}$$

for $i = 1, 2, \ldots, N$, where Φ^{-1}denotes the inverse of the standard normal cumulative distribution function. The first and last two $Z_{(i)}$ are multiplied by a constant so that the first five sample moments of the $Z_{(i)}$ closely match the first five theoretical moments of the normal distribution. In the general case, the constant 1.04 is used. In the specific case $N = 400$ (see Sections 3.3 and 5.3) the constant is 1.0448.

1.4 Equal Weight Sample

Consider the intake data Y_{ij} and associated sampling weights w_{ij} from (1.2). The N' distinct observations in the sample of intakes are ranked and re-indexed. The distinct values are denoted by $Y_{(k)}$, $k = 1, 2, \ldots, N'$. The corresponding weights $w_{(k)}$ for the distinct values are

$$w_{(k)} = \sum_{\{Y_{ij} = Y_{(k)}\}} w_{ij} \ .$$

The empirical cumulative distribution function $\widetilde{F}$, defined at $N' + 2$ points, is

$$\begin{aligned} \widetilde{F}\left(Y_{(k)}\right) &= \sum_{i<k} w_{(i)} + \frac{1}{2} w_{(k)} \ , \\ \widetilde{F}^{-1}(0) &= \max\left(Y_{(1)} - \frac{Y_{(2)} - Y_{(1)}}{\widetilde{F}\left(Y_{(2)}\right) - \widetilde{F}\left(Y_{(1)}\right)} \widetilde{F}\left(Y_{(1)}\right), 0\right) \ , \end{aligned}$$

$$\widetilde{F}^{-1}(1) \;=\; Y_{(N')} + \frac{Y_{(N')} - Y_{(N'-1)}}{\widetilde{F}\left(Y_{(N')}\right) - \widetilde{F}_{(N'-1)}} \left(1 - \widetilde{F}\left(Y_{(N')}\right)\right) .$$

Let $\widetilde{F}(y)$ be the function that is linear between the $N'+2$ points. Thus, $\widetilde{F}$ is obtained from the usual step-function empirical cumulative distribution function $\hat{F}$ by connecting the midpoints of the rises of adjacent steps with straight lines. The lines passing through the first and last steps are extended to 0 and 1, respectively. Let $N = \sum_{i=1}^{n} k_i$. An equal weight sample $\{Y_i^{(e)}\}_{i=1}^{N}$ is constructed by inverting $\widetilde{F}$ at the points $\{p_i\}_{i=1}^{N}$, where $p_i = N^{-1}(i - 0.5)$. The equal weight sample has the property that the empirical cumulative distribution function $\hat{F}$ constructed from the N distinct values $Y_i^{(e)}$, using weights N^{-1} for each i, is essentially the same as $\widetilde{F}$ computed from the original, weighted data.

1.5 Power Transformation Selection

A power or log transformation is applied to intake data so that the transformed data appears roughly normally-distributed. To avoid taking logarithms of observed zero intakes, a small fraction of the (weighted) mean intake is added to each observation before any power transformations are applied. See Equation (1.11). Given an equal weight sample of strictly positive observations $\{Y_{(i)}\}_{i=1}^{N}$, ordered from smallest to largest, a value of α is chosen to minimize

$$\sum_{i=1}^{N} \left(Z_{(i)} - \hat{\beta}_0 - \hat{\beta}_1 Y_{(i)}^{\alpha}\right)^2 , \tag{1.8}$$

where $\{Z_{(i)}\}_{i=1}^{N}$ are the normal scores defined in Section 1.3, $\hat{\beta}_0$ and $\hat{\beta}_1$ are the usual least-squares estimates for the intercept and slope in the regression of $Z_{(i)}$ on $\mathrm{Y}_{(i)}^{\alpha}$, and $\alpha \in \{0, 1, 1.5^{-1}, 2^{-1}, 2.5^{-1}, \ldots, (R-0.5)^{-1}, R^{-1}\}$. The value of R can be any positive integer greater than one, and is typically taken to be 10. Note that the $\left\{Z_{(i)}\right\}_{i=1}^{N}$ have sample mean zero by construction. Formulas for

$\hat{\beta}_0$ and $\hat{\beta}_1$ are

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{N} Z_{(i)} \left(Y_{(i)}^{\alpha} - N^{-1} \sum_{i=1}^{N} Y_{(i)}^{\alpha} \right)}{\sum_{i=1}^{N} \left(Y_{(i)}^{\alpha} - N^{-1} \sum_{i=1}^{N} Y_{(i)}^{\alpha} \right)^2} , \\ \hat{\beta}_0 &= -\hat{\beta}_1 \left(N^{-1} \sum_{i=1}^{N} Y_{(i)}^{\alpha} \right) . \end{aligned} \tag{1.9}$$

From well-known results from least-squares theory, an equivalent form of (1.8) is given by

$$\sum_{i=1}^{N} \left(Z_{(i)} - \hat{\beta}_0 - \hat{\beta}_1 Y_{(i)}^{\alpha} \right)^2 = \sum_{i=1}^{N} Z_{(i)}^2 - \hat{\beta}_1^2 \sum_{i=1}^{N} \left(Y_{(i)}^{\alpha} - N^{-1} \sum_{i=1}^{N} Y_{(i)}^{\alpha} \right)^2 . \tag{1.10}$$

In the case $\alpha = 0, Y_{(i)}^{\alpha}$ should be taken to mean $\ln(Y_{(i)})$, where $\ln(\cdot)$ denotes the natural logarithm transformation.

Once the best power α is selected, a scale factor β is chosen so that the values of $(Y_{ij}^{\alpha}) \times 10^{-\beta}$ are not too large in absolute value. The value of β is

$$\beta = \begin{cases} \left\lfloor \log_{10} \left\{ \max \left(\left| Y_{(1)}^{\alpha} \right| , \left| Y_{(N)}^{\alpha} \right| \right) \right\} \right\rfloor & \text{if } \alpha \neq 0 , \\ \left\lfloor \log_{10} \left\{ \max \left(\left| \ln Y_{(1)} \right| , \left| \ln Y_{(N)} \right| \right) \right\} \right\rfloor & \text{if } \alpha = 0 , \end{cases}$$

where $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x. The scaled values $(Y_{ij}^{\alpha}) \times 10^{-\beta}$ are used in the remaining calculations.

1.6 Ratio-adjustment to Remove Nuisance Effects

Consider the sample of original observations $Y_{ij}^{(o)}$ and weights W_{ij} described in Section 1.1. The quantity

$$\delta = \varepsilon \sum_{i=1}^{n} \sum_{j=1}^{k_i} W_{ij} Y_{ij} \tag{1.11}$$

is added to each of the $Y_{ij}^{(o)}$, where ε is a small positive number, typically 0.0001. Denote these shifted observations by $Y_{ij}^{(s)}$. Suppose that the data set being analyzed contains observations taken on variables representing nuisance effects, such as day of week, interview sequence, or

metabolic rate. Such nuisance effects may be either discrete or continuous. Let $V_{1ij}, V_{2ij}, \ldots, V_{Dij}$ and $U_{1ij}, U_{2ij}, \ldots, U_{Cij}$ denote the collections of discrete and continuous nuisance variables, respectively. Also, let $l_1, l_2, \ldots, l_D$ denote the number of levels for each of the discrete variables.

Because all of the preliminary smoothing procedures are applied to power-transformed data, a best power α and scale factor β are chosen as described in Section 1.5 before the ratio-adjustment is performed. The power α is applied to the original $Y_{ij}^{(s)}$. Let

$$X_{ij}^{(s)} = \begin{cases} \left(\mathrm{Y}_{ij}^{(s)}\right)^{\alpha} \times 10^{-\beta} & \text{if } \alpha \neq 0 \,, \\ \ln\left(Y_{ij}^{(s)}\right) \times 10^{-\beta} & \text{if } \alpha = 0 \,, \end{cases} \tag{1.12}$$

for $i = 1, \ldots, n, \quad j = 1, \ldots, k_i$. The ratio-adjustment of the $\{X_{ij}^{(s)}\}$ for the effects of $\{V_{1ij}, \ldots, V_{Dij}, U_{1ij}, \ldots, U_{Cij}\}$ is as follows.

1. Let $\mathbf{D}_k = [\mathbf{d}_{2k}\ \mathbf{d}_{3k}\ \ldots\ \mathbf{d}_{l_k k}]$ for $k = 1, 2, \ldots, D$, where for $m = 2, 3, \ldots, l_k$, $\mathbf{d}_{mk}$ is a column vector with elements

$$d_{mkij} = \begin{cases} 1 & \text{if } V_{kij} \text{ is at level } m \,, \\ 0 & \text{if } V_{kij} \text{ is not at level } m \,, \end{cases}$$

   That is, $\mathbf{D}_k$ is a full-rank design matrix for the categorical variable V_{kij}.

2. Let $\mathbf{1}$ denote a column vector of ones, and $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_C$ denote the column vectors of observations for the continuous variables $U_{1ij}, U_{2ij}, \ldots, U_{Cij}$. Write

$$\mathbf{M} = [\mathbf{1}\ \mathbf{D}_1\ \mathbf{D}_2\ \ldots\ \mathbf{D}_D\ \mathbf{U}_1\ \mathbf{U}_2\ \ldots\ \mathbf{U}_C] \,. \tag{1.13}$$

   The matrix $\mathbf{M}$ is N rows by $C + 1 + \sum_{k=1}^{D} (l_k - 1)$ columns. The matrix $\mathbf{M}$ need not be of full column rank.

3. A weighted least-squares regression with model matrix $\mathbf{M}$ and response variable $\{X_{ij}^{(s)}\}$ is performed, where the weights in the regression are the W_{ij} from (1.1). The predicted values $\hat{X}_{ij}^{(s)}$ from the regression are

$$\hat{\mathbf{X}}^{(s)} = \mathbf{M}\hat{\boldsymbol{\beta}} \,, \tag{1.14}$$

where

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{M}' \operatorname{diag}(W_{ij})\, \mathbf{M}\right)^{-} \mathbf{M}' \operatorname{diag}(W_{ij})\, \mathbf{X}^{(s)}, \tag{1.15}$$

$\mathbf{X}^{(s)}$ is the column vector of observations $X_{ij}^{(s)}$, $\hat{\mathbf{X}}^{(s)}$ is the column vector of predicted values $\hat{X}_{ij}^{(s)}$, and $\mathbf{A}^{-}$ denotes the generalized inverse of $\mathbf{A}$.

4. Let $\bar{X}_1$ denote the weighted average of the observations taken in the first time period of the survey. It may be that not all of the n individuals represented in the survey had a valid observation in the first time period, in which case fewer than n observations are used to compute $\bar{X}_1$. Then

$$\bar{X}_1 = \left(\sum_{i \in I} W_{i1}\right)^{-1} \sum_{i \in I} W_{i1} X_{i1}^{(s)} \,,$$

where $I \equiv \{i : \text{individual } i \text{ has a valid observation in period 1}\}$.

5. The ratio-adjusted observations $X_{ij}^{(r)}$ are

$$X_{ij}^{(r)} = \bar{X}_1 \left(\frac{X_{ij}^{(s)}}{\hat{X}_{ij}^{(s)}}\right). \tag{1.16}$$

6. The ratio-adjusted intakes in the original scale are

$$Y_{ij}^{(r)} = \begin{cases} \max\left(0, \left(X_{ij}^{(r)} \times 10^{\beta}\right)^{\frac{1}{\alpha}} - \delta\right) & \text{if } \alpha \neq 0 \,, \\ \max\left(0, \exp\left(X_{ij}^{(r)} \times 10^{\beta}\right) - \delta\right) & \text{if } \alpha = 0 \,, \end{cases} \tag{1.17}$$

for $i = 1, \ldots, n,\ j = 1, \ldots, k_i$, where δ is the quantity in (1.11).

If it is desired to ratio-adjust to the grand mean rather than to the period one mean, let the grand mean be

$$\bar{X}_{\cdot\cdot} = \left(\sum_{i=1}^{n} \sum_{j=1}^{k_i} W_{ij}\right)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{k_i} W_{ij} X_{ij}^{(s)}.$$

Then the $X_{ij}^{(r)}$ of (1.16) are

$$X_{ij}^{(r)} = \bar{X}_{\cdot\cdot} \left(\frac{X_{ij}^{(s)}}{\hat{X}_{ij}^{(s)}}\right)$$

and the $Y_{ij}^{(r)}$ are obtained from the $X_{ij}^{(r)}$ as in (1.17).

1.7 Regression for Individual and Period Effects

This section describes the computations of a regression that will be used to adjust for time-in-sample (period) effects. The adjustment is described in the next section.

Given data y_{ij} and weights $w_{ij},\ i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, k_i$, let $k = \max(k_i)$ and consider the regression model

$$y_{ij} = \alpha_i + \beta_j + \varepsilon_{ij} \ , \tag{1.18}$$

where α_i is the effect due to individual i , and β_j is the effect of taking an observation in the j^{th} time period (day), with the restriction $\beta_1 = 0$. Furthermore, assume that the ε_{ij} have mean zero. Equation (1.18) can be written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \ , \tag{1.19}$$

where $\mathbf{Y}$ and $\boldsymbol{\varepsilon}$ are column vectors of y_{ij} and ε_{ij},

$$\boldsymbol{\beta} = [\alpha_1, \alpha_2, \ldots, \alpha_n, \beta_2, \beta_3, \ldots, \beta_k]' \ ,$$

and the row of $\mathbf{X}$ corresponding to y_{ij} has a one in column i, and, if $j > 1$, another one in column $n+j-1$. Every other entry in $\mathbf{X}$ is zero. Let $\mathbf{W} = \text{diag}(w_{ij})$. It is desired to compute the weighted least-squares estimate of $\boldsymbol{\beta}$,

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y} \ . \tag{1.20}$$

The inversion of $\mathbf{X}'\mathbf{W}\mathbf{X}$ is not feasible, due to the high dimension $(n + k - 1)$ of $\mathbf{X}$. By taking advantage of the special structure of the $\mathbf{X}$ matrix and results concerning inverses of partitioned matrices, a solution is easily derived.

Write

$$\mathbf{X}'\mathbf{W}\mathbf{X} = \begin{bmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}' & \mathbf{D} \end{bmatrix}, \quad \mathbf{X}'\mathbf{W}\mathbf{Y} = \begin{bmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{bmatrix} \ , \tag{1.21}$$

where

$$\begin{aligned}
\mathbf{A} &= \operatorname{diag}(w_{i\cdot}) \quad i = 1, 2, \ldots, n \ , \\
\mathbf{D} &= \operatorname{diag}(w_{\cdot j}) \quad j = 2, 3, \ldots, k \ , \\
\mathbf{C} &= \{C_{ij}\} = \begin{cases} w_{ij} & \text{if individual } i \text{ had intake on day } j \ , \\ 0 & \text{otherwise} \ , \end{cases} \\
\mathbf{Z}_1 &= \left[\sum_{j=1}^{k_i} w_{ij} Y_{ij} \right]_{i=1}^{n} \ , \\
\mathbf{Z}_2 &= \left[\sum_{i=1}^{n} w_{ij} Y_{ij} \right]_{j=2}^{k} \ .
\end{aligned}$$

The dot subscript indicates summation over the dotted index. For nonsingular $\mathbf{X}'\mathbf{W}\mathbf{X}$,

$$\begin{bmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}' & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} -\mathbf{A}^{-1}\mathbf{C} \\ \mathbf{I} \end{bmatrix} \left(\mathbf{D} - \mathbf{C}'\mathbf{A}^{-1}\mathbf{C} \right)^{-1} \begin{bmatrix} -\mathbf{C}'\mathbf{A}^{-1} & \mathbf{I} \end{bmatrix} . \tag{1.22}$$

Note that $\mathbf{A}^{-1} = \operatorname{diag}(w_{i\cdot}^{-1})$, and $\mathbf{D}$ is typically of much lower order than $\mathbf{A}$. Hence, inversion of $(\mathbf{D} - \mathbf{C}'\mathbf{A}^{-1}\mathbf{C})$ is easily performed.

Let $\mathbf{U}$ be the Cholesky root of $(\mathbf{D} - \mathbf{C}'\mathbf{A}^{-1}\mathbf{C})^{-1}$, i.e. $\mathbf{U}'\mathbf{U} = (\mathbf{D} - \mathbf{C}'\mathbf{A}^{-1}\mathbf{C})^{-1}$, and let

$$\mathbf{Q}' = \begin{bmatrix} -\mathbf{A}^{-1}\mathbf{C} \\ \mathbf{I} \end{bmatrix} \mathbf{U}^{'},$$

so that

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y} = \left(\begin{bmatrix} \mathbf{A}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \mathbf{Q}'\mathbf{Q} \right) \begin{bmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{A}^{-1}\mathbf{Z}_1 \\ \mathbf{0} \end{bmatrix} + \mathbf{Q}'\mathbf{Q} \begin{bmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{A}^{-1}\mathbf{Z}_1 \\ \mathbf{0} \end{bmatrix} + \mathbf{Q}' \left(\mathbf{Q} \begin{bmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{bmatrix} \right) \ .
\end{aligned} \tag{1.23}$$

In the formulation of the model (1.19), the overall mean response (adjusted for individual and period effects) is not directly estimable. An intuitively reasonable estimate for the overall mean is given by

$$\hat{\mu} = \left(\sum_{i=1}^{n} w_{i\cdot} \right)^{-1} \sum_{i=1}^{n} w_{i\cdot} \hat{\alpha}_i \ , \tag{1.24}$$

where the α_i are the first n elements of $\boldsymbol{\beta}$.

1.8 Adjustment to Remove Period Effects

The estimation of usual intake distributions requires that the distribution of intakes for the first time period be the same as that of intakes for the second and subsequent periods. The ratio-adjusted intakes $Y_{ij}^{(r)}$ defined in (1.17) are further adjusted so that the mean and variance for each period's observations are the same, where period indexes the interview number. Period one is the first day-of-interview, period two is the second day-of-interview, etc.

The initial shift, power, and scale factors δ, α, and β used in the ratio-adjustment procedure (Section 1.6) are applied to the $Y_{ij}^{(r)}$ to get

$$X_{ij}^{(r)} = \begin{cases} 10^{-\beta} \left(Y_{ij}^{(r)} + \delta \right)^{\alpha} & \text{if } \alpha \neq 0 \ , \\ 10^{-\beta} \ln \left(Y_{ij}^{(r)} + \delta \right) & \text{if } \alpha = 0 \ , \end{cases}$$

for $i = 1, \ldots, n, \quad j = 1, \ldots, k_i$. The weighted regression calculations described in Section 1.7 are performed, with data $X_{ij}^{(r)}$ and weights w_{ij} from (1.2), to obtain estimates $\hat{\mu}$ and $\hat{\beta}_2, \hat{\beta}_3, \ldots, \hat{\beta}_k$ of the period one mean and the period effects for the subsequent periods. Here, k denotes the maximum number of time periods represented in the sample.

For $i = 1, \ldots, n, \quad j = 1, \ldots, k_i$, define

$$\delta_{ij} = \begin{cases} 1 & \text{if the } i\text{th individual had an observation in the } j \text{ th period,} \\ 0 & \text{otherwise .} \end{cases}$$

If the original data (from Section 1.1) are equally weighted, then let

$$n_j \ = \ \sum_{i=1}^{n} \delta_{ij} \ ,$$

$$\begin{aligned} \bar{X}_j &= n_j^{-1} \sum_{i=1}^{n} \delta_{ij} X_{ij}^{(r)} \ , \\ s_j^2 &= (n_j - 1)^{-1} \sum_{i=1}^{n} \delta_{ij} \left(X_{ij}^{(r)} - \bar{X}_j \right)^2 \ , \end{aligned}$$

for $j = 1, 2, \ldots, k$. Otherwise,

$$\begin{aligned} W_{\cdot j} &= \sum_{i=1}^{n} \delta_{ij} W_{ij} \ , \\ \bar{X}_j &= W_{\cdot j}^{-1} \sum_{i=1}^{n} \delta_{ij} W_{ij} X_{ij}^{(r)} \ , \\ s_j^2 &= W_{\cdot j}^{-1} \sum_{i-1}^{n} \delta_{ij} W_{ij} \left(X_{ij}^{(r)} - \bar{X}_j \right)^2 \ , \end{aligned}$$

for $j = 1, 2, \ldots, k$, where W_{ij} are the weights defined in (1.1). Let

$$\begin{aligned} \hat{\mu}_1 &= \hat{\mu} \ , \\ s_1 &= \sqrt{s_1^2} \ , \end{aligned}$$

and for $j = 2, \ldots, k$, let

$$\begin{aligned} \hat{\mu}_j &= \hat{\mu} + \hat{\beta}_j \ , \\ s_j &= \sqrt{s_j^2} \ , \\ c_j &= s_j^{-1} s_1 \ , \\ a_j &= \hat{\mu}_j - c_j^{-1} \hat{\mu}_1 \ , \\ b_j &= \hat{\mu}_1 - c_j \hat{\mu}_j \ . \end{aligned}$$

The new, adjusted data $X_{ij}^{(h)}$ are

$$X_{ij}^{(h)} = \begin{cases} X_{ij}^{(*)} - b_j \left[1 - (2\,|a_j|)^{-1} X_{ij}^{(r)} \right] I\left(X_{ij}^{(*)} \le 2\,|a_j| \right) & \text{if } \alpha \neq 0 \text{ and } j \neq 1 \ , \\ X_{ij}^{(*)} & \text{if } \alpha = 0 \text{ or } j = 1 \ , \end{cases} \tag{1.25}$$

where

$$X_{ij}^{(*)} = \begin{cases} X_{ij}^{(r)} & \text{if } j = 1 \ , \\ c_j \left(X_{ij}^{(r)} - \hat{\mu}_j \right) + \hat{\mu}_1, & \text{if } j \neq 1 \ , \end{cases} \tag{1.26}$$

for $i = 1, \ldots, n, \ j = 1, \ldots, k_i$. The constants a_j and b_j are the points of intersection of the line defined in the second component of (1.26) and the $X^{(r)}$ and $X^{(*)}$ axes, respectively. The indicator function in the first component of (1.25) is a modification to the linear transformation in (1.26) to ensure that adjusted transformed intakes are positive and that zero intakes are transformed into zero intakes. Empirical results indicate that very few, if any, observations fall into the $[0, 2\,|a_j|]$ interval.

Finally, the adjusted data $Y_{ij}^{(h)}$ in the original scale are

$$Y_{ij}^{(h)} = \begin{cases} \max\left(0, \left(10^{\beta} X_{ij}^{(h)}\right)^{\frac{1}{\alpha}} - \delta\right) & \text{if } \alpha \neq 0 \ , \\ \max\left(0, \exp\left(10^{\beta} X_{ij}^{(h)}\right) - \delta\right) & \text{if } \alpha = 0 \ , \end{cases} \tag{1.27}$$

for $i = 1, \ldots, n, \ j = 1, \ldots, k_i$.

If the data in Section 1.1 were assumed to be equally weighted, then the semiparametric normality transformations and subsequent analysis are performed on the $Y_{ij}^{(h)}$ observations. Otherwise, an equal weight sample $Y_{ij}^{(e)}$ is constructed using the procedure described in Section 1.4, intake data $Y_{ij}^{(h)}$, and sampling weights w_{ij} from (1.2).

If adjustment to the grand mean, rather than to the day one mean, is desired, let

$$\begin{aligned} \hat{\mu}_G &= \begin{cases} n^{-1} \displaystyle\sum_{i=1}^{n} k_i^{-1} \sum_{j=1}^{k_i} X_{ij}^{(r)} & \text{if the data are equally weighted ,} \\ \left(\displaystyle\sum_{i=1}^{n} k_i^{-1} \sum_{j=1}^{k_i} W_{ij} \right)^{-1} \displaystyle\sum_{i=1}^{n} k_i^{-1} \sum_{j=1}^{k_i} W_{ij} X_{ij}^{(r)} & \text{otherwise ,} \end{cases} \\ s_p^2 &= \left(\sum_{j=1}^{k} n_j - k \right)^{-1} \sum_{j=1}^{k} (n_j - 1)\, s_j^2 \ , \\ s_p &= \sqrt{s_p^2} \, . \end{aligned}$$

Take $\hat{\beta}_1$ to be zero, and for $j = 1, \ldots, k$, let

$$\begin{aligned} \hat{\mu}_j &= \hat{\mu} + \hat{\beta}_j, \\ c_j &= s_j^{-1} s_p, \\ a_j &= \hat{\mu}_j - c_j^{-1} \hat{\mu}_G, \\ b_j &= \hat{\mu}_G - c_j \hat{\mu}_j. \end{aligned}$$

The grand-mean adjusted data are given by

$$X_{ij}^{(h)} = \begin{cases} X_{ij}^{(*)} - b_j \left[1 - (2\,|a_j|)^{-1} X_{ij}^{(r)}\right] I\left(X_{ij}^{(*)} \leq 2\,|a_j|\right) & \text{if } \alpha \neq 0\ , \\ X_{ij}^{(*)} & \text{if } \alpha = 0\ , \end{cases} \tag{1.28}$$

where

$$X_{ij}^{(*)} = c_j \left(X_{ij}^{(r)} - \hat{\mu}_j\right) + \hat{\mu}_G \tag{1.29}$$

for $i = 1, \ldots, n,\ j = 1, \ldots, k_i$. Equation (1.27) is used to obtain the adjusted data $Y_{ij}^{(h)}$ from the $X_{ij}^{(h)}$ defined in (1.28).

Chapter 2

Semiparametric Normality Transformations

2.1 Anderson-Darling Test for Normality

Let $\{X_i\}_{i=1}^N$ be an equal weight sample of observations from some distribution F. It is desired to test $H_0 : F$ is the cumulative distribution function of a normal distribution vs. H_A : not H_0. Let

$$\begin{aligned} \bar{X} &= N^{-1} \sum_{i=1}^{N} X_i \, , \\ s^2 &= (N-1)^{-1} \sum_{i=1}^{N} \left(X_i - \bar{X} \right)^2 , \\ Z_i &= \Phi \left(\frac{X_i - \bar{X}}{s} \right) \, , \end{aligned}$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. Let $Z_{(i)}$ denote the i^{th} order statistic of the standardized observations, and let

$$\lambda_i = \max \left(1 \times 10^{-7}, Z_{(i)} \left(1 - Z_{(N+1-i)} \right) \right) \, . \tag{2.1}$$

The Anderson-Darling test statistic is

$$A = - \left(1 + \frac{4}{N} - \frac{25}{N^2} \right) \sum_{i=1}^{N} 1 + N^{-1} \left(2i - 1 \right) \ln(\lambda_i) \, . \tag{2.2}$$

See Stephens (1974). Note that the construction of the λ_i avoids taking the logarithm of very small numbers. If A exceeds a given critical value in Table 2.1, the null hypothesis is rejected at the nominal level given in the table. A critical value marked with an asterisk should be considered approximate.

Critical Value	Nominal Type I Error Rate
0*	0.9999
0.127*	0.99
0.193*	0.90
0.250*	0.75
0.340*	0.50
0.465*	0.25
0.576	0.15
0.656	0.10
0.787	0.05
0.918	0.025
1.092	0.01

Table 2.1: Critical values for the Anderson-Darling test.

2.2 Join Point Determination

The semiparametric normality transformation performed by the C-SIDE software uses a piecewise polynomial to approximate the nonlinear function that relates observed quantiles of an intake distribution to the corresponding quantiles of the standard normal distribution. An important step in the transformation procedure is to determine *join points*, places where the polynomial will change shape. The number of join points is variable, and is denoted by p. The p join points partition the domain of the piecewise polynomial into $p+1$ regions. The constructed polynomial is to be linear on the first and last regions and cubic over the $p-1$ interior regions.

Let the data be $\left\{\left(T_{(i)},\ Z_{(i)}\right)\right\}_{i=1}^{N}$, where the $T_{(i)}$ are the ordered observations from an intake distribution and the $Z_{(i)}$ are the Blom normal scores (1.7) for a sample of size N. Let $m = \lfloor \tau N \rfloor$ and $M = N - m + 1$, where τ is the pre-specified proportion of observations to be placed in each of the end regions. Typically, τ is chosen to make $m = 2$. If necessary, the values of m and M are adjusted so that both sets of observations $\left\{T_{(1)},\ \ldots,\ T_{(m)}\right\}$ and $\left\{T_{(M)},\ \ldots,\ T_{(N)}\right\}$ contain at least two distinct values. Join points $\{A_j\}_{j=1}^{p}$ are defined by

$$A_1 \quad = \quad \frac{1}{2}\left(Z_{(m)} + Z_{(m+1)}\right) \ ,$$

$$\begin{aligned} A_p &= \frac{1}{2}\left(Z_{(M-1)} + Z_{(M)}\right) , \\ A_j &= A_1 + (p-1)^{-1}(j-1)(A_p - A_1) , \end{aligned} \tag{2.3}$$

for $j = 2, \ldots, p-1$. Join points chosen in this fashion are equally-spaced in the normal scale.

2.3 Grafted Polynomial Fitting

Given ordered equal weight data $\{Y_{(i)}\}_{i=1}^{N}$, consider the problem of estimating a smooth function H such that $\{H(Y_{(i)})\}_{i=1}^{N}$ is very nearly normally distributed. The function should be such that $H(Y_{(i)}) \doteq Z_{(i)}$ for all $i = 1, 2, \ldots, N$, where $Z_{(i)}$ is the i^{th} Blom normal score (as defined in Section 1.3) for a sample of size N. A plot of the $(Z_{(i)}, Y_{(i)})$ pairs is commonly known as a *normal probability plot*. C-SIDE constructs a smooth estimate of the normal probability plot, and uses the smooth estimate to define a transformation that carries the intakes into normality. Because distributions of observed intakes are often decidedly nonnormal, power or log transformations are applied to yield more nearly linear normal probability plots that are easier to smooth. Let $\{T_{(i)}\}_{i=1}^{N}$ be the power-transformed values defined by the power selection algorithm of Section 1.5. The $(Z_{(i)}, T_{(i)})$ pairs are used to construct a transformation g from Z to T. The function g is restricted so that

1. g is a piecewise function defined over $p+1$ regions,

2. g is a linear function over the first and last regions, but is a cubic polynomial over the $p-1$ interior regions, and

3. g has two continuous derivatives and is monotone .

Because of conditions 1-3, the inverse of g exists and also satisfies 1-3. The inverse of g is used in conjunction with the inverse of the power transformation to define H. C-SIDE estimates g by a regression with functions of normal scores as independent variables and power-transformed intakes as the dependent variable.

1. Let $A_1, A_2, \ldots, A_p$ be join points in the normal scale as determined by the procedure in Section 2.2. Define for $i = 1, 2, \ldots, N$

$$
\begin{aligned}
x_{i1} &= 1 \,, && (2.4)\\
x_{i2} &= Z_{(i)} \,, && (2.5)\\
G_{ij} &= \begin{cases} 0 & \text{if } Z_{(i)} \le A_{j-2}, \\ (Z_i - A_{j-2})^3 & \text{if } Z_{(i)} > A_{j-2}, \end{cases} \quad \text{for } j = 3, 4, \ldots, p+2, && (2.6)\\
x_{ij} &= G_{ij} + C_{2j} G_{i,\,p+1} + C_{3j} G_{i,\,p+2} \qquad \text{for } j = 3, 4, \ldots, p \,, && (2.7)
\end{aligned}
$$

where

$$
\begin{aligned}
C_{2j} &= \frac{A_p - A_{j-2}}{A_{p-1} - A_p} \,, && (2.8)\\
C_{3j} &= \frac{-(A_{p-1} - A_{j-2})}{A_{p-1} - A_p} \,. && (2.9)
\end{aligned}
$$

The $x_{ij},\ j = 1, 2, \ldots, p$ are the independent variables in the regression. Write

$$\mathbf{X} = [\mathbf{x}_1\ \mathbf{x}_2\ \ldots\ \mathbf{x}_p] \,, \tag{2.10}$$

where $\mathbf{x}_j$ denotes a column vector of the $\{x_{ij}\}_{i=1}^N$, and compute

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-} \mathbf{X}'\mathbf{T} \,, \tag{2.11}$$

where $\mathbf{T}$ is a column vector of the $\{T_{(i)}\}_{i=1}^N$. The values of $\{A_j\}_{j=1}^p$ and $\hat{\boldsymbol{\beta}}$ define a grafted polynomial $g(\mathbf{Z}) = \mathbf{X}\hat{\boldsymbol{\beta}}$, where the elements of $\mathbf{X}$ depend on the values of $\mathbf{Z}$ and $\{A_j\}_{j=1}^p$.

2.4 Evaluation and Differentiation of Grafted Polynomials

The calculations in Section 2.3 give rise to a piecewise polynomial $g(Z)$ of the form

$$g(Z) = \begin{cases} \hat{\beta}_1 + \hat{\beta}_2 Z & Z < A_1 \ , \\ \hat{\beta}_1 + \hat{\beta}_2 Z + \hat{\beta}_3 (Z - A_1)^3 & A_1 < Z \leq A_2 \ , \\ \hat{\beta}_1 + \hat{\beta}_2 Z + \hat{\beta}_3 (Z - A_1)^3 + \hat{\beta}_4 (Z - A_2)^3 & A_2 < Z \leq A_3 \ , \\ \hat{\beta}_1 + \hat{\beta}_2 Z + \sum_{j=3}^{k} \hat{\beta}_j (Z - A_{j-2})^3 & A_{k-2} < Z \leq A_{k-1} \ , \\ \hat{\beta}_1 + \hat{\beta}_2 Z + \sum_{j=3}^{p} \hat{\beta}_j \left[(Z - A_{j-2})^3 + C_{2j} (Z - A_{p-1})^3 \right] & A_{p-1} < Z \leq A_p \ , \\ \hat{\beta}_1 + \hat{\beta}_2 Z + \sum_{j=3}^{p} \hat{\beta}_j \left[(Z - A_{j-2})^3 + C_{2j} (Z - A_{p-1})^3 \right. & \\ \qquad \left. + C_{3j} (Z - A_p)^3 \right] & A_p < Z \ , \end{cases} \tag{2.12}$$

for $k = 5, 6, \ldots, p$, where C_{2j} and C_{3j} are defined in (2.8) and (2.9). Evaluation and differentiation of $g(Z)$ with respect to Z is simple, due to the polynomial nature of g.

Operationally, it is often necessary to evaluate g and its derivative for a vector of normal scores $\mathbf{Z} = \{Z_i\}_{i=1}^n$. When evaluation of g is required, the expression

$$g(\mathbf{Z}) = \mathbf{X}\hat{\boldsymbol{\beta}}$$

is used, where the model matrix $\mathbf{X}$ is described in (2.4)-(2.10).

When the derivative g is required, the quantities (2.4)-(2.6) are differentiated to get

$$\begin{aligned} & u_{i1} = 0 \ , \\ & u_{i2} = 1 \ , \\ & H_{ij} = \begin{cases} 0 & \text{if } Z_i \leq A_{j-2} \ , \\ 3(Z_i - A_{j-2})^2 & \text{if } Z_i > A_{j-2} \ , \end{cases} \qquad \text{for } j = 3, 4, \ldots, p+2 \ , \\ & u_{ij} = H_{ij} + C_{2j} H_{i,\,p+1} + C_{3j} H_{i,\,p+2} \ , \qquad \text{for } j = 3, 4, \ldots, p \ , \end{aligned}$$

where C_{2j} and C_{3j} are defined in (2.8) and (2.9), respectively. Let

$$\mathbf{U} = [\mathbf{u}_1 \ \ \mathbf{u}_2 \ \ \ldots \ \ \mathbf{u}_p] \, ,$$

where $\mathbf{u}_j$ denotes a column vector of the $\{u_{ij}\}_{i=1}^n$. The derivative $g'(\mathbf{Z})$ of g at the points $\mathbf{Z} = \{Z_i\}_{i=1}^n$ is the vector $g'(\mathbf{Z}) = \mathbf{U}\hat{\boldsymbol{\beta}}$.

Note that g and g' are functions from Z to T , where Z denotes the normal scale and T denotes the shifted, power transformed, scaled intake scale. To get a function from Z to Y , where Y denotes the original intake scale, results on the composition of functions are used to obtain

$$Y = f(T) = f\left(g\left(Z\right)\right) = h\left(Z\right) \ ,$$

where

$$f\left(t\right) = \begin{cases} \left(10^{\beta}t\right)^{\frac{1}{\alpha}} - \delta & \text{if } \alpha \neq 0 \ , \\ \exp(10^{\beta}t) - \delta & \text{if } \alpha = 0 \ , \end{cases} \tag{2.13}$$

and

$$\begin{aligned} Y' &= f'(T) = f'\left(g\left(Z\right)\right) g'\left(Z\right) = h'\left(Z\right) \\ &= \begin{cases} \alpha^{-1}10^{\beta}\left[10^{\beta}g\left(Z\right)\right]^{\frac{1}{\alpha}-1} g'\left(Z\right) & \text{if } \alpha \neq 0 \ , \\ 10^{\beta}\exp\left(10^{\beta}g\left(Z\right)\right) g'\left(Z\right) & \text{if } \alpha = 0 \ . \end{cases} \end{aligned}$$

Note that it is possible to obtain negative values for $g(Z)$ for very small values of Z , and that the subtraction of δ in (2.13) can also produce negative estimates for intakes, which are by definition nonnegative. Negative values of $g(Z)$ are replaced with zeros when $\alpha \neq 0$, so that the exponentiation of $g(Z)$ can be performed by computer using the identity $a^x = \exp\left(x \ \ln a\right)$. Likewise, negative estimates of intake are replaced with zeros.

2.5 Inversion of the Grafted Polynomial

Sections 2.3-2.4 detail the construction of a function $h : Z \to Y$ as $f \circ g$, where $f : T \to Y$ and $g : Z \to T$. The function g is a grafted cubic polynomial, and the function f is a combination of a shift, a power or log transformation, and a scale factor. The transformation $H : Y \to Z$ mentioned in Section 2.3 is given by $H = g^{-1}\left(f^{-1}\left(Y\right)\right)$.

Let $f(t)$ be determined by α, β, and δ as in (2.13). Then

$$f^{-1}(y) = \begin{cases} 10^{-\beta}\left(y + \delta\right)^{\alpha} & \text{if } \alpha \neq 0 \ , \\ 10^{-\beta}\ln\left(y + \delta\right) & \text{if } \alpha = 0 \ . \end{cases}$$

There is no simple explicit form for g^{-1}, but g^{-1} can be computed at any point t_0, because g is a cubic polynomial. Let $g^{-1}(t_0) = g^{-1}(f^{-1}(Y_0))$. Expanding the terms of g (see (2.12)) yields

$$g(Z) = C_1 Z^3 + C_2 Z^2 + C_3 Z + C_4 ,$$

where C_1, C_2, C_3, and C_4 depend on $\hat{\boldsymbol{\beta}}$ from (2.11) and $\{A_j\}_{j=1}^p$ from (2.3). Computation of the values of C_1, C_1, C_3, and C_4 is a tedious exercise in algebra involving repeated use of the expansion

$$(Z - A_j)^3 = Z^3 - 3Z^2 A_j + 3Z A_j^2 - A_j^3 ,$$

and is omitted.

Now, $g^{-1}(t_0)$ is the value Z_0 such that $g(Z_0) = C_1 Z_0^3 + C_2 Z_0^2 + C_3 Z_0 + C_4 = t_0$, or equivalently such that

$$\tilde{g}(Z_0) = C_1 Z_0^3 + C_2 Z_0^2 + C_3 Z_0 + (C_4 - t_0) = 0 . \tag{2.14}$$

If $C_1 = C_2 = 0$, i.e. g is linear in Z, then either $t_0 > g(A_p)$ or $t_0 < g(A_1)$, and Z_0 is

$$Z_0 = \frac{t_0 - C_4}{C_3} .$$

Suppose g is not linear in Z. Then by construction, g is cubic in Z. Solutions to (2.14) are solutions to

$$\begin{aligned} \bar{g}(Z_0) &= Z_0^3 + \frac{C_2}{C_1} Z_0^2 + \frac{C_3}{C_1} Z_0 + \frac{(C_4 - t_0)}{C_1} \\ &= Z_0^3 + p Z_0^2 + q Z_0 + r \\ &= 0 . \end{aligned}$$

The cubic equation

$$Z^3 + pZ^2 + qZ + r = 0$$

is reduced to the normal form

$$x^3 + ax + b = 0 , \tag{2.15}$$

where

$$a = \frac{1}{3}\left(3q - p^2\right) \ , \quad b = \frac{1}{27}\left(2p^3 - 9pq + 27r\right) \ ,$$

by the substitution

$$Z = \left(x - \frac{p}{3}\right) \ . \tag{2.16}$$

The solutions x_1, x_2, and x_3 to (2.15) are

$$\begin{aligned} x_1 &= A + B \ , \\ x_2, x_3 &= -\frac{1}{2}\left(A + B\right) \pm \frac{i\sqrt{3}}{2}\left(A - B\right) \ , \end{aligned}$$

where

$$\begin{aligned} A &= \sqrt[3]{-\frac{b}{2} + \sqrt{\frac{b^2}{4} + \frac{a^3}{27}}}, \\ B &= \sqrt[3]{-\frac{b}{2} - \sqrt{\frac{b^2}{4} + \frac{a^3}{27}}}. \end{aligned}$$

Case I: If $\frac{b^2}{4} + \frac{a^3}{27} > 0$, there is one real root, x_1.

Case II: If $\frac{b^2}{4} + \frac{a^3}{27} = 0$, there are three real roots, of which two are equal.

In this case, the roots are given by

$$(x_1,\ x_2,\ x_3) = \left(\mp 2\sqrt{-\frac{a}{3}},\ \pm\sqrt{-\frac{a}{3}},\ \pm\sqrt{-\frac{a}{3}}\right) \ ,$$

where the upper sign is used for b positive, and the lower sign is used for b negative.

Case III: If $\frac{b^2}{4} + \frac{a^3}{27} < 0$, there are three unequal real roots. The roots are

$$x_k = 2\sqrt{-\frac{a}{3}}\cos\left(\frac{\phi + 2\pi(k-1)}{3}\right), \quad k = 1, 2, 3,$$

where

$$\cos\phi = \mp\sqrt{-\frac{27b^2}{4a^3}},$$

and the upper sign is used for b positive and the lower sign is used for b negative. Recalling the substitution in (2.16), there are up to three real numbers, say Z_1, Z_2, Z_3, such that $g(Z_i) = t_0$, given by

$$Z_i = \left(x_i - \frac{p}{3}\right) ,$$

where the $\{x_i\}$ are the roots computed above.

Recall that g is one-to-one. On each interval $[A_{j-1}, A_j]$, g behaves like the cubic polynomial

$$a_j Z^3 + b_j Z^2 + c_j Z + d_j,$$

for which there may be multiple Z values in $(-\infty, \infty)$ that give a value of t_0. However, there is only one Z value in $[A_{j-1}, A_j]$ for which g has a function value of t_0. In the absence of round-off error, it is simple to inspect the potential roots Z_1, Z_2, and Z_3 to determine which root is the desired one. However, tiny numerical inaccuracies in the root-finding computation require a more robust approach. In practice, g is evaluated at each potential root Z_i, and the root for which $|g(Z_i) - t_0|$ is minimum is taken as the unique $Z_0 = g^{-1}(t_0)$. Because the "wrong" roots can be very far from the "right" root, the evaluation of g for the potential roots requires additional bound checking to prevent overflow errors.

2.6 Construction of Normality Transformations

Consider the adjusted, equal weight data $Y_{ij}^{(e)}$ obtained after performing the procedures of Chapter 1. Let N be the total number of observations and denote the ordered data by $\{Y_{(k)}\}_{k=1}^N$. Let $\{Z_{(k)}\}_{k=1}^N$ be the associated Blom normal scores defined by (1.7). The procedure described in Section 1.4 is performed to obtain a shift δ, power α, and scale factor β so that the set of observations $\{T_{(k)}\}_{k=1}^N$ is roughly normally distributed, where

$$T_{(k)} = \begin{cases} 10^{-\beta} \left(Y_{(k)} + \delta\right)^{\alpha} & \text{if } \alpha \neq 0 \ , \\ 10^{-\beta} \ln\left(Y_{(k)} + \delta\right) & \text{if } \alpha \neq 0 \ . \end{cases}$$

Let p be the maximum number of join points permitted (typically between 10 and 12), and let $p_0 = 3$. The procedure of Section 2.2 is applied to the $\{(T_{(k)},\ Z_{(k)})\}$ to obtain join points $\{A_j\}_{j=1}^{p_0}$. The $(T_{(k)}, Z_{(k)})$ data and the $\{A_j\}$ are used to fit a grafted polynomial $g_{p_0}(Z)$ as described in Section 2.3. The grafted polynomial depends on the $\{A_j\}$ and the vector of p_0 parameter estimates $\hat{\boldsymbol{\beta}}$ in (2.11). Recall that the derivative of g is to be nonnegative over the range of the $\{Z_{(k)}\}$. As a partial check for this condition, the fitted polynomials are required to have a positive derivative at each join point.

If g_{p_0} has a negative derivative at any of the $\{A_j\}$, the value of p_0 is increased by 1, and the grafted polynomial fitting is repeated using a new set of $\{A_j\}$. Otherwise, g_{p_0} is checked to see if it yields an acceptable normality transformation. Let $\hat{Z}_{(k)} = g_{p_0}^{-1}(T_{(k)})$ where $g_{p_0}^{-1}(\cdot)$ is computed using results from Section 2.5.

The Anderson-Darling test for normality described in Section 2.1 is applied to the $\{\hat{Z}_{(k)}\}$. If the test is significant at some user-specified level (typically 0.15), the value of p_0 is increased by 1, and the grafted polynomial fitting is repeated using a new set of $\{A_j\}$. This process continues until g_{p_0} yields an acceptable normality transformation with a nonnegative derivative at each join point, or until $p_0 = p$. If $p_0 = p$ and g_p does not satisfy the nonnegativity/normality criteria, C-SIDE reports an error and stops.

The polynomial $g_{p_0}(\cdot)$ with the fewest join points, with a nonnegative derivative at each join point, and with an Anderson-Darling test statistic less than the specified value, is chosen as the transformation from Z to T. The inverse $g_{p_0}^{-1}(\cdot)$ is the transformation from T to Z.

Thus, the transformation H that transforms the equal weight observations $\{Y_{ij}\}$ into normally-distributed observations $\{\hat{Z}_{ij}\}$ is

$$\hat{Z}_{ij} \equiv H\left(Y_{ij}\right) = \begin{cases} g_{p_0}^{-1}\left(10^{-\beta}\left(Y_{ij}+\delta\right)^{\alpha}\right) & \text{if } \alpha \neq 0\ , \\ g_{p_0}^{-1}\left(10^{-\beta}\ln\left(Y_{ij}+\delta\right)\right) & \text{if } \alpha = 0\ , \end{cases} \tag{2.17}$$

for $i = 1, \ldots, n,\ j = 1, \ldots, k_i$.

Chapter 3

Usual Intake Distributions

3.1 Variance Components for Independent Within-Individual Observations

Assume that the transformation H derived in Section 2.6 produces transformed observations $\{X_{ij}\}$ that satisfy the model

$$\begin{aligned} X_{ij} &= x_i + u_{ij} \ , \\ u_{ij} &= \sigma_i e_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, k_i \ , \end{aligned} \tag{3.1}$$

where

$$x_i \sim NI\left(\mu_x, \sigma_x^2\right), \ e_{ij} \sim NI\left(0, 1\right), \ \sigma_i^2 \sim \left(\mu_A, \sigma_A^2\right) \ . \tag{3.2}$$

Assume that x_i, σ_i^2, and u_{kj} are independent for all i, k, j. It follows that

$$\mathrm{E}\{u_{ij}^2\} = \mathrm{E}\{\sigma_i^2 e_{ij}^2\} = \mu_A \ .$$

Let

$$\overline{X}_{i\cdot} = k_i^{-1} \sum_{j=1}^{k_i} X_{ij}, \quad \hat{\mu}_x = n^{-1} \sum_{i=1}^{n} \overline{X}_{i\cdot}, \quad N = \sum_{i=1}^{n} k_i, \ n_0 = N - N^{-1} \sum_{i=1}^{n} k_i^2 \ ,$$

and consider the ANOVA of Table 3.1, in which $k = \max\left(k_i\right)$ and the subtraction of $(k-1)$ from the residual and total degrees of freedom corrects for the removal of period effects described in Section 1.8. Equating mean squares to their expectations yields

$$\hat{\mu}_A = \left[N - n - (k-1)\right]^{-1} \sum_{i=1}^{n} \sum_{j=1}^{k_i} \left(X_{ij} - \overline{X}_{i\cdot}\right)^2 , \tag{3.3}$$

$$\hat{\sigma}_x^2 = n_0^{-1} \left[\sum_{i=1}^{n} k_i \left(\overline{X}_{i\cdot} - \hat{\mu}_x\right)^2 - (n-1)\, \hat{\mu}_A \right] . \tag{3.4}$$

Source	**df**	**SS**	**E{MS}**
Individual	$n-1$	$\sum_{i=1}^{n} k_i \left(\overline{X}_{i\cdot} - \hat{\mu}_x\right)^2$	$(n-1)^{-1} n_0 \sigma_x^2 + \mu_A$
Residual	$N-n-(k-1)$	$\sum_{i=1}^{n}\sum_{j=1}^{k_i} \left(X_{ij} - \overline{X}_{i\cdot}\right)^2$	μ_A
Total	$N-1-(k-1)$		

Table 3.1: ANOVA for one-way classification with unbalanced data.

Assume that $k_i > 1$ for m of the n individuals and let

$$A_i = (k_i - 1)^{-1} \sum_{j=1}^{m} \left(X_{ij} - \overline{X}_{i\cdot}\right)^2 , \tag{3.5}$$

$$s_i = \sqrt{A_i} , \tag{3.6}$$

$$d_i = (k_i - 1) , \tag{3.7}$$

where $i = 1, \ldots, m$ is used to index the individuals with multiple observations. To investigate the possibility that the A_i are correlated with the $\overline{X}_{i\cdot}$, two sets of regression calculations are performed.

1. Let $\mathbf{1}$ denote a column vector of ones, $\mathbf{S}$ a column vector of the s_i, $\mathbf{W}$ a diagonal matrix of the d_i, and compute

$$\begin{aligned} \hat{\beta}_1 &= \left(\mathbf{1'W1}\right)^{-1} \mathbf{1'WS} , \\ SS_R &= \hat{\beta}_1^2 \left(\mathbf{1'W1}\right) . \end{aligned}$$

2. Let $\bar{\mathbf{X}}$ denote a column vector of the $\overline{X}_{i\cdot}$, let $\mathbf{T} = [\mathbf{1}\ \bar{\mathbf{X}}]$, and define

$$\begin{aligned} \hat{\boldsymbol{\beta}}_2 &= \left(\mathbf{T'WT}\right)^{-} \mathbf{T'WS} , \\ SS_F &= \left(\mathbf{T}\hat{\boldsymbol{\beta}}_2\right)' \mathbf{W} \left(\mathbf{T}\hat{\boldsymbol{\beta}}_2\right) , \\ SS_E &= \left(\mathbf{S} - \mathbf{T}\hat{\boldsymbol{\beta}}_2\right)' \mathbf{W} \left(\mathbf{S} - \mathbf{T}\hat{\boldsymbol{\beta}}_2\right) , \\ MS_E &= \frac{SS_E}{(m-2)} , \end{aligned}$$

$$\begin{aligned} \widehat{\mathrm{Cov}}\left(\hat{\boldsymbol{\beta}}_2\right) &= MS_E\left(\mathbf{T}'\mathbf{W}\mathbf{T}\right)^{-}, \\ F &= \frac{(SS_F - SS_R)}{MS_E}. \end{aligned}$$

Then a test of H_0 : Individual standard deviations are uncorrelated with individual means vs. H_A : Individual standard deviations are linearly related to individual means has p-value $1 - \Pr\left(F^1_{m-2} \leq F\right)$, where F^1_{m-2} denotes a random variable distributed as Snedecor's F with 1 numerator and $(m-2)$ denominator degrees of freedom.

Recall from (3.2) that $\sigma_i^2 \sim (\mu_A, \sigma_A^2)$. An estimate of μ_A is obtained from (3.3). The fourth moment of u_{ij} is

$$\mathrm{E}\left\{u_{ij}^4\right\} = \mathrm{E}\left\{3\sigma_i^4\right\} .$$

Now,

$$\begin{aligned} \mathrm{E}\left\{A_i^2|i\right\} &= \sigma_i^4 + \mathrm{Var}\left\{A_i^2\right\} \\ &= \sigma_i^4\left[1 + 2\left(k_i - 1\right)^{-1}\right] . \end{aligned}$$

It follows that an unbiased estimator of $\mathrm{E}\left\{u_{ij}^4\right\}$ is

$$3m^{-1}\sum_{i=1}^{m}\left[1 + 2\left(k_i - 1\right)^{-1}\right]^{-1} A_i^2 ,$$

and an approximately unbiased estimator of σ_A^2 is

$$\hat{\sigma}_A^2 = m^{-1}\sum_{i=1}^{m}\left[1 + 2\left(k_i - 1\right)^{-1}\right]^{-1} A_i^2 - \hat{\mu}_A^2 . \tag{3.8}$$

Let $\hat{M}_{A4}$ be the standardized estimator

$$\hat{M}_{A4} = 3\hat{\mu}_A^{-2} m^{-1}\sum_{i=1}^{m}\left[1 + 2\left(k_i - 1\right)^{-1}\right]^{-1} A_i^2 . \tag{3.9}$$

If $\sigma_A^2 = 0$, then $\hat{M}_{A4}$ estimates 3, the fourth moment of the standard normal distribution, and

$$\begin{aligned}
\hat{\mu}_A^{-2}\hat{\sigma}_A^2 &= m^{-1}\sum_{i=1}^{m}\left[1+2\left(k_i-1\right)^{-1}\right]^{-1}\left(\frac{A_i}{\hat{\mu}_A}\right)^2 - 1 \\
&= m^{-1}\sum_{i=1}^{m}\left[1+2\left(k_i-1\right)^{-1}\right]^{-1}\left(k_i-1\right)^{-2}\left[\frac{\left(k_i-1\right)A_i}{\hat{\mu}_A}\right]^2 - 1 \\
&= \frac{1}{3}\left(\hat{M}_{A4}-3\right) .
\end{aligned}$$

Now, assuming $\hat{\mu}_A = \mu_A$,

$$C_i = \frac{\left(k_i-1\right)A_i}{\hat{\mu}_A} \sim \chi^2_{(k_i-1)} .$$

The variance of C_i^2 is given by

$$\mathrm{Var}\left\{C_i^2\right\} = \mu_{4i} - \mu_{2i}^2 , \tag{3.10}$$

where μ_{ki} denotes the k^{th} noncentral moment of a chi-square random variable with degrees of freedom $d_i = (k_i - 1)$.

Now,

$$\mu_{2i}^2 = \left(2d_i + d_i^2\right)^2 = d_i^4 + 4d_i^3 + 4d_i^2$$

and

$$\begin{aligned}
\mu_{4i} &= 2^4\frac{\Gamma\left(4+\frac{d_i}{2}\right)}{\Gamma\left(\frac{d_i}{2}\right)} = 2^4\left(\frac{d_i}{2}+3\right)\left(\frac{d_i}{2}+2\right)\left(\frac{d_i}{2}+1\right)\left(\frac{d_i}{2}\right) \\
&= \left(d_i+6\right)\left(d_i+4\right)\left(d_i+2\right)\left(d_i\right) \\
&= d_i^4 + 12d_i^3 + 44d_i^2 + 48d_i .
\end{aligned}$$

Hence, (3.10) becomes

$$\mathrm{Var}\left\{C_i^2\right\} = 8d_i^3 + 40d_i^2 + 48d_i ,$$

and

$$\begin{aligned}
\mathrm{Var}\left\{\hat{\mu}_A^{-2}\hat{\sigma}_A^2\right\} &= m^{-2}\sum_{i=1}^{m}\left[1+2d_i^{-1}\right]^{-2} d_i^{-4}\left[8d_i^3 + 40d_i^2 + 48d_i\right] \\
&= m^{-2}\sum_{i=1}^{m}\left[1+2d_i^{-1}\right]^{-2}\left[8d_i^{-1} + 40d_i^{-2} + 48d_i^{-3}\right] .
\end{aligned} \tag{3.11}$$

It follows that

$$\mathrm{Var}\left\{\hat{M}_{A4}\right\} = 9\mathrm{Var}\left\{\hat{\mu}_A^{-2}\hat{\sigma}_A^2\right\},$$

and that an approximate size α test for $H_0 : M_{A4} = 3$ vs. $H_A : M_{A4} \neq 3$ is to reject H_0 if

$$\frac{\left|\hat{M}_{A4} - 3\right|}{\sqrt{\mathrm{Var}\left\{\hat{M}_{A4}\right\}}} \geq t_{\alpha/2}^{m-1}, \tag{3.12}$$

where t_α^d denotes the upper α percentile of a Student's t distribution with d degrees of freedom. In most cases, m is very large, and $t_{\alpha/2}^{m-1}$ can be replaced with the corresponding $\alpha/2$ percentile from the normal distribution.

3.2 Variance Components for Correlated Within-Individual Observations

As in Section 3.1, assume that the transformation H derived in Section 2.6 produces transformed observations $\{X_{ij}\}$ that satisfy the model

$$\begin{aligned} X_{ij} &= x_i + u_{ij}\ , \\ u_{ij} &= \sigma_i e_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, k_i\ , \end{aligned} \tag{3.13}$$

where

$$x_i \sim NI\left(\mu_x, \sigma_x^2\right), e_{ij} \sim N\left(0, 1\right),\ \sigma_i^2 \sim \left(\mu_A, \sigma_A^2\right)\ .$$

As in Section 3.1, assume x_i and σ_i^2 are independent for all i. However the assumption that the u_{ij} are independent is relaxed. Assume that the k_i observations for individual i are taken on a daily basis, and that $k = \max\left(k_i\right)$ days are represented in the survey. It is possible that some individuals may not have all k daily observations recorded. Suppose, for a known correlation coefficient ρ, that for $j, m = 1, \ldots, k$,

$$\mathrm{Cov}\left\{u_{ij}, u_{lm}\right\} = \begin{cases} 0 & \text{if } i \neq l\ , \\ \sigma_i^2 \rho^{|j-m|} & \text{if } i = l\ . \end{cases} \tag{3.14}$$

C-SIDE performs calculations for the structure (3.14) only if $k \leq 3$ **and** the value of ρ is supplied. Carriquiry et al. (1995) provide estimates of the between-day correlation coefficient ρ for several common nutrients. If $k = 3$, and $k_i = 2$ for some i, it must be known whether the two observations are taken on consecutive days or if the two observations are separated by an additional day.

Case I: $k = 2$, and the first m individuals have two observations.

Let

$$\bar{X}_{i\cdot} = k_i^{-1} \sum_{j=1}^{k_i} X_{ij},\ \hat{\mu}_x = n^{-1} \sum_{i=1}^{n} \bar{X}_{i\cdot}\ .$$

It follows that

$$\mathrm{E}\left\{\overline{X}_{i\cdot}^2\right\} = \begin{cases} \mu_x^2 + \sigma_x^2 + \mu_A & \text{if } k_i = 1\ , \\ \mu_x^2 + \sigma_x^2 + \frac{1}{2}\left(1+\rho\right)\mu_A & \text{if } k_i = 2\ , \end{cases}$$

and

$$\mathrm{E}\left\{\sum_{j=1}^{k_i}\left(X_{ij} - \overline{X}_{i\cdot}\right)^2\right\} = \begin{cases} 0 & \text{if } k_i = 1\ , \\ \mu_A\left(1-\rho\right) & \text{if } k_i = 2\ . \end{cases}$$

Then an estimator of μ_A is

$$\hat{\mu}_A = \left[m\left(1-\rho\right) - 1\right]^{-1} \sum_{i=1}^{n}\sum_{j=1}^{k_i}\left(X_{ij} - \overline{X}_{i\cdot}\right)^2 . \tag{3.15}$$

The subtraction of 1 in the reciprocal term corrects for the adjustment in Section 1.8.

Furthermore, assuming $\hat{\mu}_x = \mu_x$,

$$\mathrm{E}\left\{k_i\left(\bar{X}_{i\cdot} - \hat{\mu}_x\right)^2\right\} = \begin{cases} \sigma_x^2 + \mu_A & \text{if } k_i = 1\ , \\ 2\sigma_x^2 + \left(1+\rho\right)\mu_A & \text{if } k_i = 2\ , \end{cases}$$

and

$$\mathrm{E}\left\{\sum_{i=1}^{n} k_i\left(\bar{X}_{i\cdot} - \hat{\mu}_x\right)^2\right\} = \left(n+m\right)\sigma_x^2 + \left(n + m\rho\right)\mu_A\ . \tag{3.16}$$

Instead of using (3.16) to estimate σ_x^2, let

$$\hat{\sigma}_x^2 = n_0^{-1}\left[\sum_{i=1}^{n} k_i\left(\bar{X}i. - \hat{\mu}_x\right)^2 - n^{-1}\left(n-1\right)\left(n + m\rho\right)\hat{\mu}_A\right]\ , \tag{3.17}$$

where $N = \sum_{i=1}^{n} k_i$ and $n_0 = N - N^{-1} \sum_{i=1}^{n} k_i^2$.

The factor $n^{-1}(n-1)$ makes (3.17) equal (3.4) when $\rho = 0$. Note that in **Case I**, $N = n + m$, and $\sum_{i=1}^{n} k_i^2 = n + 3m$, so that $n_0 = \left[n + m - 1 - 2\,(n+m)^{-1}\, m\right]$. For large values of n and m, the effect of using n_0 instead of $(n+m)$ is negligible.

For correlated X_{ij} and $k = 2$, the quantities

$$A_i = \frac{(X_{i1} - X_{i2})^2}{2\,(1-\rho)}, \quad i = 1, 2, \ldots, m \tag{3.18}$$

are analogous to the $\{A_i\}_{i=1}^{m}$ of (3.5). The regression of individual standard deviations on individual means and the estimation of $\hat{M}_{A4}$ and $\hat{\sigma}_A^2$ are carried out exactly as in Section 3.1, with (3.18) replacing (3.5).

Case II: $k = 3$.

Let

$$\overline{X}_{i\cdot} = k_i^{-1} \sum_{j=1}^{k_i} X_{ij},\ \hat{\mu}_x = n^{-1} \sum_{i=1}^{n} \overline{X}_{i\cdot}\ .$$

It follows that

$$\mathrm{E}\left\{X_{i\cdot}^2\right\} = \begin{cases} \mu_x^2 + \sigma_x^2 + \mu_A & \text{if } k_i = 1\ , \\ \mu_x^2 + \sigma_x^2 + \frac{1}{2}(1+\rho)\mu_A & \text{if } k_i = 2 \text{ consecutive days}\ , \\ \mu_x^2 + \sigma_x^2 + \frac{1}{2}(1+\rho^2)\mu_A & \text{if } k_i = 2 \text{ nonconsecutive days}\ , \\ \mu_x^2 + \sigma_x^2 + \frac{1}{9}(3+4\rho+2\rho^2)\mu_A & \text{if } k_i = 3\ , \end{cases}$$

and

$$\mathrm{E}\left\{\sum_{j=1}^{k_i} \left(X_{ij} - \overline{X}_{i\cdot}\right)^2\right\} = \begin{cases} 0 & \text{if } k_i = 1\ , \\ \mu_A\,(1-\rho) & \text{if } k_i = 2 \text{ consecutive days}\ , \\ \mu_A\left(1-\rho^2\right) & \text{if } k_i = 2 \text{ nonconsecutive days}\ , \\ \mu_A\left(2 - \frac{4}{3}\rho - \frac{2}{3}\rho^2\right) & \text{if } k_i = 3\ . \end{cases}$$

Let

$$m_1 \;=\; \sum_{i=1}^{n} I\,(k_i = 1)\ ,$$

$$\begin{aligned} m_{2c} &= \sum_{i=1}^{n} I\left(k_i = 2 \text{ consecutive days}\right) , \\ m_{2n} &= \sum_{i=1}^{n} I\left(k_i = 2 \text{ nonconsecutive days}\right) , \\ m_3 &= \sum_{i=1}^{n} I\left(k_i = 3\right) . \end{aligned}$$

Then

$$\begin{aligned} \mathrm{E}\left\{\sum_{i=1}^{n}\sum_{j=1}^{k_i}\left(X_{ij} - \overline{X}_{i\cdot}\right)^2\right\} &= \left[m_{2c}\left(1-\rho\right) + m_{2n}\left(1-\rho^2\right) + m_3\left(2 - \frac{4}{3}\rho - \frac{2}{3}\rho^2\right)\right]\mu_A \\ &= K_1\mu_A , \end{aligned}$$

which yields

$$\hat{\mu}_A = \left(K_1 - 2\right)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{k_i}\left(X_{ij} - \overline{X}_{i\cdot}\right)^2 . \tag{3.19}$$

The subtraction of 2 in the reciprocal term corrects for the adjustment in Section 1.8.

Furthermore, assuming $\hat{\mu}_x = \mu_x$,

$$\mathrm{E}\left\{k_i\left(\overline{X}_{i\cdot} - \hat{\mu}_x\right)^2\right\} = \begin{cases} \sigma_x^2 + \mu_A & \text{if } k_i = 1 , \\ 2\sigma_x^2 + \left(1+\rho\right)\mu_A & \text{if } k_i = 2 \text{ consecutive days}, \\ 2\sigma_x^2 + \left(1+\rho^2\right)\mu_A & \text{if } k_i = 2 \text{ nonconsecutive days}, \\ 3\sigma_x^2 + \left(1 + \frac{4}{3}\rho + \frac{2}{3}\rho^2\right)\mu_A & \text{if } k_i = 3 , \end{cases}$$

and thus

$$\begin{aligned} \mathrm{E}\left\{\sum_{i=1}^{n} k_i\left(\overline{X}_{i\cdot} - \hat{\mu}_x\right)^2\right\} &= \left[m_1 + m_{2c}\left(1+\rho\right) + m_{2n}(1+\rho^2)\right. \\ &\quad \left. + m_3\left(1 + \tfrac{4}{3}\rho + \tfrac{2}{3}\rho^2\right)\right]\mu_A + \\ &\quad \left[m_1 + 2\left(m_{2c} + m_{2n}\right) + 3m_3\right]\sigma_x^2 . \end{aligned} \tag{3.20}$$

As in **Case I**, let $N = \sum_{i=1}^{n} k_i$ and estimate σ_x^2 by

$$\hat{\sigma}_x^2 = \left(N - N^{-1}\sum_{i=1}^{n} k_i^2\right)^{-1}\left(\sum_{i=1}^{n} k_i\left(\overline{X}_{i\cdot} - \hat{\mu}_x\right)^2 - n^{-1}\left(n-1\right)K_2\hat{\mu}_A\right) ,$$

where $K_2 = m_1 + m_{2c}(1+\rho) + m_{2n}(1+\rho^2) + m_3(1+\frac{4}{3}\rho+\frac{2}{3}\rho^2)$. The quantities A_i for the correlated data in **Case II** are

$$A_i = \begin{cases} \dfrac{(X_{i1}-X_{i2})^2}{2(1-\rho)} & \text{if } k_i = 2 \text{ consecutive days,} \\ \dfrac{(X_{i1}-X_{i3})^2}{2(1-\rho^2)} & \text{if } k_i = 2 \text{ nonconsecutive days,} \\ \dfrac{1}{2}\left[\dfrac{(X_{i1}-2X_{i2}+X_{i3})^2}{6-8\rho+2\rho^2} + \dfrac{(X_{i1}-X_{i3})^2}{2-2\rho^2}\right] & \text{if } k_i = 3\ . \end{cases} \tag{3.21}$$

The regression of individual standard deviations on individual means and the estimation of M_{A4} and $\hat{\sigma}_A^2$ are carried out exactly as in Section 3.1, with (3.21) replacing (3.5).

3.3 Usual Intake Transformation

Assume that a transformation $H(y)$ transforms adjusted equal weight intakes $\{Y_{ij}\}$ to normal-scale intakes $\{X_{ij}\}$. Assume the $\{X_{ij}\}$ satisfy a measurement-error model

$$\begin{aligned} X_{ij} &= x_i + u_{ij}\ , \\ u_{ij} &= \sigma_i e_{ij},\ i = 1,\ldots,n,\ j = 1,\ldots,k_i\ , \end{aligned}$$

where

$$\begin{aligned} x_i &\sim NI(\mu_x, \sigma_x^2),\ e_{ij} \sim N(0,1),\ \sigma_i^2 \sim \left(\mu_A, \sigma_A^2\right)\ , \\ \mathrm{Cov}\left\{u_{ij}, u_{lm}\right\} &= \begin{cases} 0 & \text{if } i \neq l\ , \\ \sigma_i^2 \rho^{|j-m|} & \text{if } i = l\ , \end{cases} \end{aligned}$$

and the x_i, σ_i^2 are independent for all i. Note that if ρ is zero, ρ^0 is taken to be one. Sections 3.1 and 3.2 develop estimators for μ_x, σ_x^2, μ_A,and σ_A^2 . The transformation g that takes usual intakes from the normal space into the original space is constructed by approximating y, the conditional expectation of Y, at 400 values of x and then fitting a smooth function to the (x, y) points. For $i = 1,\ldots,400$, let

$$x_i = \sqrt{\hat{\sigma}_x^2}\, Z_{(i)}\ ,$$

where $Z_{(i)}$ is the i^{th} Blom normal score defined in (1.7). A set of points $\{(c_j, w_j)\}_{j=1}^{9}$ is used to approximate the measurement error distribution. Then an estimate of the conditional expectation at transformed usual intake x_i is

$$y_i = \sum_{i=1}^{9} w_j y_{ij} \ , \tag{3.22}$$

where

$$y_{ij} = \max\left(0, H^{-1}(x_i + c_j)\right) \ . \tag{3.23}$$

To allow for heterogeneous within-individual variances, the (c_j, w_j) are constructed so that the first five moments of the nine-point discrete distribution match the first five estimated moments of the measurement error distribution. The set of (c_j, w_j) for the standard normal distribution are shown in Table 3.2. These $(c_j,\ w_j)$ satisfy

$$\sum_{j=1}^{9} w_j = 1 \ , \tag{3.24}$$

$$\sum_{j=1}^{9} w_j c_j^k = 0 \text{ for } k = 1, 3, 5 \ , \tag{3.25}$$

$$\sum_{j=1}^{9} w_j c_j^2 = 1 \ , \tag{3.26}$$

$$\sum_{j=1}^{9} w_j c_j^4 = 3 \ . \tag{3.27}$$

Recall that $\hat{M}_{A4}$ of (3.9) is a standardized estimate of the fourth moment of the measurement error distribution. If $\hat{M}_{A4}$ is greater than 7.5, it is replaced with 7.5. To modify the (c_j, w_j) of Table 3.2 to get the correct fourth moment in (3.27), the weights for $(\pm 1.3, \pm 0.8, \pm 0.5)$ are multiplied by a

c_j	0	± 0.5	± 0.8	± 1.3	± 2.1
w_j	0.252489	0.159698	0.070458	0.080255	0.063345

Table 3.2: Values of $(c_j,\ w_j)$ for the standard normal distribution

constant a and the absolute value of ± 2.1 is replaced by b. The system of equations defining a and b is

$$\begin{aligned} 0.2206474993a + 0.0633452382b &= 0.50 \ , \\ 0.268055472a + 0.0633452382b^2 &= 0.50\hat{M}_{A4} \ , \end{aligned}$$

where 0.2206474993 is the weighted sum of squares for $(1.3, 0.8, 0.5)$, 0.268055472 is the weighted sum of fourth powers for $(1.3, 0.8, 0.5)$, a is the multiplier to be applied to the w_j and b is the value to replace $(2.1)^2$. Solving for b in terms of a in the first equation yields

$$b = 7.893253129 - 3.483253128a \ ,$$

and substituting this expression into the second equation yields a as the (sensible) solution to

$$0.768571092a^2 - 3.215197658a + 3.9466265649 - 0.5\hat{M}_{A4} = 0 \ .$$

Thus

$$a = (1.537142184)^{-1}\left[3.215197658 - \left(1.537142184\hat{M}_{A4} - 1.795556375\right)^{1/2}\right] \ .$$

The new largest value is $b^{1/2}$, and the new set of (c_j, w_j) are given in Table 3.3. The first five moments of the discrete distribution determined by the new (c_j, w_j) are $\left(0, \hat{\mu}_A, 0, \hat{\mu}_A^2\hat{M}_{A4}, 0\right)$. The approximations y_i given by (3.23) and (3.22) for $i = 1, 2, \ldots, 400$ form a "representative sample" of usual intakes in the original scale, just as the x_i, $i = 1, 2, \ldots, 400$ form a representative sample of $N\left(0, \ \hat{\sigma}_x^2\right)$ variates. Recall that the construction of the Blom scores (Section 1.3) matches the first five sample moments of the scores to the first five theoretical moments of the normal distribution. Hence, the first five sample moments of the y_i are estimates of the first five moments of the usual intake distribution in the original scale.

c_j	0	$\pm 0.5\hat{\mu}_A$	$\pm 0.8\hat{\mu}_A$	$\pm 1.3\hat{\mu}_A$	$\pm\sqrt{b}\hat{\mu}_A$
w_j	$0.873310 - 0.620820a$	$0.159698a$	$0.070458a$	$0.080255a$	0.063345

Table 3.3: Values of $(c_j, \ w_j)$ for the measurement error distribution.

To construct a smooth function $g(x)$ that carries $N\left(0,\ \hat{\sigma}_x^2\right)$ intakes back to original scale usual intakes, the methodology developed in Chapter 2 is used. The values of $\delta,\ \alpha$, and β used in the construction of H applied to the $\{y_i\}_{i=1}^{400}$ yield

$$t_i = \begin{cases} 10^{-\beta}\left(y_i+\delta\right)^{\alpha} & \text{if } \alpha \neq 0\ , \\ 10^{-\beta}\ln\left(y_i+\delta\right) & \text{if } \alpha = 0\ . \end{cases}$$

A grafted polynomial $g_1(x)$, with the same number of join points as was used to construct H, is fit to the $(t_i,\ x_i)$ pairs. Because the y_i are weighted averages of points on a smooth function, the grafted polynomial fit to the $(t_i,\ x_i)$ is almost always very good. On the rare cases that $g_1(x)$ has negative derivatives, the number of join points is reduced by one until an acceptable grafted polynomial is obtained.

The distribution of original-scale usual intake y is obtained from the distribution of normal scale intake x by the transformation

$$y \equiv g(x) = \begin{cases} \max\left(0, \left(10^{\beta} g_1(x)\right)^{\frac{1}{\alpha}} - \delta\right) & \text{if } \alpha \neq 0\ , \\ \max\left(0, \exp\left(10^{\beta} g_1(x)\right) - \delta\right) & \text{if } \alpha = 0\ . \end{cases} \tag{3.28}$$

3.4 Quantiles and Cumulative Distribution Function Values for Usual Intake

As explained in Section 3.3, estimates of the first five moments of the usual intake (y) distribution can be constructed from the representative sample $\{y_i\}_{i=1}^{400}$. Other characteristics of the usual intake distribution are also of interest. Consider, for $p \in [0,\ 1]$, estimation of Q_p such that

$$\Pr\left(y \leq Q_p\right) = p. \tag{3.29}$$

Alternatively, consider for $A \in [0,\ \infty]$, estimation of

$$p_0 = F(A) = \Pr\left(y \leq A\right). \tag{3.30}$$

Recall that, by construction, the y_i of (3.22) can be thought of as a sample of usual intakes in the original scale, and hence, estimates required in (3.29) and (3.30) can be obtained from the

estimated cumulative distribution function $\hat{F}$ of the $\{y_i\}_{i=1}^{400}$. The estimate $\hat{F}$ is constructed by taking

$$\hat{F}(y_i) = \widehat{\Pr}(y \leq y_i) = \frac{i - 3/8}{400.25}, \tag{3.31}$$

where $y_1 < y_2 < \ldots < y_{400}$. The slopes of the first and last step are extended to 0 and 1, respectively, to obtain $\hat{F}^{-1}(0)$ and $\hat{F}^{-1}(1)$. Linear interpolation of $\hat{F}$ is used to compute quantiles and cumulative distribution function values.

3.5 Estimated Density of Usual Intakes

Assume that X is a $N\left(\mu_x,\, \sigma_x^2\right)$ random variable, and that $y = g(x)$ is a one-to-one transformation from $(-\infty,\, \infty)$ to $[0,\, \infty)$, with derivative $g'(x) = \frac{d}{dx}g(x)$ and inverse transformation $x = g^{-1}(y)$. Assume that the derivative $\frac{d}{dy}g^{-1}(y)$ is continuous and nonzero on $[0,\, \infty)$. Then the density function of $Y = g(X)$ is

$$\begin{aligned} f_Y(y) &= \left(\sqrt{2\pi}\sigma_x\right)^{-1} \exp\left\{-\frac{1}{2}\frac{\left(g^{-1}(y) - \mu_x\right)^2}{\sigma_x^2}\right\} \left|\frac{d}{dy}g^{-1}(y)\right| \\ &= \left(\sqrt{2\pi}\sigma_x\right)^{-1} \exp\left\{-\frac{1}{2}\frac{\left(g^{-1}(y) - \mu_x\right)^2}{\sigma_x^2}\right\} \left|g'\left(g^{-1}(y)\right)\right|^{-1}. \end{aligned} \tag{3.32}$$

Equation (3.32) is used to estimate the density function of the usual intakes for a grid of points $\{x_i\}_{i=1}^{M}$, where M is specified by the user.

Chapter 4

Variances of Usual Intake Quantiles

4.1 Taylor Approximation Standard Errors

The function g of (3.28) is a differentiable transformation from x to y, and can be used to obtain approximate standard errors for percentiles and cumulative distribution function values. The following derivation assumes

1. $g(\cdot)$ is a fixed, known function, and
2. the X_{ij} consist of k independent measurements taken on each of a simple random sample of n individuals.

Suppose that the $\{X_{ij}\}$ of Section 3.1 follow the model

$$X_{ij} = x_i + u_{ij} \qquad i = 1,\ n,\ j = 1,\ \ldots,\ k\ ,$$

where

$$x_i \sim NI\left(\mu_x,\ \sigma_x^2\right), \quad u_{ij} \sim NI\left(0,\ \sigma_u^2\right)\ ,$$

and u_{ij} is independent of x_k for all $i,\ j,\ k$. Associated with the model is the ANOVA of Table 4.1. The usual estimators of $\mu_x,\ \sigma_u^2$, and σ_x^2 are

$$\begin{aligned}
\hat{\mu}_x &= \bar{X}_{\cdot\cdot} = (nk)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{k} X_{ij}\ , \\
\hat{\sigma}_u^2 &= [n(k-1)]^{-1} \sum_{i=1}^{n} \sum_{j=1}^{k} \left(X_{ij} - \bar{X}_{\cdot}\right)^2, \\
\hat{\sigma}_x^2 &= k^{-1} \left[(n-1)^{-1} k \sum_{i=1}^{n} \left(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot}\right)^2 - \hat{\sigma}_u^2 \right]\ ,
\end{aligned}$$

where $\bar{X}_{\cdot} = k^{-1} \sum_{j=1}^{k} X_{ij}$.

Under the model,

$$\mathrm{Var}\left\{\hat{\mu}_x\right\} = \frac{1}{n}\sigma_x^2 + \frac{1}{nk}\sigma_u^2 \,,$$

so

$$\widehat{\mathrm{Var}}\left\{\hat{\mu}_x\right\} = \frac{1}{n}\hat{\sigma}_x^2 + \frac{1}{nk}\,\hat{\sigma}_u^2 \tag{4.1}$$

is an unbiased estimator of $\mathrm{Var}\{\hat{\mu}_x\}$. Note that

$$\begin{aligned} T_1 &\equiv \sigma_u^{-2} \sum_{i=1}^{n}\sum_{j=1}^{k} \left(X_{ij} - \bar{X}_{i\cdot}\right)^2 \sim \chi^2_{n(k-1)} \,, \\ T_2 &\equiv \left(\sigma_u^2 + k\sigma_x^2\right)^{-1} k \sum_{i=1}^{n} \left(\bar{X}_{\cdot} - \bar{X}_{\cdot\cdot}\right)^2 \sim \chi^2_{n-1} \,, \end{aligned}$$

and T_1 and T_2 are independent. Then

$$\begin{aligned} \mathrm{Var}\left\{\hat{\sigma}_x^2\right\} &= \mathrm{Var}\left\{\frac{\sigma_u^2 + k\sigma_x^2}{k\,(n-1)}T_2 - \frac{\sigma_u^2}{kn\,(k-1)}T_1\right\} \\ &= \left[\frac{\sigma_u^2 + k\sigma_x^2}{k\,(n-1)}\right]^2 2\,(n-1) + \left[\frac{\sigma_u^2}{kn\,(k-1)}\right]^2 2n\,(k-1) \\ &= k^{-2}\left\{2\left(\sigma_u^2 + k\sigma_x^2\right)^2 (n-1)^{-1} + 2\sigma_u^4\,[n\,(k-1)]^{-1}\right\} \,. \end{aligned}$$

Therefore,

$$\widehat{\mathrm{Var}}\left\{\hat{\sigma}_x^2\right\} = k^{-2}\left\{2\left(\hat{\sigma}_u^2 + k\hat{\sigma}_x^2\right)^2 (n-1)^{-1} - 2\hat{\sigma}_u^4\,[n\,(k-1)]^{-1}\right\} \,. \tag{4.2}$$

Source	**df**	**SS**	**E{MS}**
Individual	$n-1$	$k\sum_{i=1}^{n}\left(\bar{X}_{j\cdot} - \bar{X}_{\cdot\cdot}\right)^2$	$\sigma_u^2 + k\sigma_x^2$
Residual	$n\,(k-1)$	$\sum_{i=1}^{n}\sum_{j=1}^{k}\left(X_{ij} - \bar{X}_{i\cdot}\right)^2$	σ_u^2
Total	$nk-1$	$\sum_{i=1}^{n}\sum_{j=1}^{k}\left(X_{ij} - \bar{X}_{\cdot\cdot}\right)^2$	

Table 4.1: ANOVA for one-way classification with balanced data.

From the Taylor expansion for the square root function at the point σ_x^2,

$$\hat{\sigma}_x \doteq \sigma_x + \left(4\hat{\sigma}_x^2\right)^{-1/2}\left(\hat{\sigma}_x^2 - \sigma_x^2\right) \ . \tag{4.3}$$

It follows that

$$\mathrm{Var}\left\{\hat{\sigma}_x\right\} \doteq \left(4\hat{\sigma}_x^2\right)^{-1}\mathrm{Var}\left\{\hat{\sigma}_x^2\right\} \ ,$$

which yields

$$\begin{aligned}\widehat{\mathrm{Var}}\left\{\hat{\sigma}_x\right\} &\doteq \left(4\hat{\sigma}_x^2\right)^{-1}\widehat{\mathrm{Var}}\left\{\hat{\sigma}_x^2\right\} \\ &= \left(2\hat{\sigma}_x^2 k^2\right)^{-1}\left\{(n-1)^{-1}\left(\hat{\sigma}_u^2 + k\hat{\sigma}_x^2\right)^2 + \left[n\left(k-1\right)\right]^{-1}\hat{\sigma}_u^4\right\} \ . \end{aligned} \tag{4.4}$$

Assume $x \sim N\left(\mu_x,\ \sigma_x^2\right)$. Then for any fixed A,

$$F\left(A\right) \equiv \Pr\left(x \le A\right) = \Phi\left(\sigma_x^{-1}\left(A - \mu_x\right)\right) \ ,$$

where $\Phi\left(\cdot\right)$ is the standard normal distribution function. The estimated probability that $x \le A$ is

$$\hat{F}\left(A\right) = \Phi\left(\hat{\sigma}_x^{-1}\left(A - \hat{\mu}_x\right)\right) \ . \tag{4.5}$$

Using a Taylor expansion about the point $\left(\mu_x,\ \sigma_x\right)$ and (4.3),

$$\begin{aligned}\hat{F}\left(A\right) &\doteq F\left(A\right) - \phi\left(\sigma_x^{-1}\left(A-\mu_x\right)\right)\left[\sigma_x^{-1}\left(\hat{\mu}_x - \mu_x\right) + \sigma_x^{-2}\left(A-\mu_x\right)\left(\hat{\sigma}_x - \sigma_x\right)\right] \\ &\doteq F\left(A\right) - \phi\left(\sigma_x^{-1}\left(A-\mu_x\right)\right)\left[\sigma_x^{-1}\left(\hat{\mu}_x - \mu_x\right) + \left(\sigma_x^2\right)^{-3/2}\left(A-\mu_x\right)\left(\hat{\sigma}_x^2 - \sigma_x^2\right)\right] \ , \end{aligned}$$

where $\phi\left(\cdot\right)$ is the standard normal density. Because $\hat{\mu}_x$ and $\hat{\sigma}_x^2$ are independent, it follows that

$$\mathrm{Var}\left\{\hat{F}\left(A\right)\right\} \doteq \phi^2\left(\sigma_x\left(A-\mu_x\right)\right)\left[\sigma_x^{-2}\mathrm{Var}\left\{\hat{\mu}_x\right\} + \left(\sigma_x^2\right)^{-3}\left(A-\mu_x\right)^2\mathrm{Var}\left\{\hat{\sigma}_x^2\right\}\right] \ ,$$

which yields an estimator for $\mathrm{Var}\left\{\hat{F}\left(A\right)\right\}$,

$$\widehat{\mathrm{Var}}\left\{\hat{F}\left(A\right)\right\} = \phi^2\left[\hat{\sigma}_x\left(A-\hat{\mu}_x\right)\right]\left\{\hat{\sigma}_x^{-2}\widehat{\mathrm{Var}}\left\{\hat{\mu}_x\right\} + \left(\sigma_x^2\right)^{-3}\left(A-\hat{\mu}_x\right)^2\widehat{\mathrm{Var}}\left\{\hat{\sigma}_x^2\right\}\right\} \ , \tag{4.6}$$

where $\widehat{\mathrm{Var}}\left\{\hat{\mu}\right\}$ and $\widehat{\mathrm{Var}}\left\{\hat{\sigma}_x^2\right\}$ are defined in (4.1) and (4.2), respectively. The estimated standard error of $\hat{F}\left(A\right)$ is the square root of $\widehat{\mathrm{Var}}\left\{\hat{F}\left(A\right)\right\}$.

Now, for fixed $p \in [0,\ 1]$, let Q_p denote the p^{th} percentile, that is, Q_p is such that

$$F\left(Q_p\right) = \Pr\left(x \leq Q_p\right) = p$$

or $Q_p = F^{-1}\left(p\right)$. From (4.5), an estimator of Q_p is

$$\hat{Q}_p = \hat{\sigma}_x \Phi^{-1}\left(p\right) + \hat{\mu}_x \tag{4.7}$$

and its variance is

$$\text{Var}\left\{\hat{Q}_p\right\} = \left[\Phi^{-1}\left(p\right)\right]^2 \text{Var}\left\{\hat{\sigma}_x\right\} + \text{Var}\left\{\hat{\mu}_x\right\}\ .$$

An estimator of $\text{Var}\left\{\hat{Q}_p\right\}$ is

$$\widehat{\text{Var}}\left\{\hat{Q}_p\right\} = \left[\Phi^{-1}\left(p\right)\right]^2 \widehat{\text{Var}}\left\{\hat{\sigma}_x\right\} + \widehat{\text{Var}}\left\{\hat{\mu}_x\right\}\ , \tag{4.8}$$

where $\widehat{\text{Var}}\left\{\hat{\sigma}_x\right\}$ and $\widehat{\text{Var}}\left\{\hat{\mu}_x\right\}$ are as in (4.4) and (4.1). Let the random variable y be defined by

$$y = g\left(x\right)\ ,$$

where $x \sim N\left(\mu_x,\ \sigma_x^2\right)$ and g is a known function with finite first derivative. Let $Q_y\left(p\right)$ denote the p^{th} percentile of y, i.e.

$$\Pr\left(y \leq Q_y\left(p\right)\right) = p\ ,$$

and let $Q_x\left(p\right)$ denote the p^{th} percentile of x. Then

$$Q_y\left(p\right) = g\left(Q_x\left(p\right)\right)\ .$$

A reasonable estimator of $Q_y\left(p\right)$ is

$$\hat{Q}_y\left(p\right) = g\left(\hat{Q}_x\left(p\right)\right)\ , \tag{4.9}$$

where $\hat{Q}_x\left(p\right)$ is the same as $\hat{Q}_p$ in (4.7).

To get an estimator of the variance of $\hat{Q}_y\left(p\right)$, the Taylor expansion of (4.9) at $Q_x\left(p\right)$ is used to obtain

$$\hat{Q}_y\left(p\right) \doteq g\left(Q_x\left(p\right)\right) + \frac{\partial g\left(x\right)}{\partial x}\left[\hat{Q}_x\left(p\right) - Q_x\left(p\right)\right]\ .$$

It follows that

$$\mathrm{Var}\left\{\hat{Q}_y(p)\right\} \doteq \left[\frac{\partial g(x)}{\partial x}\right]^2 \mathrm{Var}\left\{\hat{Q}_x(p)\right\} . \tag{4.10}$$

Equation (4.8) is used to obtain an estimator of (4.10) as

$$\widehat{\mathrm{Var}}\left\{\hat{Q}_y(p)\right\} \doteq \left[\frac{\partial g(x)}{\partial x}\right]^2 \left\{\left[\Phi^{-1}(p)\right]^2 \widehat{\mathrm{Var}}\{\hat{\sigma}_x\} + \widehat{\mathrm{Var}}\{\hat{\mu}_x\}\right\} . \tag{4.11}$$

Small modifications of the procedure are necessary when the number of measurements per individual are not constant. Instead of (4.1), let

$$\widehat{\mathrm{Var}}\{\hat{\mu}_x\} = \frac{1}{n}\hat{\sigma}_x^2 + \frac{1}{n}\hat{\sigma}_u^2, \tag{4.12}$$

where the multiplier for $\hat{\sigma}_u^2$ reflects the adjustment of Section 1.8.

The ANOVA of Table 4.1 yields

$$\begin{aligned}
SSB &= k\sum_{i=1}^{n}\left(X_{i.} - \bar{X}_{..}\right)^2 , \\
MSB &= (n-1)^{-1} SSB = \hat{\sigma}_u^2 + k\hat{\sigma}_x^2 , \\
d_e &= n(k-1) , \\
SSE &= \sum_{i=1}^{n}\sum_{j=1}^{k}\left(X_{ij} - \bar{X}_{.}\right)^2 , \\
MSE &= d_e^{-1} SSE = \hat{\sigma}_u^2 .
\end{aligned}$$

Then (4.2) can be written as

$$\widehat{\mathrm{Var}}\left\{\hat{\sigma}_x^2\right\} = k^{-2}\left\{2(n-1)^{-1} MSB^2 + 2d_e^{-1} MSE^2\right\} . \tag{4.13}$$

For the case where each individual has k_i independent measurements, the ANOVA of Table 3.1 yields

$$SSB = \sum_{i=1}^{n} k_i \left(\bar{X}_{i.} - \hat{\mu}_x\right)^2 , \tag{4.14}$$

$$MSB = (n-1)^{-1} SSB , \tag{4.15}$$

$$
\begin{aligned}
N &= \sum_{i=1}^{n} k_i \ , && (4.16)\\
m_0 &= \left(N - N^{-1} \sum_{i=1}^{n} k_i^2 \right) n^{-1}, && (4.17)\\
d_e &= N - n - (k-1) \ \text{, where } k = \max k_i \ , && (4.18)\\
SSE &= \sum_{i=1}^{n} \sum_{j=1}^{k_i} \left(X_{ij} - \bar{X}_{i.} \right)^2 , && (4.19)\\
MSE &= (N-n)^{-1} SSE \ , && (4.20)\\
\widehat{\text{Var}} \left\{ \hat{\sigma}_x^2 \right\} &= m_0^{-2} \left\{ 2 (n-1)^{-1} MSB^2 + 2 d_e^{-1} MSE^2 \right\} \ . && (4.21)
\end{aligned}
$$

For the case where each individual has k_i correlated measurements, the expression (4.21) is used with

$$
d_e = \begin{cases} (N-n)(1-\rho) - 1, & \text{if } k = 2 \ , \\ (m_{2c} + m_{2n} + 2m_3) - \left(m_{2c} + \frac{4}{3} m_3 \right) \rho - & \\ \left(m_{2n} + \frac{2}{3} m_3 \right) \rho^2 - 2, & \text{if } k = 3 \ , \end{cases} \qquad (4.22)
$$

where ρ is the correlation between measurements on the same individual and

$$
\begin{aligned}
m_{2c} &= \sum_{i=1}^{n} I \left(k_i = 2 \text{ consecutive days} \right) \ , \\
m_{2n} &= \sum_{i=1}^{n} I \left(k_i = 2 \text{ nonconsecutive days} \right) \ , \\
m_3 &= \sum_{i=1}^{n} I \left(k_i = 3 \right) \ .
\end{aligned}
$$

Operationally, C-SIDE obtains point estimates $\hat{Q}_y(p)$ and $\hat{F}(A)$ via linear interpolation of (3.31), but uses the derivative of g when required in (4.11).

4.2 Variance Estimation Using Replicate Weights

If the data from Section 1.1 come from a simple random sample of individuals, the results from Section 4.1 can be used to estimate the variances of estimated usual intake quantiles and cumulative

distribution function values. A similar Taylor linearization technique can be used to estimate the variances of the estimated usual intake moments. The derivations in Section 4.1 do not allow for stratification or clustering of the sampled individuals, and hence are not applicable when the data are obtained under a complex survey design.

C-SIDE supports the use of the jackknife and the balanced repeated replication (BRR) method for approximating variances of estimates obtained from survey data. Both the jackknife and the BRR method compute estimates of the parameter of interest from each of several subsamples of the parent sample. The variance of the parent sample is estimated by the variability between the subsample estimates. In what follows, let θ represent the parameter of interest.

Suppose that the data from Section 1.1 come from a stratified cluster sample with L strata and two clusters per stratum. For more complicated surveys, it may be necessary to group individuals into L "pseudo-strata," each with two "pseudo-clusters," before using the jackknife or BRR variance estimation procedures. Denote the parent sample weight for individual i, who belongs to the k^{th} cluster of the j^{th} stratum, by $W_{ijk}^{(0)}$, $j = 1, \ldots, L,\ k = 1, 2$.

- **For the jackknife method**, L sets of replicate weights $\left\{W_{ijk}^{(l)}\right\}_{i=1}^{n}$, $l = 1, \ldots, L$, are created. The l^{th} set of replicate weights assigns zero weight to the observations from a randomly selected cluster in the l^{th} stratum and doubles the weights for the observations in the remaining cluster of the l^{th} stratum.

  A more formal description of the procedure will highlight the difference between the jackknife method and the BRR method. Let $\mathbf{g}$ be an $L \times 1$ vector with l^{th} element selected at random from the set $\{-1, 1\}$. Then the jackknife weights are

$$W_{ij1}^{(l)} = \begin{cases} 2W_{ij1}^{(0)} & \text{if } j = l \text{ and } \mathbf{g}(l) = -1\ , \\ 0 & \text{if } j = l \text{ and } \mathbf{g}(l) = 1\ , \\ W_{ij1}^{(0)} & \text{if } j \neq l\ , \end{cases} \tag{4.23}$$

and

$$W_{ij2}^{(l)} \begin{cases} 0 & \text{if } j = l \text{ and } \mathbf{g}(l) = -1 \ , \\ 2W_{ij2}^{(0)} & \text{if } j = l \text{ and } \mathbf{g}(l) = 1 \ , \\ W_{ij2}^{(0)} & \text{if } j \neq l \ . \end{cases} \tag{4.24}$$

If necessary, each set of jackknife replicate weights $\left\{W_{ijk}^{(l)}\right\}_{i=1}^{n}$ can be adjusted to the same control totals as the parent sample weights $\left\{W_{ijk}^{(0)}\right\}_{i=1}^{n}$. Let $\hat{\theta}$ be the estimate of θ obtained using the parent sample weights, and let $\hat{\theta}_l,\ l = 1, \ldots, L$ be the estimate obtained using the l^{th} set of replicate weights. The jackknife estimator of the variance of $\hat{\theta}$ is given by

$$\widehat{\text{Var}}_{JK}\left(\hat{\theta}\right) = \sum_{l=1}^{L}\left(\hat{\theta}_l - \hat{\theta}\right)^2 . \tag{4.25}$$

This estimator of variance can be considered to have L degrees of freedom.

- **For the BRR Method**, M sets of replicate weights $\left\{W_{ijk}^{(l)}\right\}_{i=1}^{n},\ l = 1, \ldots, M$, are created, where M is the smallest multiple of 4 that is greater than the number of strata L. The l^{th} set of replicate weights assigns zero weight to the observations from one cluster in each stratum and doubles the weights for the observations in the remaining clusters. For a particular replicate, this construction modifies the parent sample weights for the entire data set in contrast to the construction of a jackknife replicate, which modifies the parent sample weights for the observations in only one stratum. The choice of which clusters receive zero weights in a BRR replicate is determined in a systematic manner, instead of being random, as is the case for a jackknife replicate. Let $\mathbf{H} = \mathbf{H}(j,\ l)$ be an $L \times M$ matrix obtained by taking L rows from a Hadamard matrix[1] of order M. Then the BRR weights are

$$W_{ij1}^{(l)} = \begin{cases} 2W_{ij1}^{(0)} & \text{if } \mathbf{H}(j,\ l) = -1 \ , \\ 0 & \text{if } \mathbf{H}(j,\ l) = 1 \ , \end{cases} \tag{4.26}$$

[1] A Hadamard matrix of order n is an $n \times n$ orthogonal matrix with entries ± 1.

and

$$W_{ij2}^{(l)} = \begin{cases} 0 & \text{if } \mathbf{H}(j,\ l) = -1\ , \\ 2W_{ij2}^{(0)} & \text{if } \mathbf{H}(j,\ l) = 1\ . \end{cases} \tag{4.27}$$

If necessary, each set of BRR replicate weights $\left\{W_{ijk}^{(l)}\right\}_{i=1}^{n}$ can be adjusted to the same control totals as the parent sample weights $\left\{W_{ijk}^{(0)}\right\}_{i=1}^{n}$. Let $\hat{\theta}$ be the estimate of θ obtained using the parent sample weights, and let $\hat{\theta}_l,\ l = 1, \ldots, m$ be the estimate obtained using the l^{th} set of replicate weights. The BRR estimator of the variance of $\hat{\theta}$ is given by

$$\widehat{\text{Var}}_{BRR}\left(\hat{\theta}\right) = \frac{1}{M}\sum_{l=1}^{M}\left(\hat{\theta}_l - \hat{\theta}\right)^2. \tag{4.28}$$

This estimator of variance should be considered to have only L degrees of freedom, even though $M > L$ replicate estimates are used in the computation of (4.28).

Empirical evidence (Dodd et al. 1996) suggests that the BRR estimator (4.28) is less biased and more stable than the jackknife estimator (4.25) when the parameter of interest is a usual intake quantile. For a general discussion of variance estimation for complex surveys, see Wolter (1985).

If the user supplies the appropriate BRR or jackknife replicate weights along with the parent sample weights for nutrient data, C-SIDE can produce variance estimates for usual intake quantiles, usual intake cumulative distribution function values, and usual intake moments using the formulas (4.25) and (4.28), where the calculations for each replicate are carried out internally. For food data, or for survey designs for which (4.25) and (4.28) do not hold, the user must run C-SIDE separately for the parent sample and each replicate, then combine the estimates using the appropriate analogues to (4.25) and (4.28).

Chapter 5

Analysis of Food Intake

5.1 Test for Correlation Between Intake and Probability of Consumption

The method for estimating usual intake distributions for foods requires an individual's usual intake to be independent of the individual's probability of consumption. A test for correlation between intake and probability of consumption is performed by C-SIDE.

At the first stage of analysis, the consumption day usual intake distribution is estimated using only the positive food intakes. The procedures described in Chapters 1-3 are applied to the data for the n individuals with at least one nonzero food intake. Let

$$\begin{aligned} k_i &= \text{number of days the food is consumed by the } i^{\text{th}} \text{ individual} , \\ \bar{k} &= \left(\sum_{i=1}^{n} \sum_{j=1}^{k_i} W_{ij} \right)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{k_i} W_{ij} k_i , \\ B_{ij} &= k_i - \bar{k} , \end{aligned}$$

for $i = 1, \ldots, n,\ j = 1, \ldots, k_i$, where the $\{W_{ij}\}$ are the weights defined in (1.1). Let $\mathbf{B}$ denote the column vector of observations for the variable B_{ij}, and let

$$\mathbf{M}_C = \left[\begin{array}{cc} \mathbf{M} & \mathbf{B} \end{array} \right] ,$$

where $\mathbf{M}$ is obtained from (1.13) if the positive intakes were ratio-adjusted for nuisance variables and is otherwise a column vector of ones. The weighted least squares regression with model matrix $\mathbf{M}_C$ and response variable $X_{ij}^{(s)}$ from (1.12) is performed to obtain regression parameter estimates

$$\hat{\boldsymbol{\beta}}_C = \left(\mathbf{M}_C' \text{diag}\left(W_{ij}\right) \mathbf{M}_C \right)^{-} \mathbf{M}_C' \text{diag}\left(W_{ij}\right) \mathbf{X}^{(s)},$$

where $\mathbf{X}^{(s)}$ denotes the column vector of observations $X_{ij}^{(s)}$. Let $\hat{e}_{ij}$ denote the deviation from regression,

$$\hat{e}_{ij} = X_{ij}^{(s)} - \mathbf{M}_{Cij}\hat{\boldsymbol{\beta}}_C \ ,$$

where $\mathbf{M}_{Cij}$ is the row of $\mathbf{M}_C$ corresponding to $X_{ij}^{(s)}$, the j^{th} observation for the i^{th} individual. Define

$$\begin{aligned} \mathbf{d}_{ij} &= W_{ij}\mathbf{M}_{Cij}\hat{e}_{ij} \ , \\ \mathbf{d}_{i.} &= \sum_{j=1}^{k_i} \mathbf{d}_{ij} \ , \\ \mathbf{G} &= \frac{n}{n-p}\sum_{i=1}^{n} \mathbf{d}_{i.}'\mathbf{d}_{i.} \ , \end{aligned}$$

where p is the dimension of the vector $\mathbf{M}_{Cij}$. The estimated variance of $\hat{\boldsymbol{\beta}}_C$ is

$$\widehat{\text{Var}}\left\{\hat{\boldsymbol{\beta}}_C\right\} = \mathbf{H}^{-}\mathbf{G}\mathbf{H}^{-}, \tag{5.1}$$

where $\mathbf{H} = \mathbf{M}_C'\text{diag}(W_{ij})\,\mathbf{M}_C$. Equation (5.1) is the Taylor variance approximation as computed in survey data analysis software such as PCCARP (Fuller et al. 1989). Let $\hat{\beta}_p$ be the coefficient of B_{ij} and let $\hat{v}_{pp}$ be the associated diagonal element of $\widehat{\text{Var}}\left\{\hat{\boldsymbol{\beta}}_C\right\}$. An approximate size 0.05 test for H_0 : Intake is uncorrelated with probability of consumption is to reject H_0 if

$$v_{pp}^{-1/2}\hat{\beta}_p \geq 2 \ .$$

This test statistic is output by C-SIDE.

5.2 Estimating the Distribution of Individual Consumption Probabilities

In estimating a usual intake distribution for a food, let $\pi_i \in [0,1]$ be the i^{th} individual's probability of consuming the food on a given day, $i = 1, \ldots, n$. Then for $l = 0, 1, \ldots, k$, the probability that individual i consumed the food on l out of k sample days is

$$\binom{k}{l}\pi_i^l\,(1-\pi_i)^{k-l} \ ,$$

under the assumption of independent days. The information available to support estimation of the distribution $\eta(\pi)$ of individual consumption probabilities is the proportion of sample days on which the food is consumed by the i^{th} individual. Let

$$\hat{\pi}_i = k^{-1} \sum_{j=1}^{k} \delta_{ij} \ ,$$

where for $j = 1, \ldots, k$,

$$\delta_{ij} = \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ individual consumed the food on day } j \ , \\ 0 & \text{otherwise .} \end{cases}$$

There are only $k+1$ possible values of $\hat{\pi}_i$. Let W_i be the sampling weight associated with the i^{th} individual. (Recall that the W_i are used in the construction of observation weights W_{ij} and w_{ij} in Section 1.1.) Then for $l = 0, 1, \ldots, k$,

$$\widehat{\Psi}_l = \left(\sum_{i=1}^{n} W_i \right)^{-1} \sum_{i=1}^{n} W_i I\left(\hat{\pi}_i = \frac{l}{k} \right) \tag{5.2}$$

is the weighted relative frequency of individuals who consume the food on l out of k days. If the data are equally weighted, each W_i is one.

The consumption probability distribution $\eta(\cdot, \boldsymbol{\theta})$ is modeled as a discrete distribution with $M+1$ probability values

$$\{p_m\}_{m=0}^{M} = \{0, p_1, \ldots, p_M\}$$

with corresponding probability masses $\{\theta_m\}_{m=0}^{M}$. Under the model, $\widehat{\Psi}_l$ arises from a mixture of the $M+1$ binomial probabilities of consumption on l out of k days, with binomial parameters k and p_m, and mixture parameters $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_M)$, where $\theta_m \in [0,1]$ and $\sum_{m=0}^{M} \theta_m = 1$. Hence, the expected value of $\widehat{\Psi}_l$ is equal to

$$\Psi_l(\boldsymbol{\theta}) = \sum_{m \in E_l} \theta_m \binom{k}{l} p_m^l \left(1 - p_m\right)^{k-l} ,$$

where

$$E_l = \begin{cases} \{0, 1, \ldots, M-1\} & \text{if } l < k \ , \\ \{1, 2, \ldots, M\} & \text{if } l = k \ . \end{cases}$$

The *minimum chi-squared estimator* (Agresti, 1990, page 471) for this problem is the value of $\boldsymbol{\theta}$ that minimizes

$$n \sum_{l=0}^{k} \left[\widehat{\Psi}_l - \Psi_l \left(\boldsymbol{\theta} \right) \right]^2 \left[\Psi_l \left(\boldsymbol{\theta} \right) \right]^{-1} . \tag{5.3}$$

The number of estimated parameters, $M + 1$, exceeds the number of terms in the chi-squared objective function, $k + 1$.

The *maximum entropy estimator* (Shannon 1948; Jaynes 1957) for this problem is the value of $\boldsymbol{\theta}$ that maximizes

$$\Gamma \left(\boldsymbol{\theta} \right) = - \sum_{m=0}^{M} \theta_m \ln \theta_m \tag{5.4}$$

subject to $\sum_{m=0}^{M} \theta_m = 1$ and constraints representing prior information on the θ_m, where $\theta_m \in [0, 1]$ and $\theta \ln \theta$ is zero for $\theta = 0$. In the absence of any prior information, $\Gamma \left(\boldsymbol{\theta} \right)$ is maximized when $\theta_m = (M + 1)^{-1}$ for $m = 0, 1, \ldots, M$. The function $\Gamma \left(\boldsymbol{\theta} \right)$ reaches a global minimum when θ_m is one for some m_0 and zero for all other values of m. Combining the chi-square and entropy criteria leads to a *modified minimum chi-squared estimator* similar to that found on page 472 of Agresti (1990). The estimator used in C-SIDE is defined as the value of $\boldsymbol{\theta}$ that minimizes

$$n \sum_{l=0}^{k} \left[\widehat{\Psi}_l - \Psi_l \left(\boldsymbol{\theta} \right) \right]^2 \widetilde{\Psi}_l^{-1} + \sum_{m=1}^{M} \frac{\theta_m}{1 - \theta_0} \ln \left(\frac{\theta_m}{1 - \theta_0} \right) \ , \tag{5.5}$$

where $\sum_{m=0}^{M} \theta_m = 1$, $\theta_m \in [0, 1]$,

$$\widetilde{\Psi}_l = \begin{cases} \max \left\{ \widehat{\Psi}_0, \ \left(1 - \overline{\Psi} \right)^k \right\} & \text{if } l = 0 \ , \\ \max \left\{ \widehat{\Psi}_k, \ \overline{\Psi}^k \right\} & \text{if } l = k \ , \\ \widehat{\Psi}_l \left(1 - \widetilde{\Psi}_0 - \widetilde{\Psi}_k \right) \left(1 - \widehat{\Psi}_0 - \widehat{\Psi}_k \right)^{-1} & \text{if } l = 1, 2, \ldots, k-1 \ , \end{cases}$$

and

$$\overline{\Psi} = \left(k \sum_{i=1}^{n} W_i \right)^{-1} \sum_{i=1}^{n} W_i \sum_{j=1}^{k} \delta_{ij} \ .$$

The first term in (5.5) contains the sample information. The modified denominator in this term prevents numerical difficulties that can arise when $\widehat{\Psi}_l = 0$. The second term in (5.5) is a maximum entropy term that smooths the probability mass across all possible values of p_m given the sample information in the first term. Note that θ_0, which is the proportion of the population that never consumes the food, is not included in the entropy term. C-SIDE uses the value 50 for M. The estimated consumption probability distribution has mass on the points $\{0, 0.02, 0.04, \ldots, 0.98, 1\}$.

For a given set of $\widehat{\Psi}_l$, the vector $\boldsymbol{\theta}_{\text{opt}}$ that minimizes (5.5) is obtained by the following algorithm, in which the symbol $\leftarrow$ means "receives the value of."

1. Let $q_m \leftarrow 0$ for $m = 0, 1, \ldots, 50$, and let $L_{\text{opt}} \leftarrow 10^6$.

2. Steps 2a-2c are performed for $i = 1, \ldots, 50,000$.

   2a. For $j = 0, 1, \ldots, 50$, let

   $$L_j = n \sum_{l=0}^{k} \left[\widehat{\Psi}_l - \Psi_l \left(\boldsymbol{\theta}^{(j)} \right) \right]^2 \widetilde{\Psi}_l^{-1} + \sum_{m=1}^{50} \frac{\theta_m^{(j)}}{1 - \theta_0^{(j)}} \ln \left(\frac{\theta_m^{(j)}}{1 - \theta_0^{(j)}} \right) ,$$

   where the j^{th} element of $\boldsymbol{\theta}^{(j)}$ is $i^{-1}(q_j + 1)$ and all other elements of $\boldsymbol{\theta}^{(j)}$ are $i^{-1} q_j$.

   2b. For the value of j corresponding to the minimum L_j value, $q_j \leftarrow q_j + 1$.

   2c. If $\min_j(L_j) < L_{\text{opt}}$, $Q_m \leftarrow q_m$ for $m = 0, 1, \ldots, 50$, $i_{\text{opt}} \leftarrow i$, and $L_{\text{opt}} \leftarrow \min_j(L_j)$.

3. For $m = 0, 1, \ldots, 50$, the m^{th} element of $\boldsymbol{\theta}_{\text{opt}}$ is

   $$\theta_m = \frac{Q_m}{i_{\text{opt}}} .$$

5.3 Estimation of Usual Food Intake Distributions

The usual intake of a food for individual i, denoted by u_i, is

$$u_i = y_i \pi_i ,$$

where y_i is the usual intake given that the food is consumed and π_i is the probability of positive consumption. The distribution of usual intakes for consumers is the distribution of $y_i\pi_i$, which can be derived from the joint distribution of y_i and π_i. In C-SIDE, the distribution is computed under the assumption that y_i is independent of π_i. Then the joint distribution of y_i and π_i is the product of the respective marginal distributions.

Let $F_y(y)$ denote the consumption day usual intake distribution, with associated density function $f_y(y)$. Let $\eta(\pi;\, \boldsymbol{\theta}) = \{(p_m,\, \theta_m)\}_{m=0}^{50}$ denote the consumption probability distribution. The cumulative distribution function of usual intake for the entire population is then

$$\begin{aligned} F_U(u) = \Pr(y\pi \leq u) &= \theta_0 + \sum_{m=1}^{50} \theta_m \int_0^{u/p_m} f_y(y)\, dy \\ &= \theta_0 + \sum_{m=1}^{50} \theta_m F_y(u/p_m) \ . \end{aligned} \tag{5.6}$$

The cumulative distribution function of usual intake for consumers is

$$F_C(u) = (1-\theta_0)^{-1}\left[F_U(u) - \theta_0\right] \ , \tag{5.7}$$

where consumers are those individuals with $\pi_i > 0$. The methodology described in Chapters 1-3 applied to the positive daily intakes yields the estimated consumption day usual intake distribution $\hat{F}_y(y)$, and the method described in Section 5.2 yields the estimated consumption probability distribution $\hat{\eta}\left(\pi;\, \hat{\boldsymbol{\theta}}\right) = \left\{\left(p_m,\, \hat{\theta}_m\right)\right\}_{m=0}^{50}$. Substituting the appropriate estimates into Equations (5.6) and (5.7) yields $\hat{F}_U(u)$ and $\hat{F}_C(u)$, respectively, where $\hat{F}_U(u)$ is the estimated usual intake distribution for the entire population and $\hat{F}_C(u)$ is the estimated usual intake distribution for consumers. The estimated distributions are used to obtain a representative sample of size $N = 400$ of estimated usual intakes for consumers. The representative sample is $\{\gamma_i\}_{i=1}^{400}$, where $\tau_i = (400.25)^{-1}\left(i - \frac{3}{8}\right)$ and γ_i is the τ_i^{th} quantile of $\hat{F}_C$ defined by

$$\begin{aligned} \tau_i &= \left(1-\hat{\theta}_0\right)^{-1}\left[\hat{F}_U(\gamma_i) - \hat{\theta}_0\right] \\ &= \left(1-\hat{\theta}_0\right)^{-1}\left[\sum_{m=1}^{50} \hat{\theta}_m \hat{F}_y(\gamma_i/p_m)\right] \ . \end{aligned}$$

The following iterative procedure is used to estimate $\gamma_1, \gamma_2, \ldots, \gamma_{400}$.

1. Let $(\bar{p}, \bar{\theta})$ be the element of $\left\{\left(p_m,\ \hat{\theta}_m\right)\right\}_{\hat{\theta}_m \neq 0}$ with the smallest nonzero second coordinate. The initial approximation $\gamma^{(0)}$ to γ_1 is

$$\gamma^{(0)} = \bar{p}\bar{y}\ ,$$

where $\bar{y}$ is the $\left(\bar{\theta}^{-1}\tau_1\left(1-\hat{\theta}_0\right)\right)^{\text{th}}$ quantile of the estimated consumption day usual intake distribution. The quantities $\gamma_b, \gamma_s, f_b, f_s$, and γ_h are initialized with 0.

2. Steps 2a-2c are performed for $i = 1, \ldots, 400$.

   2a. Denote the current approximation to γ_i by $\gamma^{(k)}$, and let

$$\begin{aligned} f &= \left(1-\hat{\theta}_0\right)^{-1} \sum_{m=1}^{50} \hat{\theta}_m \hat{F}_y\left(\gamma^{(k)}/p_m\right)\ , \\ D &= f - \tau_i\ . \end{aligned}$$

   If $|D| < 10^{-6}$, go to 2b. If $|D| \geq 10^{-6}$, go to 2c.

   2b. $\left(|D| < 10^{-6}\right)$

$$\gamma_i \leftarrow \gamma^{(k)}.$$

   The initial approximation $\gamma^{(0)}$ to γ_{i+1} is then obtained.

$$\begin{aligned} \gamma_b &\leftarrow 0\ , \\ f_b &\leftarrow 0\ , \\ \gamma_s &\leftarrow \gamma^{(k)}, \\ f_s &\leftarrow f\ , \\ m &\leftarrow \begin{cases} f^{-1}\gamma^{(k)} & \text{if } i = 1\ , \\ (f-\tau_{i-1})^{-1}\left(\gamma^{(k)} - \gamma_{i-1}\right) & \text{if } i > 1\ , \end{cases} \\ \gamma^{(0)} &\leftarrow \gamma^{(k)} - m\left(f - \tau_{i+1}\right)\ , \\ i &\leftarrow i+1\ . \end{aligned}$$

Go back to Step 2a.

2c. $\left(|D| \geq 10^{-6}\right)$ A new approximation $\gamma^{(k+1)}$ to γ_i is obtained.

If $D > 0$,

$$\begin{aligned} f_b &\leftarrow f\ , \\ \gamma_b &\leftarrow \gamma^{(k)}, \end{aligned}$$

and if $f > \tau_{400}$ and $i < 400$,

$$\gamma_h \leftarrow \gamma^{(k)}.$$

If $D < 0$,

$$\begin{aligned} f_s &\leftarrow f\ , \\ \gamma_s &\leftarrow \gamma^{(k)}. \end{aligned}$$

If $k < 10$,

$$\begin{aligned} m &\leftarrow \left(f_b - f_s\right)^{-1}\left(\gamma_b - \gamma_s\right)\ , \\ \gamma^{(k+1)} &\leftarrow \gamma^{(k)} - mD\ , \\ k &\leftarrow k + 1\ . \end{aligned}$$

Go back to Step 2a.

If $k \geq 10$,

- If $\gamma_b = 0$ and $\gamma_h > 0$, $\gamma_b \leftarrow \gamma_h$.
- If $\gamma_b = 0$ and $\gamma_h = 0$, $\gamma_h \leftarrow \ddot{y}$, where $\ddot{y}$ is the estimated τ_{400}^{th} quantile of the consumption day usual intake distribution. Then $\gamma_b \leftarrow \gamma_h$.

$$\begin{aligned} \gamma^{(k+1)} &\leftarrow \gamma^{(k)} + \frac{1}{2}\left(\gamma_b - \gamma_s\right) , \\ k &\leftarrow k+1 , \end{aligned}$$

Go back to Step 2a.

The $\{\gamma_i\}_{i=1}^{400}$ are a representative sample of estimated usual intakes for consumers. Percentiles of the F_C distribution are estimated via linear interpolation of the empirical cumulative distribution function of the representative sample, and moments for the usual intake distribution for consumers are estimated by the moments of the representative sample. The estimated proportion of consumers with usual intake below γ_i is $\tau_i = (400.25)^{-1}\left(i - \frac{3}{8}\right)$. It follows that γ_i estimates the $\left(\tau_i\left(1-\hat{\theta}_0\right)+\hat{\theta}_0\right)^{\text{th}}$ quantile of the usual food intake distribution, F_U, for the entire population.

Let $x_i = \Phi^{-1}(\tau_i)$ for $i = 1, \ldots, 400$, where $\Phi^{-1}(\cdot)$ denotes the inverse of the standard normal cumulative distribution function. The x_i are the Blom normal scores for a sample of size 400 from the standard normal distribution. The methodology of Chapter 2 is used to estimate a transformation $G(u)$ that carries usual intakes u for consumers to standard normality. The density $f_C(u)$ of the consumer distribution is estimated using Equation (3.32) with $\mu_x = 0$, $\sigma_x^2 = 1$ and $g^{-1}(\cdot) = G(\cdot)$. The usual intake distribution $F_U(u)$ for the entire population is a mixture of a single value (0) and a continuous distribution ($F_C(u)$) with mixing parameters θ_0 and $(1-\theta_0)$, respectively.

Let u denote the usual intake of a food for a randomly selected individual on an arbitrary day. Let C be a Bernoulli random variable:

$$C = \begin{cases} 0 & \text{if the person never consumes the food ,} \\ 1 & \text{otherwise .} \end{cases}$$

The probability that $C = 0$ is θ_0. Denote the mean of the consumer distribution by

$$\mu_C = \mathrm{E}\left\{u | C = 1\right\} .$$

An estimator of μ_C computed from the $\{\gamma_i\}_{i=1}^{400}$ is

$$\hat{\mu}_C = \frac{1}{400}\sum_{i=1}^{400}\gamma_i \,, \tag{5.8}$$

The mean of the whole-population distribution is

$$\mu_U = 0\Pr\left(C=0\right) + \mathrm{E}\left\{u|C=1\right\}\Pr\left(C=1\right) \tag{5.9}$$

and the estimator for μ_U is

$$\hat{\mu}_U = \hat{\mu}_C\left(1-\hat{\theta}_0\right) \,. \tag{5.10}$$

Denote the variance of the consumer distribution by

$$\sigma_C^2 = \mathrm{Var}\left\{u|C=1\right\} \,.$$

An estimator of σ_C^2 computed from the $\{\gamma_i\}_{i=1}^{400}$ is

$$\hat{\sigma}_C^2 = \frac{1}{399}\sum_{i=1}^{400}\left(\gamma_i - \hat{\mu}_C\right)^2 . \tag{5.11}$$

The estimated variance of u is obtained from the formula

$$\mathrm{Var}\left\{u\right\} = \mathrm{Var}\left\{\mathrm{E}\left\{u|C\right\}\right\} + \mathrm{E}\left\{\mathrm{Var}\left\{u|C\right\}\right\} \,.$$

Now,

$$\mathrm{Var}\left\{u|C\right\} = \begin{cases} 0 & \text{if } C=0 \,, \\ \sigma_C^2 & \text{if } C=1 \,, \end{cases}$$

so that

$$\mathrm{E}\left\{\mathrm{Var}\left\{u|C\right\}\right\} = \sigma_C^2\left(1-\theta_0\right) \,.$$

Also,

$$\begin{aligned} \mathrm{Var}\left\{\mathrm{E}\left\{u|C\right\}\right\} &= \mathrm{E}\left\{\mathrm{E}\left\{u|C\right\}^2\right\} - \mathrm{E}\left\{\mathrm{E}\left\{u|C\right\}\right\}^2 \\ &= \mu_C^2\left(1-\theta_0\right) - \left[\mu_C\left(1-\theta_0\right)\right]^2 \\ &= \mu_C^2\left[\left(1-\theta_0\right) - \left(1-\theta_0\right)^2\right] \\ &= \mu_C^2\theta_0\left(1-\theta_0\right) \,. \end{aligned}$$

Thus, the variance of the whole-population distribution is

$$\sigma_U^2 = \sigma_C^2 \left(1 - \theta_0\right) + \mu_C^2 \theta_0 \left(1 - \theta_0\right) \ , \tag{5.12}$$

and its estimator is

$$\hat{\sigma}_U^2 = \hat{\sigma}_C^2 \left(1 - \hat{\theta}_0\right) + \hat{\mu}_C^2 \hat{\theta}_0 \left(1 - \hat{\theta}_0\right) \ . \tag{5.13}$$

Higher order moments are estimated in a similar fashion, using $\hat{\theta}_0$ and the estimated moments of the consumer distribution. Formulas similar to (1.5) and (1.6) are used to estimate the third and fourth moments of the consumer distribution. The replicate weighting methods described in Section 4.2 may be used to calculate approximate standard errors for the estimated moments, although the C-SIDE internal replicate weighting is not available for foods.

References

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

Blom, G. (1958). *Statistical Estimates and Transformed Beta Variables*. New York: John Wiley & Sons. Inc.

Carriquiry, A. L., W. A. Fuller, J. J. Goyeneche, and H. H. Jensen (1995). "Estimated Correlations Among Days for the Combined 1989-91 CSFII." Dietary Assessment Research Series 4. CARD Staff Report 95-SR77. Center for Agricultural and Rural Development. Iowa State University, Ames.

Department of Statistics and the Center for Agricultural and Rural Development, Iowa State University (1996). *A User's Guide to C-SIDE: Software for Intake Distribution Estimation Version 1.0*. Technical Report 96-TR 31. Center for Agricultural and Rural Development, Iowa State University, Ames.

Dodd, K. W., A. L. Carriquiry, and W. A. Fuller (1996). "Replicate Weighting Methods for Quantile Variance Estimation." *Proceedings of the Section on Survey Research Methods*. American Statistical Association. Washington. D.C. (in press).

Fisher, R. A. (1963). *Statistical Methods for Research Workers*, 13th edition. New York: Hafner Publishing Company.

Fuller, W. A., W. Kennedy, D. Schnell, G. Sullivan, H. J. Park (1989). *PCCARP*. Ames: Iowa State University.

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

Nusser, S. M., A. L. Carriquiry, K. W. Dodd, and W. A. Fuller (1996a). "A semiparametric transformation approach to estimating usual daily intake distributions." *Journal of the American Statistical Association*. (in press).

Nusser, S. M., W. A. Fuller, and P. G. Guenther (1996b). "Estimating usual dietary intake distributions: Adjusting for measurement error and non-normality in 24-hour food intake data." In L. Lyberg, P. Beimer, M. Collins, E. De Leeuw, C. Dippo, N. Schwartz, and D. Trewin (ed.) *Survey Measurement and Process Quality*. New York: John Wiley & Sons. Inc. (in press).

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

Stephens, M. A. (1974). "EDF statistics for goodness of fit and some comparisons." *Journal of the American Statistical Association* 69:730-737.

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